Complete Dictionary Recovery over the Sphere

Ju Sun, Qing Qu, and John Wright

{js4038, qq2105, jw2966}@columbia.edu

Department of Electrical Engineering, Columbia University, New York, USA

April 26, 2015

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 the 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. In contrast, prior results based on efficient algorithms provide recovery guarantees when X_0 has only $O(\sqrt{n})$ nonzeros per column.

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. To show this apparently hard problem is tractable, we first provide a geometric characterization of the high-dimensional objective landscape, which shows that with high probability there are no "spurious" local minima. This particular geometric structure allows us to design a Riemannian trust region algorithm over the sphere that provably converges to one local minimizer with an arbitrary initialization, despite the presence of saddle points. The geometric approach we develop here may also shed light on other problems arising from nonconvex recovery of structured signals.

Keywords. Dictionary learning, Nonconvex optimization, Spherical constraint, Trust region method, Manifold optimization, Function landscape, Second-order geometry, Inverse problem, Structured signal, Nonlinear approximation

Mathematics Subject Classification. 68P30, 58C05, 94A12, 94A08, 68T05, 90C26, 90C48, 90C55

Contents

1   Introduction

Given p signal samples from $\mathbb{R}^n$, i.e., $\boldsymbol{Y} \doteq [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_p]$, is it possible to construct a dictionary $\boldsymbol{A} \doteq [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m]$ with m much smaller than p, such that $\boldsymbol{Y} \approx \boldsymbol{AX}$ and the coefficient matrix $\boldsymbol{X}$ has as few nonzeros as possible? In other words, this model *dictionary learning* (DL) problem seeks a concise representation for a collection of input signals. Concise signal representations play a central role in compression, and also prove useful for many other important tasks, such as signal acquisition, denoising, and classification.

Traditionally, concise signal representations have relied heavily on explicit analytic bases constructed in nonlinear approximation and harmonic analysis. This constructive approach has proved highly successfully; the numerous theoretical advances in these fields (see, e.g., [DeV98, Tem03, DeV09, Can02, MP10a] for summary of relevant results) provide ever more powerful representations, ranging from the classic Fourier to modern multidimensional, multidirectional, multiresolution bases, including wavelets, curvelets, ridgelets, and so on. However, two challenges confront practitioners in adapting these results to new domains: which function class best describes signals at hand, and consequently which representation is most appropriate. These challenges are coupled, as function classes with known "good" analytic bases are rare. [1]

Around 1996, neuroscientists Olshausen and Field discovered that sparse coding, the principle of encoding a signal with few atoms from a learned dictionary, reproduces important properties of the receptive fields of the simple cells that perform early visual processing [OF96, OF97]. The discovery has spurred a flurry of algorithmic developments and successful applications for DL in the past two decades, spanning classical image processing, visual recognition, compressive signal acquisition, and also recent deep architectures for signal classification (see, e.g., [Ela10, MBP14] for review this development).

The learning approach is particularly relevant to modern signal processing and machine learning, which deal with data of huge volume and great variety (e.g., images, audios, graphs, texts, genome sequences, time series, etc). The proliferation of problems and data seems to preclude analytically deriving optimal representations for each new class of data in a timely manner. On the other hand, as datasets grow, learning dictionaries directly from data looks increasingly attractive and promising. When armed with sufficiently many data samples of one signal class, by solving the model DL problem, one would expect to obtain a dictionary that allows sparse representation for the whole class. This hope has been borne out in a number of successful examples [Ela10, MBP14] and theories [MP10b, VMB11, MG13, GJB$^+$13].

---

[1] As Donoho et al [DVDD98] put it, "...in effect, uncovering the optimal codebook structure of naturally occurring data involves more challenging empirical questions than any that have ever been solved in empirical work in the mathematical sciences."

1.1 Theoretical and Algorithmic Challenges

In contrast to the above empirical successes, the theoretical study of dictionary learning is still developing. For applications in which dictionary learning is to be applied in a "hands-free" manner, it is desirable to have efficient algorithms which are guaranteed to perform correctly, when the input data admit a sparse model. There have been several important recent results in this direction, which we will review in Section 1.5, after our sketching main results. Nevertheless, obtaining algorithms that provably succeed under broad and realistic conditions remains an important research challenge.

To understand where the difficulties arise, we can consider a model formulation, in which we attempt to obtain the dictionary A and coefficients X which best trade-off sparsity and fidelity to the observed data:

$$\text{minimize}_{A\in\mathbb{R}^{n\times m}, X\in\mathbb{R}^{m\times p}} \ \lambda\|X\|_1 + \frac{1}{2}\|AX - Y\|_F^2, \ \text{subject to} \ A \in \mathcal{A}. \tag{1.1}$$

Here, $\|X\|_1 \doteq \sum_{i,j}|X_{ij}|$ promotes sparsity of the coefficients, $\lambda \geq 0$ trades off the level of coefficient sparsity and quality of approximation, and $\mathcal{A}$ imposes desired structures on the dictionary.

This formulation is nonconvex: the admissible set $\mathcal{A}$ is typically nonconvex (e.g., orthogonal group, matrices with normalized columns)[2], while the most daunting nonconvexity comes from the bilinear mapping: $(A, X) \mapsto AX$. Because (A, X) and $(A\Pi\Sigma, \Sigma^{-1}\Pi^*X)$ result in the same objective value for the conceptual formulation (1.1), where Π is any permutation matrix, and Σ any diagonal matrix with diagonal entries in $\{\pm 1\}$, and $(\cdot)^*$ denotes matrix transpose. Thus, we should expect the problem to have combinatorially many global minima. Because there are multiple isolated global minima, the problem does not appear to be amenable to convex relaxation (see similar discussions in, e.g., [GS10] and [GW11]).[3] This contrasts sharply with problems in sparse recovery and compressed sensing, in which simple convex relaxations are often provably effective [DT09, OH10, CLMW11, DGM13, MT14, MHWG13, CRPW12, CSV13, ALMT14, Can14]. Is there any hope to obtain global solutions to the DL problem?

1.2 An Intriguing Numerical Experiment with Real Images

We provide empirical evidence in support of a positive answer to the above question. Specifically, we learn orthogonal bases (orthobases) for real images patches. Orthobases are of interest because typical hand-designed dictionaries such as discrete cosine (DCT) and wavelet bases are orthogonal, and orthobases seem competitive in performance for applications such as image denoising, as compared to overcomplete dictionaries [BCJ13][4].

---

[2]For example, in nonlinear approximation and harmonic analysis, orthonormal basis or (tight-)frames are preferred; to fix the scale ambiguity discussed in the text, a common practice is to require that A to be column-normalized. There is no obvious reason to believe that convexifying these constraint sets would leave the optima unchanged. For example, the convex hull of the orthogonal group O_n is the operator norm ball $\{X \in \mathbb{R}^{n\times n} : \|X\| \leq 1\}$. If there are no effective symmetry breaking constraints, any convex objective function tends to have minimizers inside the ball, which obviously will not be orthogonal matrices. Other ideas such as lifting may not play together with the objective function, nor yield tight relaxations (see, e.g., [BKS13, BR14]).

[3]Semidefinite programming (SDP) lifting may be one useful general strategy to convexify bilinear inverse problems, see, e.g., [ARR14, CM14]. However, for problems with general nonlinear constraints, it is unclear whether the lifting always yield tight relaxation, consider, e.g., [BKS13, BR14] again.

[4]See Section 1.3 for more detailed discussions of this point. [LGBB05] also gave motivations and algorithms for learning (union of) orthobases as dictionaries.
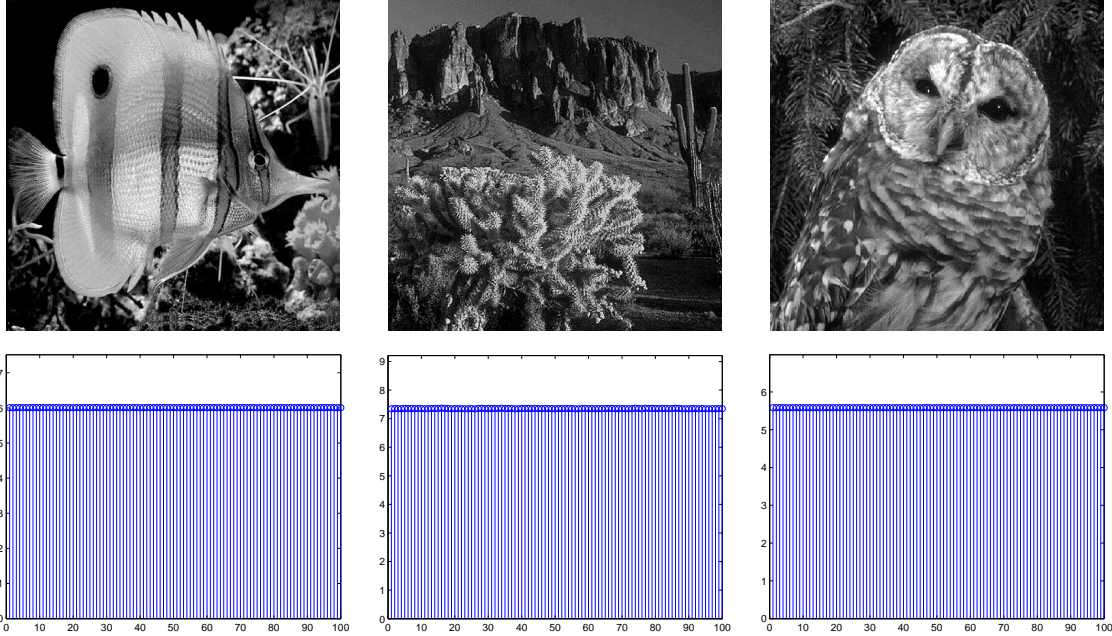
Figure 1: Alternating direction method for (1.2) **on uncompressed real images seems to always produce the same solution! Top**: Each image is 512×512 in resolution and encoded in the uncompressed `pgm` format (uncompressed images to prevent possible bias towards standard bases used for compression, such as DCT or wavelet bases). Each image is evenly divided into 8×8 non-overlapping image patches (4096 in total), and these patches are all vectorized and then stacked as columns of the data matrix $\boldsymbol{Y}$. **Bottom**: Given each $\boldsymbol{Y}$, we solve (1.2) 100 times with independent and randomized (uniform over the orthogonal group) initialization $\boldsymbol{A}_0$. The plots show the values of $\|\boldsymbol{A}_\infty^* \boldsymbol{Y}\|_1$ across the independent repetitions. They are virtually the same and the relative differences are less than 10^{-3}!

We divide a given greyscale image into 8×8 non-overlapping patches, which are converted into 64-dimensional vectors and stacked column-wise into a data matrix $\boldsymbol{Y}$. Specializing (1.1) to this setting, we obtain the optimization problem:

$$\text{minimize}_{\boldsymbol{A}\in\mathbb{R}^{n\times n}, \boldsymbol{X}\in\mathbb{R}^{n\times p}} \ \lambda \|\boldsymbol{X}\|_1 + \frac{1}{2} \|\boldsymbol{A}\boldsymbol{X} - \boldsymbol{Y}\|_F^2, \text{ subject to } \boldsymbol{A} \in O_n. \tag{1.2}$$

To derive a concrete algorithm for (1.2), one can deploy the alternating direction method (ADM)[5], i.e., alternately minimizing the objective function with respect to (w.r.t.) one variable while fixing the other. The iteration sequence actually takes very simple form: for $k = 1, 2, 3, \dots$,

$$\boldsymbol{X}_k = \mathcal{S}_\lambda \left[\boldsymbol{A}_{k-1}^* \boldsymbol{Y} \right], \qquad \boldsymbol{A}_k = \boldsymbol{U}\boldsymbol{V}^* \text{ for } \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^* = \texttt{SVD}\left(\boldsymbol{Y}\boldsymbol{X}_k^*\right)$$

where $\mathcal{S}_\lambda \left[\cdot \right]$ denotes the well-known soft-thresholding operator acting elementwise on matrices, i.e., $\mathcal{S}_\lambda \left[x \right] \doteq \text{sign}\left(x\right) \max\left(|x| - \lambda, 0\right)$ for any scalar x.

Figure 1 shows what we obtained using the simple ADM algorithm, with *independent and randomized initializations*:

---

[5]This method is also called alternating minimization or (block) coordinate descent method. see, e.g., [BT89, Tse01] for classic results and [ABRS10, BST14] for several interesting recent developments.

The algorithm seems to always produce the same solution, regardless of the initialization.

This observation implies the heuristic ADM algorithm may *always converge to one global minimizer*! [6] Equally surprising is that the phenomenon has been observed on real images[7]. One may imagine only random data typically have "favorable" structures; in fact, almost all existing theories for DL pertain only to random data [SWW12, AAJ+13, AGM13, AAN13, ABGM14, AGMM15].

1.3 Dictionary Recovery and Our Results

In this paper, we take a step towards explaining the surprising effectiveness of nonconvex optimization heuristics for DL. We focus on the *dictionary recovery* (DR) setting: given a data matrix Y generated as $Y = A_0 X_0$, where $A_0 \in \mathcal{A} \subseteq \mathbb{R}^{n \times m}$ and $X_0 \in \mathbb{R}^{m \times p}$ is "reasonably sparse", try to recover A_0 and X_0. Here recovery means to return any pair $\left(A_0 \Pi \Sigma, \Sigma^{-1} \Pi^* X_0 \right)$, where Π is a permutation matrix and Σ is a nonsingular diagonal matrix, i.e., recovering up to sign, scale, and permutation.

To define a reasonably simple and structured problem, we make the following assumptions:

- The target dictionary A_0 is complete, i.e., square and invertible ($m = n$). In particular, this class includes orthogonal dictionaries. Admittedly overcomplete dictionaries tend to be more powerful for modeling and to allow sparser representations. Nevertheless, most classic hand-designed dictionaries in common use are orthogonal. Orthobases are competitive in performance for certain tasks such as image denoising [BCJ13], and admit faster algorithms for learning and encoding. [8]

- The coefficient matrix X_0 follows the Bernoulli-Gaussian (BG) model with rate θ: $[X_0]_{ij} = \Omega_{ij} V_{ij}$, with $\Omega_{ij} \sim \mathrm{Ber}(\theta)$ and $V_{ij} \sim \mathcal{N}(0, 1)$, where all the different random variables are mutually independent. We write compactly $X_0 \sim_{i.i.d.} \mathrm{BG}(\theta)$.

We prove the following result:

Theorem 1.1 (Informal statement of our results) *For any $\theta \in (0, 1/3)$, given $Y = A_0 X_0$ with A_0 a complete dictionary and $X_0 \sim_{i.i.d.} \mathrm{BG}(\theta)$, there is a polynomial time algorithm that recovers A_0 and X_0 with high probability (at least $1 - O(p^{-6})$) whenever $p \geq p_\star (n, 1/\theta, \kappa(A_0), 1/\mu)$ for a fixed polynomial $p_\star(\cdot)$, where $\kappa(A_0)$ is the condition number of A_0 and μ is a parameter that can be set as $cn^{-5/4}$ for a fixed positive numerical constant c.*

Obviously, even if X_0 is known, one needs $p \geq n$ to make the identification problem well posed. Under our particular probabilistic model, a simple coupon collection argument implies that one needs $p \geq \Omega\left(\frac{1}{\theta} \log n\right)$ to ensure all atoms in A_0 are observed with high probability (w.h.p.). To

---

[6]Technically, the converge to global solutions is surprising because even convergence of ADM to critical points is atypical, see, e.g., [ABRS10, BST14] and references therein. Section 6 includes more detailed discussions on this point.

[7]Actually the same phenomenon is also observed for simulated data when the coefficient matrix obeys the Bernoulli-Gaussian model, which is defined later. The result on real images supports that previously claimed empirical successes over two decades may be non-incidental.

[8]Empirically, there is no systematic evidence supporting that overcomplete dictionaries are strictly necessary for good performance in all published applications (though [OF97] argues for the necessity from neuroscience perspective). Some of the ideas and tools developed here for complete dictionaries may also apply to certain classes of structured overcomplete dictionaries, such as tight frames. See Section 6 for relevant discussion.

ensure that an efficient algorithm exists may demand more. Our result implies when p is polynomial in n, $1/\theta$ and $\kappa(\boldsymbol{A}_0)$, recovery with efficient algorithm is possible.

The parameter θ controls the sparsity level of $\boldsymbol{X}_0$. Intuitively, the recovery problem is easy for small θ and becomes harder for large θ.[9] It is perhaps surprising that an efficient algorithm can succeed up to constant θ, i.e., linear sparsity in $\boldsymbol{X}_0$. Compared to the case when $\boldsymbol{A}_0$ is known, there is only at most a constant gap in the sparsity level one can deal with.

For DL, our result gives the first efficient algorithm that provably recovers complete $\boldsymbol{A}_0$ and $\boldsymbol{X}_0$ when $\boldsymbol{X}_0$ has $O(n)$ nonzeros per column under appropriate probability model. Section 1.5 provides detailed comparison of our result with other recent recovery results for complete and overcomplete dictionaries.

1.4 Main Ingredients and Innovations

In this section we describe three main ingredients that we use to obtain the stated result.

1.4.1 A Nonconvex Formulation

Since $\boldsymbol{Y} = \boldsymbol{A}_0\boldsymbol{X}_0$ and $\boldsymbol{A}_0$ is complete, $\text{row}(\boldsymbol{Y}) = \text{row}(\boldsymbol{X}_0)$ ($\text{row}(\cdot)$ denotes the row space of a matrix) and hence rows of $\boldsymbol{X}_0$ are sparse vectors in the known (linear) subspace $\text{row}(\boldsymbol{Y})$. We can use this fact to first recover the rows of $\boldsymbol{X}_0$, and subsequently recover $\boldsymbol{A}_0$ by solving a system of linear equations. In fact, for $\boldsymbol{X}_0 \sim_{i.i.d.} \text{BG}(\theta)$, rows of $\boldsymbol{X}_0$ are the n *sparsest* vectors (directions) in $\text{row}(\boldsymbol{Y})$ w.h.p. whenever $p \geq \Omega(n \log n)$ [SWW12]. Thus one might try to recover rows of $\boldsymbol{X}_0$ by solving

$$\text{minimize } \|\boldsymbol{q}^*\boldsymbol{Y}\|_0 \text{ subject to } \boldsymbol{q} \neq \boldsymbol{0}. \tag{1.3}$$

The objective is discontinuous, and the domain is an open set. In particular, the homogeneous constraint is nonconventional and tricky to deal with. Since the recovery is up to scale, one can remove the homogeneity by fixing the scale of $\boldsymbol{q}$. Known relaxations [SWW12, DH14] fix the scale by setting $\|\boldsymbol{q}^*\boldsymbol{Y}\|_\infty = 1$, where $\|\cdot\|_\infty$ is the elementwise ℓ^∞ norm. The optimization problem reduces to a sequence of convex programs, which recover $(\boldsymbol{A}_0, \boldsymbol{X}_0)$ for very sparse $\boldsymbol{X}_0$, but provably break down when columns of $\boldsymbol{X}_0$ has more than $O(\sqrt{n})$ nonzeros, or $\theta \geq \Omega(1/\sqrt{n})$. Inspired by our previous image experiment, we work with a *nonconvex* alternative[10]:

$$\text{minimize } f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}) \doteq \frac{1}{p} \sum_{k=1}^{p} h_\mu(\boldsymbol{q}^*\widehat{\boldsymbol{y}}_k), \text{ subject to } \|\boldsymbol{q}\| = 1, \tag{1.4}$$

where $\widehat{\boldsymbol{Y}} \in \mathbb{R}^{n \times p}$ is a proxy for $\boldsymbol{Y}$ (i.e., after appropriate processing), k indexes columns of $\widehat{\boldsymbol{Y}}$, and $\|\cdot\|$ is the usual ℓ^2 norm for vectors. Here $h_\mu(\cdot)$ is chosen to be a convex smooth approximation to $|\cdot|$, namely,

$$h_\mu(z) = \mu \log\left(\frac{\exp(z/\mu) + \exp(-z/\mu)}{2}\right) = \mu \log \cosh(z/\mu), \tag{1.5}$$

---

[9]Indeed, when θ is small enough such that columns of $\boldsymbol{X}_0$ are predominately 1-sparse, one directly observes scaled versions of the atoms (i.e., columns of $\boldsymbol{X}_0$); when $\boldsymbol{X}_0$ is fully dense corresponding to $\theta = 1$, recovery is never possible as one can easily find another complete $\boldsymbol{A}_0'$ and fully dense $\boldsymbol{X}_0'$ such that $\boldsymbol{Y} = \boldsymbol{A}_0'\boldsymbol{X}_0'$ with $\boldsymbol{A}_0'$ not equivalent to $\boldsymbol{A}_0$.

[10]A similar formulation has been proposed in [ZP01] in the context of blind source separation; see also [QSW14].
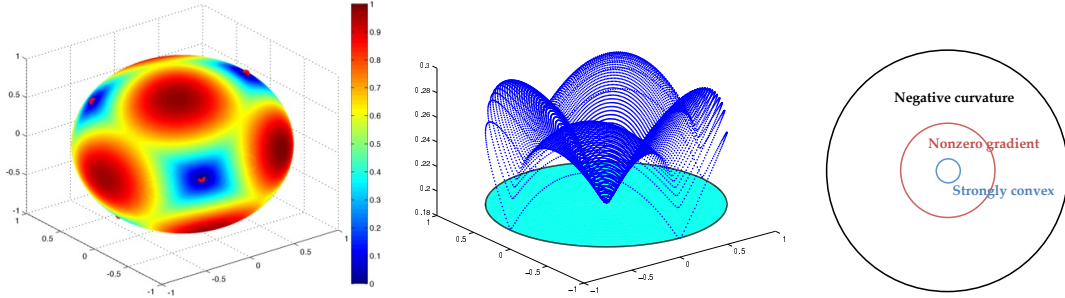
7

Figure 2: Why is dictionary learning over $\mathbb{S}^{n-1}$ tractable? Assume the target dictionary $\boldsymbol{A}_0$ is orthogonal. **Left:** Large sample objective function $\mathbb{E}_{\boldsymbol{X}_0}[f(\boldsymbol{q})]$. The only local minima are the columns of $\boldsymbol{A}_0$ and their negatives. **Center:** the same function, visualized as a height above the plane $\boldsymbol{a}_1^{\perp}$ ($\boldsymbol{a}_1$ is the first column of $\boldsymbol{A}_0$). **Right:** Around the optimum, the function exhibits a small region of positive curvature, a region of large gradient, and finally a region in which the direction away from $\boldsymbol{a}_1$ is a direction of negative curvature.

which is infinitely differentiable and μ controls the smoothing level.[11] The spherical constraint is nonconvex. Hence, a-priori, it is unclear whether (1.4) admits efficient algorithms that attain global optima. Surprisingly, simple descent algorithms for (1.4) exhibit very striking behavior: on many practical numerical examples[12], they appear to produce global solutions. Our next section will uncover interesting geometrical structures underlying the phenomenon.

1.4.2 A Glimpse into High-dimensional Function Landscape

For the moment, suppose $\boldsymbol{A}_0$ is orthogonal, and take $\widehat{\boldsymbol{Y}} = \boldsymbol{Y} = \boldsymbol{A}_0 \boldsymbol{X}_0$ in (1.4). Figure 2 (left) plots $\mathbb{E}_{\boldsymbol{X}_0}[f(\boldsymbol{q};\boldsymbol{Y})]$ over $\boldsymbol{q} \in \mathbb{S}^2$ ($n = 3$). Remarkably, $\mathbb{E}_{\boldsymbol{X}_0}[f(\boldsymbol{q};\boldsymbol{Y})]$ has no spurious local minima. In fact, every local minimizer $\widehat{\boldsymbol{q}}$ produces a row of $\boldsymbol{X}_0$: $\widehat{\boldsymbol{q}}^* \boldsymbol{Y} = \alpha \boldsymbol{e}_i^* \boldsymbol{X}_0$ for some $\alpha \neq 0$.

To better illustrate the point, we take the particular case $\boldsymbol{A}_0 = \boldsymbol{I}$ and project the upper hemisphere above the equatorial plane $\boldsymbol{e}_3^{\perp}$ onto $\boldsymbol{e}_3^{\perp}$. The projection is bijective and we equivalently define a reparameterization $g : \boldsymbol{e}_3^{\perp} \mapsto \mathbb{R}$ of f. Figure 2 (center) plots the graph of g. Obviously the only local minimizers are $\boldsymbol{0}, \pm\boldsymbol{e}_1, \pm\boldsymbol{e}_2$, and they are also global minimizers. Moreover, the apparent nonconvex landscape has interesting structures around $\boldsymbol{0}$: when moving away from $\boldsymbol{0}$, one sees successively a strongly convex region, a nonzero gradient region, and a region where at each point one can always find a direction of negative curvature, as shown schematically in Figure 2 (right). This geometry implies that at any nonoptimal point, there is always at least one direction of descent. Thus, any algorithm that can take advantage of the descent directions will likely converge to one global minimizer, irrespective of initialization.

Two challenges stand out when implementing this idea. For geometry, one has to show similar structure exists for general complete $\boldsymbol{A}_0$, in high dimensions ($n \geq 3$), when the number of observations p is finite (vs. the expectation in the experiment). For algorithms, we need to be able to

---

[11]In fact, there is nothing special about this choice and we believe that any valid smooth (twice continuously differentiable) approximation to $|\cdot|$ would work and yield qualitatively similar results. We also have some preliminary results showing the latter geometric picture remains the same for certain nonsmooth functions, such as a modified version of the Huber function, though the analysis involves handling a different set of technical subtleties. The algorithm also needs additional modifications.

[12]... not restricted to the model we assume here for $\boldsymbol{A}_0$ and $\boldsymbol{X}_0$.

take advantage of this structure without knowing A_0 ahead of time. In Section 1.4.3, we describe a Riemannian trust region method which addresses the latter challenge.

Geometry for orthogonal A_0. In this case, we take $\widehat{Y} = Y = A_0 X_0$. Since $f(q; A_0 X_0) = f(A_0^* q; X_0)$, the landscape of $f(q; A_0 X_0)$ is simply a rotated version of that of $f(q; X_0)$, i.e., when $A_0 = I$. Hence we will focus on the case when $A_0 = I$. Among the $2n$ symmetric sections of $\mathbb{S}^{n-1}$ centered around the signed basis vectors $\pm e_1, \ldots, \pm e_n$, we work with the symmetric section around e_n as an example. The result will carry over to all sections with the same argument; together this provides a complete characterization of the function $f(q; X_0)$ over $\mathbb{S}^{n-1}$.

We again invoke the projection trick described above, this time onto the equatorial plane $e_n^{\perp}$. This can be formally captured by the reparameterization mapping:

$$q(w) = \left(w, \sqrt{1 - \|w\|^2}\right), \ w \in \mathbb{B}^{n-1}, \tag{1.6}$$

where w is the new variable in $e_n^{\perp} \cap \mathbb{B}^{n-1}$ and $\mathbb{B}^{n-1}$ is the unit ball in $\mathbb{R}^{n-1}$. We first study the composition $g(w; X_0) \doteq f(q(w); X_0)$ over the set

$$\Gamma \doteq \left\{w : \|w\| < \sqrt{\tfrac{4n-1}{4n}}\right\}. \tag{1.7}$$

It can be verified the section we chose to work with is contained in this set[13].

Our analysis characterizes the properties of $g(w; X_0)$ by studying three quantities

$$\nabla^2 g(w; X_0), \quad \frac{w^* \nabla g(w; X_0)}{\|w\|}, \quad \frac{w^* \nabla^2 g(w; X_0) w}{\|w\|^2}$$

respectively over three consecutive regions moving away from the origin, corresponding to the three regions in Figure 2 (right). In particular, through typical expectation-concentration style argument, we show that there exists a positive constant c such that

$$\nabla^2 g(w; X_0) \succeq \frac{1}{\mu} c\theta I, \quad \frac{w^* \nabla g(w; X_0)}{\|w\|} \geq c\theta, \quad \frac{w^* \nabla^2 g(w; X_0) w}{\|w\|^2} \leq -c\theta \tag{1.8}$$

over the respective regions w.h.p., confirming our low-dimensional observations described above. In particular, the favorable structure we observed for $n = 3$ persists in high dimensions, w.h.p., even when p is large *yet finite*, for the case A_0 is orthogonal. Moreover, the local minimizer of $g(w; X_0)$ over Γ is very close to $\mathbf{0}$, within a distance of $O(\mu)$.

Geometry for complete A_0. For general complete dictionaries A_0, we hope that the function f retains the nice geometric structure discussed above. We can ensure this by "preconditioning" Y such that the output looks as if being generated from a certain orthogonal matrix, possibly plus

---

[13]Indeed, if $\langle q, e_n \rangle \geq |\langle q, e_i \rangle|$ for any $i \neq n$, $1 - \|w\|^2 = q_n^2 \geq 1/n$, implying $\|w\|^2 \leq \frac{n-1}{n} < \frac{4n-1}{4n}$. The reason we have defined an open set instead of a closed (compact) one is to avoid potential trivial local minimizers located on the boundary.

a small perturbation. We can then argue that the perturbation does not significantly affect the properties of the graph of the objective function. Write

$$\overline{\boldsymbol{Y}} = \left(\tfrac{1}{p\theta}\boldsymbol{Y}\boldsymbol{Y}^*\right)^{-1/2}\boldsymbol{Y}. \tag{1.9}$$

Note that for $\boldsymbol{X}_0 \sim_{i.i.d.} \mathrm{BG}\,(\theta)$, $\mathbb{E}\left[\boldsymbol{X}_0\boldsymbol{X}_0^*\right]/(p\theta) = \boldsymbol{I}$. Thus, one expects $\tfrac{1}{p\theta}\boldsymbol{Y}\boldsymbol{Y}^* = \tfrac{1}{p\theta}\boldsymbol{A}_0\boldsymbol{X}_0\boldsymbol{X}_0^*\boldsymbol{A}_0^*$ to behave roughly like $\boldsymbol{A}_0\boldsymbol{A}_0^*$ and hence $\overline{\boldsymbol{Y}}$ to behave like

$$\left(\boldsymbol{A}_0\boldsymbol{A}_0^*\right)^{-1/2}\boldsymbol{A}_0\boldsymbol{X}_0 = \boldsymbol{U}\boldsymbol{V}^*\boldsymbol{X}_0 \tag{1.10}$$

where we write the SVD of $\boldsymbol{A}_0$ as $\boldsymbol{A}_0 = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$. It is easy to see $\boldsymbol{U}\boldsymbol{V}^*$ is an orthogonal matrix. Hence the preconditioning scheme we have introduced is technically sound.

Our analysis shows that $\overline{\boldsymbol{Y}}$ can be written as

$$\overline{\boldsymbol{Y}} = \boldsymbol{U}\boldsymbol{V}^*\boldsymbol{X}_0 + \boldsymbol{\Xi}\boldsymbol{X}_0, \tag{1.11}$$

where $\boldsymbol{\Xi}$ is a matrix with small magnitude. Simple perturbation argument shows that the constant c in (1.8) is at most shrunk to $c/2$ for all $\boldsymbol{w}$ when p is sufficiently large. Thus, the qualitative aspects of the geometry have not been changed by the perturbation.

1.4.3 A Second-order Algorithm on Manifold: Riemannian Trust Region Method

We do not know $\boldsymbol{A}_0$ ahead of time, so our algorithm needs to take advantage of the structure described above without knowledge of $\boldsymbol{A}_0$. Intuitively, this seems possible as the descent direction in the $\boldsymbol{w}$ space appears to also be a local descent direction for f over the sphere. Another issue is that although the optimization problem has no spurious local minima, it does have many saddle points (Figure 2). We can use second-order information to guarantee to escape saddle points. We derive an algorithm based on the Riemannian trust region method (TRM) [ABG07, AMS09] over the sphere for this purpose.

For a function $f : \mathbb{R}^n \to \mathbb{R}$ and an unconstrained optimization problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f\,(\boldsymbol{x})\,,$$

typical (second-order) TRM proceeds by successively forming second-order approximations to f at the current iterate,

$$\widehat{f}\left(\boldsymbol{\delta}; \boldsymbol{x}^{(k-1)}\right) \doteq f\left(\boldsymbol{x}^{(k-1)}\right) + \nabla^* f\left(\boldsymbol{x}^{(k-1)}\right)\boldsymbol{\delta} + \tfrac{1}{2}\boldsymbol{\delta}^*\boldsymbol{Q}\left(\boldsymbol{x}^{(k-1)}\right)\boldsymbol{\delta}, \tag{1.12}$$

where $\boldsymbol{Q}\left(\boldsymbol{x}^{(k-1)}\right)$ is a proxy for the Hessian matrix $\nabla^2 f\left(\boldsymbol{x}^{(k-1)}\right)$, which encodes the second-order geometry. The next movement direction is determined by seeking a minimum of $\widehat{f}\left(\boldsymbol{\delta}; \boldsymbol{x}^{(k-1)}\right)$ over a small region, normally a norm ball $\|\boldsymbol{\delta}\|_p \leq \Delta$, called the trust region, inducing the well studied trust-region subproblem:

$$\boldsymbol{\delta}^{(k)} \doteq \operatorname*{arg\,min}_{\boldsymbol{\delta}\in\mathbb{R}^n, \|\boldsymbol{\delta}\|_p \leq \Delta} \widehat{f}\left(\boldsymbol{\delta}; \boldsymbol{x}^{(k-1)}\right), \tag{1.13}$$

where Δ is called the trust-region radius that controls how far the movement can be made. A ratio

$$\rho_k \doteq \frac{f\left(\boldsymbol{x}^{(k-1)}\right) - f\left(\boldsymbol{x}^{(k-1)} + \boldsymbol{\delta}^{(k)}\right)}{\widehat{f}\left(\boldsymbol{0}\right) - \widehat{f}\left(\boldsymbol{\delta}^{(k-1)}\right)} \tag{1.14}$$

is defined to measure the progress and typically the radius Δ is updated dynamically according to ρ_k to adapt to the local function behavior. Detailed introductions to the classical TRM can be found in the texts [CGT00a, NW06].
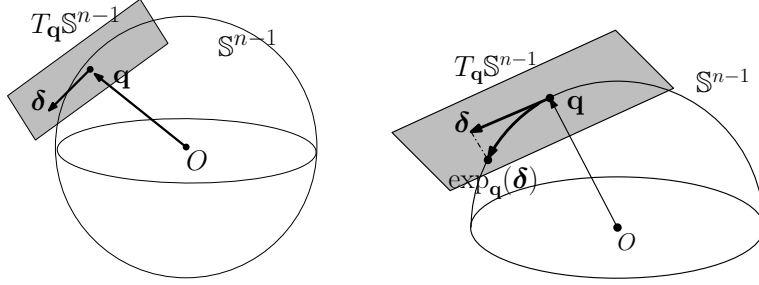
Figure 3: Illustrations of the tangent space $T_{\boldsymbol{q}}\mathbb{S}^{n-1}$ and exponential map $\exp_{\boldsymbol{q}}\left(\boldsymbol{\delta}\right)$ defined on the sphere $\mathbb{S}^{n-1}$.

To generalize the idea to smooth manifolds, one natural choice is to form the approximation over the tangent spaces [ABG07, AMS09]. Specific to our spherical manifold, for which the tangent space at an iterate $\boldsymbol{q}^{(k)} \in \mathbb{S}^{n-1}$ is $T_{\boldsymbol{q}^{(k)}}\mathbb{S}^{n-1} \doteq \left\{\boldsymbol{v} : \boldsymbol{v}^*\boldsymbol{q}^{(k)} = 0\right\}$ (see Figure 3), we work with a "quadratic" approximation $\widehat{f} : T_{\boldsymbol{q}^{(k)}}\mathbb{S}^{n-1} \mapsto \mathbb{R}$ defined as

$$\widehat{f}(\boldsymbol{\delta}; \boldsymbol{q}^{(k)}) \doteq f(\boldsymbol{q}^{(k)}) + \left\langle \nabla f(\boldsymbol{q}^{(k)}), \boldsymbol{\delta} \right\rangle + \frac{1}{2}\boldsymbol{\delta}^* \left(\nabla^2 f(\boldsymbol{q}^{(k)}) - \left\langle \nabla f(\boldsymbol{q}^{(k)}), \boldsymbol{q}^{(k)} \right\rangle \boldsymbol{I}\right) \boldsymbol{\delta}. \tag{1.15}$$

To interpret this approximation, let $\mathcal{P}_{T_{\boldsymbol{q}^{(k)}}\mathbb{S}^{n-1}} \doteq \left(\boldsymbol{I} - \boldsymbol{q}^{(k)}\left(\boldsymbol{q}^{(k)}\right)^*\right)$ be the orthoprojector onto $T_{\boldsymbol{q}^{(k)}}\mathbb{S}^{n-1}$ and write (3.2) into an equivalent form:

$$\widehat{f}(\boldsymbol{\delta}; \boldsymbol{q}^{(k)}) \doteq f(\boldsymbol{q}^{(k)}) + \left\langle \mathcal{P}_{T_{\boldsymbol{q}^{(k)}}\mathbb{S}^{n-1}} \nabla f(\boldsymbol{q}^{(k)}), \boldsymbol{\delta} \right\rangle$$
$$+ \frac{1}{2}\boldsymbol{\delta}^* \mathcal{P}_{T_{\boldsymbol{q}^{(k)}}\mathbb{S}^{n-1}} \left(\nabla^2 f(\boldsymbol{q}^{(k)}) - \left\langle \nabla f(\boldsymbol{q}^{(k)}), \boldsymbol{q}^{(k)} \right\rangle \boldsymbol{I}\right) \mathcal{P}_{T_{\boldsymbol{q}^{(k)}}\mathbb{S}^{n-1}} \boldsymbol{\delta}.$$

The two terms

$$\mathrm{grad} f\left(\boldsymbol{q}^{(k)}\right) \doteq \mathcal{P}_{T_{\boldsymbol{q}^{(k)}}\mathbb{S}^{n-1}} \nabla f(\boldsymbol{q}^{(k)}),$$
$$\mathrm{Hess} f\left(\boldsymbol{q}^{(k)}\right) \doteq \mathcal{P}_{T_{\boldsymbol{q}^{(k)}}\mathbb{S}^{n-1}} \left(\nabla^2 f(\boldsymbol{q}^{(k)}) - \left\langle \nabla f(\boldsymbol{q}^{(k)}), \boldsymbol{q}^{(k)} \right\rangle \boldsymbol{I}\right) \mathcal{P}_{T_{\boldsymbol{q}^{(k)}}\mathbb{S}^{n-1}}$$

are the Riemannian gradient and Riemannian Hessian of f w.r.t. $\mathbb{S}^{n-1}$, respectively [ABG07, AMS09]; the above approximation is reminiscent of the usual quadratic approximation described in (1.12).

Then the Riemannian trust-region subproblem is

$$\min_{\boldsymbol{\delta} \in T_{\boldsymbol{q}^{(k)}}\mathbb{S}^{n-1},\, \|\boldsymbol{\delta}\| \leq \Delta} \widehat{f}\left(\boldsymbol{\delta}; \boldsymbol{q}^{(k)}\right), \tag{1.16}$$

11

where we take the simple ℓ^2 norm ball for the trust region. This can be transformed into a classical trust region subprolem: indeed, taking any orthonormal basis $\boldsymbol{U}_{\boldsymbol{q}^{(k)}}$ for $T_{\boldsymbol{q}^{(k)}}\mathbb{S}^{n-1}$, the above problem is equivalent to

$$\min_{\|\boldsymbol{\xi}\| \leq \Delta} \widehat{f}\left(\boldsymbol{U}_{\boldsymbol{q}^{(k)}}\boldsymbol{\xi}, \boldsymbol{q}^{(k)}\right), \tag{1.17}$$

where the objective is quadratic in $\boldsymbol{\xi}$. This is the classical trust region problem (with ℓ^2 norm ball constraint) that admits very efficient numerical algorithms [MS83, HK14]. Once we obtain the minimizer $\boldsymbol{\xi}_\star$, we set $\boldsymbol{\delta}_\star = \boldsymbol{U}\boldsymbol{\xi}_\star$, which solves (1.16).

One additional issue as compared to the Euclidean setting is that now $\boldsymbol{\delta}_\star$ is one vector in the tangent space and additive update leads to a point outside the sphere. We resort to the natural exponential map to pull the tangent vector to a point on the sphere:

$$\boldsymbol{q}^{(k+1)} \doteq \exp_{\boldsymbol{q}^{(k)}}\left(\boldsymbol{\delta}_\star\right) = \boldsymbol{q}^{(k)}\cos\|\boldsymbol{\delta}_\star\| + \tfrac{\boldsymbol{\delta}_\star}{\|\boldsymbol{\delta}_\star\|}\sin\|\boldsymbol{\delta}_\star\|. \tag{1.18}$$

As seen from Figure 3, the movement to the next iterate is "along the direction"[14] of $\boldsymbol{\delta}_\star$ while staying over the sphere.

Using the above geometric characterizations, we prove that w.h.p., the algorithm converges to a local minimizer when the parameter Δ is sufficiently small[15]. In particular, we show that (1) the trust region step induces at least a fixed amount of decrease to the objective value in the negative curvature and nonzero gradient region; (2) the trust region iterate sequence will eventually move to and stay in the strongly convex region, and converge to the local minimizer contained in the region with an asymptotic quadratic rate. In short, the geometric structure implies that from *any initialization*, the iterate sequence converges to a close approximation to the target solution in a polynomial number of steps.

1.5 Prior Arts and Connections

It is far too ambitious to include here a comprehensive review of the exciting developments of DL algorithms and applications after the pioneer work [OF96]. We refer the reader to Chapter 12 - 15 of the book [Ela10] and the survey paper [MBP14] for summaries of relevant developments in image analysis and visual recognition. In the following, we focus on reviewing recent developments on the theoretical side of dictionary learning, and draw connections to problems and techniques that are relevant to the current work.

Theoretical Dictionary Learning. The theoretical study of DL in the recovery setting started only very recently. [AEB06] was the first to provide an algorithmic procedure to correctly extract the generating dictionary. The algorithm requires exponentially many samples and has exponential running time; see also [HS11]. Subsequent work [GS10, GW11, Sch14a, Sch14b, Sch15] studied when the target dictionary is a local optimum of natural recovery criteria. These meticulous analyses show that polynomially many samples are sufficient to ensure local correctness under natural assumptions. However, these results do not imply that one can design efficient algorithms to obtain the desired local optimum and hence the dictionary.

---

[14]Technically, moving along the geodesic whose velocity at time zero is $\boldsymbol{\delta}_\star$.

[15]For simplicity of analysis, we have assumed Δ is fixed throughout the analysis. In practice, dynamic updates to Δ lead to faster convergence.

[SWW12] initiated the on-going research effort to provide efficient algorithms that globally solve DR. They showed that one can recover a complete dictionary A_0 from $Y = A_0 X_0$ by solving a certain sequence of linear programs, when X_0 is a sparse random matrix with $O(\sqrt{n})$ nonzeros per column. [AAJ+13, AAN13] and [AGM13, AGMM15] give efficient algorithms that provably recover overcomplete ($m \geq n$) and incoherent dictionaries, based on a combination of {clustering or spectral initialization} and local refinement. These algorithms again succeed when X_0 has $\widetilde{O}(\sqrt{n})$ [16] nonzeros per column. Recent work, including [ABGM14] and [BKS14], guarantees recovery with (almost) $O(n)$ nonzeros per column, but runs in super-polynomial (quasipolynomial) time. By comparison, we give the first *polynomial-time* algorithm that provably recovers complete dictionary A_0 when X_0 has $O(n)$ nonzeros per column.

Aside from efficient recovery, other theoretical work on DL includes results on identifiability [AEB06, HS11], generalization bounds [MP10b, VMB11, MG13, GJB+13], and noise stability [GJB14].

Finding Sparse Vectors in a Linear Subspace. We have followed [SWW12] and cast the core problem as finding the sparsest vectors in a given linear subspace, which is also of independent interest. Under a planted sparse model[17], [DH14] shows solving a sequence of linear programs similar to [SWW12] can recover sparse vectors with sparsity up to $O(p/\sqrt{n})$, sublinear in the vector dimension. [QSW14] improved the recovery limit to $O(p)$ by solving a nonconvex spherical constrained problem similar to (1.4)[18] via an ADM algorithm. The idea of seeking rows of X_0 sequentially by solving the above core problem sees precursors in [ZP01] for blind source separation, and [GN10] for matrix sparsification. [ZP01] also proposed a nonconvex optimization similar to (1.4) here and that employed in [QSW14].

Nonconvex Optimization Problems. For other nonconvex optimization problems of recovery of structured signals[19], including low-rank matrix completion/recovery [KMO10, JNS13, Har14, HW14, NNS+14, JN14], phase retreival [NJS13, CLS14], tensor recovery [JO14, AGJ14], mixed regression [YCS13, LWB13], structured element pursuit [QSW14], and recovery of simultaneously structured signals [LWB13], the initialization plus local refinement strategy adopted in theoretical DL [AAJ+13, AAN13, AGM13, AGMM15, ABGM14] is also crucial: nearness to the target solution enables exploiting the local geometry of the target to analyze the local refinement.[20] By comparison, we provide a complete characterization of the global geometry, which admits efficient algorithms without any special initialization. The idea of separating the geometric analysis and algorithmic design may also prove valuable for other nonconvex problems discussed above.

Optimization over Riemannian Manifolds. Our trust-region algorithm on the sphere builds on the extensive research efforts to generalize Euclidean numerical algorithms to (Riemannian) manifold settings. We refer the reader to the monographs [Udr94, HMG94, AMS09] for survey

---

[16]The $\widetilde{O}$ suppresses some logarithm factors.

[17]... where one sparse vector embedded in an otherwise random subspace.

[18]The only difference is that they chose to work with the Huber function as a proxy of the $\|\cdot\|_1$ function.

[19]This is a body of recent work studying nonconvex recovery up to statistical precision, including, e.g., [LW11, LW13, WLL14, BWY14, WGNL14, LW14, Loh15, SLLC15].

[20]The powerful framework [ABRS10, BST14] to establish local convergence of ADM algorithms to critical points applies to DL/DR also, see, e.g., [BJQS14, BQJ14, BJS14]. However, these results do not guarantee to produce global optima.

of developments in this field. In particular, [EAS98] developed Newton and conjugate-gradient methods for the Stiefel manifolds, of which the spherical manifold is a special case. [ABG07] generalized the trust-region methods to Riemannian manifolds. We cannot, however, adopt the existing convergence results that concern either global convergence (convergence to critical points) or local convergence (convergence to a local minimum within a radius). The particular geometric structure forces us to piece together different arguments to obtain the global result.

(a) Correlated Gaussian, $\theta = 0.1$ **(b)** Correlated Uniform, $\theta = 0.1$ **(c)** Independent Uniform, $\theta = 0.1$

(d) Correlated Gaussian, $\theta = 0.9$ **(e)** Correlated Uniform, $\theta = 0.9$ **(f)** Independent Uniform, $\theta = 1$
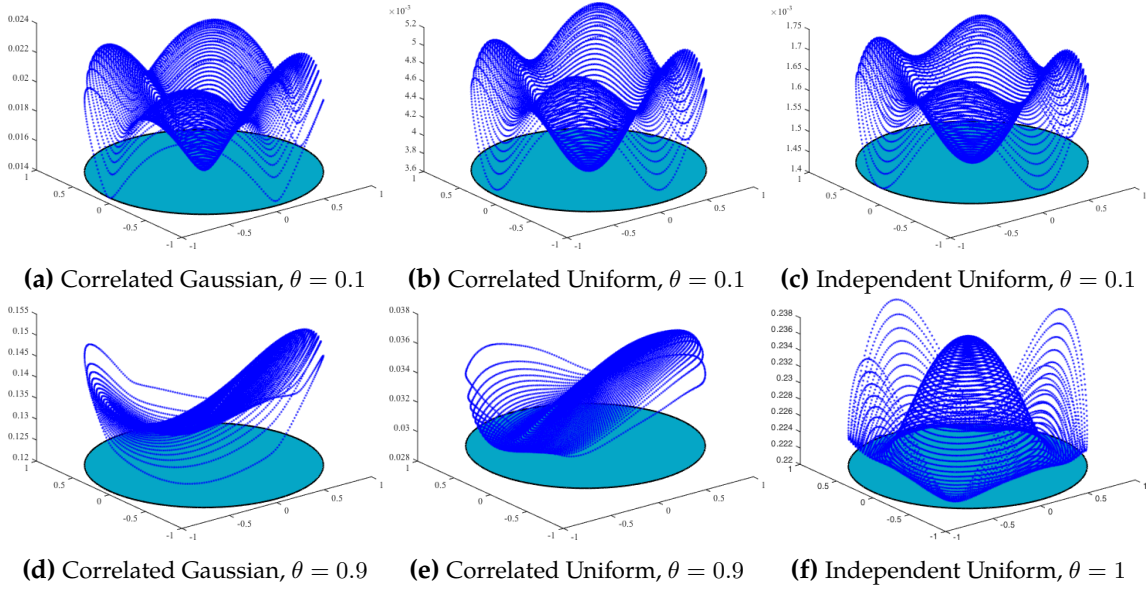
Figure 4: Asymptotic function landscapes when rows of X_0 are not independent. W.l.o.g., we again assume $A_0 = I$. In (a) and (d), $X_0 = \Omega \odot V$, with $\Omega \sim_{i.i.d.} \mathrm{Ber}(\theta)$ and columns of X_0 i.i.d. Gaussian vectors obeying $v_i \sim \mathcal{N}(0, \Sigma^2)$ for symmetric Σ with 1's on the diagonal and i.i.d. off-diagonal entries distributed as $\mathcal{N}(0, \sqrt{2}/20)$. Similarly, in (b) and (e), $X_0 = \Omega \odot W$, with $\Omega \sim_{i.i.d.} \mathrm{Ber}(\theta)$ and columns of X_0 i.i.d. vectors generated as $w_i = \Sigma u^i$ with $u_i \sim_{i.i.d.} \mathrm{Uniform}[-0.5, 0.5]$. For comparison, in (c) and (f), $X_0 = \Omega \odot W$ with $\Omega \sim_{i.i.d.} \mathrm{Ber}(\theta)$ and $W \sim_{i.i.d.} \mathrm{Uniform}[-0.5, 0.5]$. Here $\odot$ denote the elementwise product, and the objective function is still based on the $\log \cosh$ function as in (1.4).

Independent Component Analysis (ICA) and Other Matrix Factorization Problems. DL can also be considered in the general framework of matrix factorization problems, which encompass the classic principal component analysis (PCA), ICA, and clustering, and more recent problems such as nonnegative matrix factorization (NMF), multi-layer neural nets (deep learning architectures). Most of these problems are NP-hard. Identifying tractable cases of practical interest and providing provable efficient algorithms are subject of on-going research endeavors; see, e.g., recent progresses on NMF [AGKM12], and learning deep neural nets [ABGM13, SA14, NP13, LSSS14].

ICA factors a data matrix Y as $Y = AX$ such that A is square and rows of X are as independent as possible [HO00, HO01]. In theoretical study of the recovery problem, it is often assumed that rows of X_0 are (weakly) independent (see, e.g., [Com94, FJK96, AGMS12]). Our i.i.d. probability model on X_0 implies rows of X_0 are independent, aligning our problem perfectly with the ICA problem. More interestingly, the $\log \cosh$ objective we analyze here was proposed as a general-purpose *contrast function* in ICA that has not been thoroughly analyzed [Hyv99], and algorithm and

analysis with another popular contrast function, the fourth-order cumulants, indeed overlap with ours considerably [FJK96, AGMS12][21]. While this interesting connection potentially helps port our analysis to ICA, it is a fundamental question to ask what is playing the vital role for DR, sparsity or independence.

Figure 4 helps shed some light in this direction, where we again plot the asymptotic objective landscape with the natural reparameterization as in Section 1.4.2. From the left and central panels, it is evident even without independence, X_0 with sparse columns induces the familiar geometric structures we saw in Figure 2; such structures are broken when the sparsity level becomes large. We believe all our later analyses can be generalized to the correlated cases we experimented with. On the other hand, from the right panel[22], it seems with independence, the function landscape undergoes a transition as sparsity level grows - target solution goes from minimizers of the objective to the maximizers of the objective. Without adequate knowledge of the true sparsity, it is unclear whether one would like to minimize or maximize the objective.[23] This suggests sparsity, instead of independence, makes our current algorithm for DR work.

1.6 Notations, Organization, 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, $\mathbb{S}^{n-1}$ for the unit sphere in $\mathbb{R}^n$, $\mathbb{B}^n$ for the unit ball in $\mathbb{R}^n$, and $[m] \doteq \{1, \ldots, m\}$ for positive integers m, n, k. 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 $(x^i)^*$. For norms, $\|\cdot\|$ is the usual ℓ^2 norm for a vector and to the operator norm (i.e., $\ell^2 \to \ell^2$) for a matrix; all other norms will be indexed by subscript, for example the Frobenius norm $\|\cdot\|_F$ for matrices and the element-wise max-norm $\|\cdot\|_\infty$. We use $x \sim \mathcal{L}$ to mean that the random variable x is distributed according to the law $\mathcal{L}$. Let $\mathcal{N}$ denote the Gaussian law. Then $x \sim \mathcal{N}(0, I)$ means that x is a standard Gaussian vector. Similarly, we use $x \sim_{i.i.d.} \mathcal{L}$ to mean elements of x are independently and identically distributed according to the law $\mathcal{L}$. So the fact $x \sim \mathcal{N}(0, I)$ is equivalent to that $x \sim_{i.i.d.} \mathcal{N}(0, 1)$. One particular distribution of interest for this paper is the Bernoulli-Gaussian with rate θ: $Z \sim B \cdot G$, with $G \sim \mathcal{N}(0, 1)$ and $B \sim \text{Ber}(\theta)$. We also write this compactly as $Z \sim \text{BG}(\theta)$. We frequently use indexed C and c for numerical constants when stating and proving technical results. The scopes of such constants are local unless otherwise noted. We use standard notations for most other cases, with exceptions clarified locally.

The rest of the paper is organized as follows. In Section 2 we present major technical results for a complete characterization of the geometry sketched in Section 1.4.2. Similarly in Section 3 we present necessary technical machinery and results for convergence proof of the Riemannian

---

[21]Nevertheless, the objective functions are apparently different. Moreover, we have provided a complete geometric characterization of the objective, in contrast to [FJK96, AGMS12]. We believe the geometric characterization could not only provide insight to the algorithm, but also help improve the algorithm in terms of stability and also finding all components.

[22]We have not showed the results on the BG model here, as it seems the structure persists even when θ approaches 1. We suspect the "phase transition" of the landscape occurs at different points for different distributions and Gaussian is the outlying case where the transition occurs at 1.

[23]For solving the ICA problem, this suggests the $\log \cosh$ contrast function, that works well empirically [Hyv99], may not work for all distributions (rotation-invariant Gaussian excluded of course).

trust-region algorithm over the sphere, corresponding to Section 1.4.3. In Section 4, we discuss the whole algorithmic pipeline for recovering complete dictionaries given $\boldsymbol{Y}$, and present the main theorems. After presenting a simple simulation to corroborate our theory in Section 5, we wrap up the main content in Section 6 by discussing possible improvement and future directions after this work. All major proofs of geometrical and algorithmic results are deferred to Section 7 and Section 8, respectively. Section 9 augments the main results. The appendices cover some recurring technical tools and auxiliary results for the proofs.

The codes to reproduce all the figures and experimental results can be found online:

https://github.com/sunju/dl_focm

2 High-dimensional Function Landscapes

To characterize the function landscape of $f(\boldsymbol{q}; \boldsymbol{X}_0)$ over $\mathbb{S}^{n-1}$, we mostly work with the function

$$g(\boldsymbol{w}) \doteq f(\boldsymbol{q}(\boldsymbol{w}); \boldsymbol{X}_0) = \frac{1}{p} \sum_{k=1}^{p} h_\mu(\boldsymbol{q}(\boldsymbol{w})^*(\boldsymbol{x}_0)_k), \tag{2.1}$$

induced by the reparametrization

$$\boldsymbol{q}(\boldsymbol{w}) = \left(\boldsymbol{w}, \sqrt{1 - \|\boldsymbol{w}\|^2}\right), \quad \boldsymbol{w} \in \mathbb{B}^{n-1}. \tag{2.2}$$

In particular, we focus our attention to the smaller set

$$\Gamma = \left\{\boldsymbol{w} : \|\boldsymbol{w}\| < \sqrt{\frac{4n-1}{4n}}\right\}, \tag{2.3}$$

because $\boldsymbol{q}(\Gamma)$ contains all points $\boldsymbol{q} \in \mathbb{S}^{n-1}$ with $n \in \arg\max_{i \in \pm[n]} \boldsymbol{q}^* \boldsymbol{e}_i$ and we can characterize other parts of f on $\mathbb{S}^{n-1}$ using projection onto other equatorial planes. Note that over Γ, $q_n = \left(1 - \|\boldsymbol{w}\|^2\right)^{1/2} \geq \frac{1}{2\sqrt{n}}$.

2.1 Main Geometric Theorems

Theorem 2.1 (High-dimensional landscape - orthogonal dictionary) *Suppose $\boldsymbol{A}_0 = \boldsymbol{I}$ and hence $\boldsymbol{Y} = \boldsymbol{A}_0 \boldsymbol{X}_0 = \boldsymbol{X}_0$. There exist positive constants $c_\star$ and C, such that for any $\theta \in (0, 1/2)$ and $\mu < \min\{c_a \theta n^{-1}, c_b n^{-5/4}\}$, whenever $p \geq \frac{C}{\mu^2 \theta^2} n^3 \log \frac{n}{\mu\theta}$, the following hold simultaneously with high probability:*

$$\nabla^2 g(\boldsymbol{w}; \boldsymbol{X}_0) \succeq \frac{c_\star \theta}{\mu} \boldsymbol{I} \qquad \forall \boldsymbol{w} \quad s.t. \quad \|\boldsymbol{w}\| \leq \frac{\mu}{4\sqrt{2}}, \tag{2.4}$$

$$\frac{\boldsymbol{w}^* \nabla g(\boldsymbol{w}; \boldsymbol{X}_0)}{\|\boldsymbol{w}\|} \geq c_\star \theta \qquad \forall \boldsymbol{w} \quad s.t. \quad \frac{\mu}{4\sqrt{2}} \leq \|\boldsymbol{w}\| \leq \frac{1}{20\sqrt{5}} \tag{2.5}$$

$$\frac{\boldsymbol{w}^* \nabla^2 g(\boldsymbol{w}; \boldsymbol{X}_0) \boldsymbol{w}}{\|\boldsymbol{w}\|^2} \leq -c_\star \theta \qquad \forall \boldsymbol{w} \quad s.t. \quad \frac{1}{20\sqrt{5}} \leq \|\boldsymbol{w}\| \leq \sqrt{\frac{4n-1}{4n}}, \tag{2.6}$$

16

and *the function $g(\boldsymbol{w}; \boldsymbol{X}_0)$ has exactly one local minimizer $\boldsymbol{w}_\star$ over the open set $\Gamma \doteq \left\{ \boldsymbol{w} : \|\boldsymbol{w}\| < \sqrt{\frac{4n-1}{4n}} \right\}$,* *which satisfies*

$$\|\boldsymbol{w}_\star - \boldsymbol{0}\| \;\leq\; \min\left\{ \frac{c_c \mu}{\theta} \sqrt{\frac{n \log p}{p}}, \frac{\mu}{16} \right\}. \tag{2.7}$$

In particular, with this choice of p, the probability the claim fails to hold is at most $4np^{-10} + \theta(np)^{-7} + \exp(-0.3\theta np) + c_d \exp\left(-c_e p \mu^2 \theta^2 / n^2\right)$. Here c_a to c_e are all positive numerical constants.

Here $\boldsymbol{q}(\boldsymbol{0}) = \boldsymbol{e}_n$, which exactly recovers the last row of $\boldsymbol{X}_0$, $\boldsymbol{x}_0^n$. Though the unique local minimizer $\boldsymbol{w}_\star$ may not be $\boldsymbol{0}$, it is very near to $\boldsymbol{0}$. Hence the resulting $\boldsymbol{q}(\boldsymbol{w}_\star)$ produces a close approximation to $\boldsymbol{x}_0^n$. Note that $\boldsymbol{q}(\Gamma)$ (strictly) contains all points $\boldsymbol{q} \in \mathbb{S}^{n-1}$ such that $n = \arg\max_{i \in \pm[n]} \boldsymbol{q}^* \boldsymbol{e}_i$. We can characterize the graph of the function $f(\boldsymbol{q}; \boldsymbol{X}_0)$ in the vicinity of other signed basis vector $\pm \boldsymbol{e}_i$ simply by changing the plane $\boldsymbol{e}_n^\perp$ to $\boldsymbol{e}_i^\perp$. Doing this $2n$ times (and multiplying the failure probability in Theorem 2.1 by $2n$), we obtain a characterization of $f(\boldsymbol{q}; \boldsymbol{X}_0)$ over the entirety of $\mathbb{S}^{n-1}$.[24] The result is captured by the next corollary.

Corollary 2.2 *Suppose $\boldsymbol{A}_0 = \boldsymbol{I}$ and hence $\boldsymbol{Y} = \boldsymbol{A}_0 \boldsymbol{X}_0 = \boldsymbol{X}_0$. There exist positive constant C, such that for any $\theta \in (0, 1/2)$ and $\mu < \min\left\{ c_a \theta n^{-1}, c_b n^{-5/4} \right\}$, whenever $p \geq \frac{C}{\mu^2 \theta^2} n^3 \log \frac{n}{\mu\theta}$, with probability at least $1 - 8n^2 p^{-10} - \theta(np)^{-7} - \exp(-0.3\theta np) - c_c \exp\left(-c_d p \mu^2 \theta^2 / n^2\right)$, the function $f(\boldsymbol{q}; \boldsymbol{X}_0)$ has exactly $2n$ local minimizers over the sphere $\mathbb{S}^{n-1}$. In particular, there is a bijective map between these minimizers and signed basis vectors $\{\pm \boldsymbol{e}_i\}_i$, such that the corresponding local minimizer $\boldsymbol{q}_\star$ and $\boldsymbol{b} \in \{\pm \boldsymbol{e}_i\}_i$ satisfy*

$$\|\boldsymbol{q}_\star - \boldsymbol{b}\| \leq \sqrt{2} \min\left\{ \frac{c_c \mu}{\theta} \sqrt{\frac{n \log p}{p}}, \frac{\mu}{16} \right\}. \tag{2.8}$$

Here c_a to c_d are numerical constants (possibly different from that in the above theorem).

Proof By Theorem 2.1, over $\boldsymbol{q}(\Gamma)$, $\boldsymbol{q}(\boldsymbol{w}_\star)$ is the unique local minimizer. Suppose not. Then there exist $\boldsymbol{q}' \in \boldsymbol{q}(\Gamma)$ with $\boldsymbol{q}' \neq \boldsymbol{q}(\boldsymbol{w}_\star)$ and $\varepsilon > 0$, such that $f(\boldsymbol{q}'; \boldsymbol{X}_0) \leq f(\boldsymbol{q}; \boldsymbol{X}_0)$ for all $\boldsymbol{q} \in \boldsymbol{q}(\Gamma)$ satisfying $\|\boldsymbol{q}' - \boldsymbol{q}\| < \varepsilon$. Since the mapping $\boldsymbol{w} \mapsto \boldsymbol{q}(\boldsymbol{w})$ is $2\sqrt{n}$-Lipschitz (Lemma 7.7), $g(\boldsymbol{w}(\boldsymbol{q}'); \boldsymbol{X}_0) \leq g(\boldsymbol{w}(\boldsymbol{q}); \boldsymbol{X}_0)$ for all $\boldsymbol{w} \in \Gamma$ satisfying $\|\boldsymbol{w}(\boldsymbol{q}') - \boldsymbol{w}(\boldsymbol{q})\| < \varepsilon/(2\sqrt{n})$, implying $\boldsymbol{w}(\boldsymbol{q}')$ is a local minimizer different from $\boldsymbol{w}_\star$, a contradiction. Let $\|\boldsymbol{w}_\star - \boldsymbol{0}\| = \eta$. Straightforward calculation shows

$$\|\boldsymbol{q}(\boldsymbol{w}_\star) - \boldsymbol{e}_n\|^2 = \left(1 - \sqrt{1 - \eta^2}\right)^2 + \eta^2 = 2 - 2\sqrt{1 - \eta^2} \leq 2\eta^2.$$

Repeating the argument $2n$ times in the vicinity of other signed basis vectors $\pm \boldsymbol{e}_i$ gives $2n$ local minimizers of f. Indeed, the $2n$ symmetric sections cover the sphere with certain overlaps, and a simple calculation shows that no such local minimizer lies in the overlapped regions (due to nearness to a signed basis vector). There is no extra local minimizer, as such local minimizer is contained in at least one of the $2n$ symmetric sections, resulting two different local minimizers in one section, contradicting the uniqueness result we obtained above. ∎

Though the $2n$ isolated local minimizers may have different objective values, they are equally good in the sense any of them produces a close approximation to a certain row of $\boldsymbol{X}_0$. As discussed

---

[24]In fact, it is possible to pull the very detailed geometry captured in (2.4) through (2.6) back to the sphere (i.e., the $\boldsymbol{q}$ space) also; analysis of the Riemannian trust-region algorithm later does part of these. We will stick to this simple global version here.

in Section 1.4.2, for cases A_0 is an orthobasis other than I, the landscape of $f(q;Y)$ is simply a rotated version of the one we characterized above.

Theorem 2.3 (High-dimensional landscape - complete dictionary) *Suppose A_0 is complete with its condition number $\kappa(A_0)$. There exist positive constants $c_\star$ and C, such that for any $\theta \in (0, 1/2)$ and $\mu < \min\left\{c_a \theta n^{-1}, c_b n^{-5/4}\right\}$, when $p \geq \frac{C}{c_\star^2 \theta} \max\left\{\frac{n^4}{\mu^4}, \frac{n^5}{\mu^2}\right\} \kappa^8(A_0) \log^4\left(\frac{\kappa(A_0)n}{\mu\theta}\right)$ and $\overline{Y} \doteq \sqrt{p\theta}\,(YY^*)^{-1/2}\,Y$, $U\Sigma V^* = \mathtt{SVD}(A_0)$, the following hold simultaneously with high probability:*

$$\nabla^2 g(w; VU^*\overline{Y}) \succeq \frac{c_\star \theta}{2\mu} I \qquad \forall\, w \quad s.t. \quad \|w\| \leq \frac{\mu}{4\sqrt{2}}, \tag{2.9}$$

$$\frac{w^* \nabla g(w; VU^*\overline{Y})}{\|w\|} \geq \frac{1}{2} c_\star \theta \qquad \forall\, w \quad s.t. \quad \frac{\mu}{4\sqrt{2}} \leq \|w\| \leq \frac{1}{20\sqrt{5}} \tag{2.10}$$

$$\frac{w^* \nabla^2 g(w; VU^*\overline{Y})w}{\|w\|^2} \leq -\frac{1}{2} c_\star \theta \qquad \forall\, w \quad s.t. \quad \frac{1}{20\sqrt{5}} \leq \|w\| \leq \sqrt{\frac{4n-1}{4n}}, \tag{2.11}$$

and the function $g(w; VU^*\overline{Y})$ has exactly one local minimizer $w_\star$ over the open set $\Gamma \doteq \left\{w : \|w\| < \sqrt{\frac{4n-1}{4n}}\right\}$, which satisfies

$$\|w_\star - 0\| \leq \frac{\mu}{7}. \tag{2.12}$$

In particular, with this choice of p, the probability the claim fails to hold is at most $4np^{-10} + \theta(np)^{-7} + \exp(-0.3\theta np) + p^{-8} + c_d \exp\left(-c_e p\mu^2\theta^2/n^2\right)$. Here c_a to c_e are all positive numerical constants.

Corollary 2.4 *Suppose A_0 is complete with its condition number $\kappa(A_0)$. There exist positive constants $c_\star$ and C, such that for any $\theta \in (0, 1/2)$ and $\mu < \min\left\{c_a \theta n^{-1}, c_b n^{-5/4}\right\}$, when $p \geq \frac{C}{c_\star^2 \theta} \max\left\{\frac{n^4}{\mu^4}, \frac{n^5}{\mu^2}\right\} \kappa^8(A_0) \log^4\left(\frac{\kappa(A_0)n}{\mu\theta}\right)$ and $\overline{Y} \doteq \sqrt{p\theta}\,(YY^*)^{-1/2}\,Y$, $U\Sigma V^* = \mathtt{SVD}(A_0)$, with probability at least $1 - 8n^2 p^{-10} - \theta(np)^{-7} - \exp(-0.3\theta np) - p^{-8} - c_d \exp\left(-c_e p\mu^2\theta^2/n^2\right)$, the function $f(q; VU^*\overline{Y})$ has exactly $2n$ local minimizers over the sphere $\mathbb{S}^{n-1}$. In particular, there is a bijective map between these minimizers and signed basis vectors $\{\pm e_i\}_i$, such that the corresponding local minimizer $q_\star$ and $b \in \{\pm e_i\}_i$ satisfy*

$$\|q_\star - b\| \leq \frac{\sqrt{2}\mu}{7}. \tag{2.13}$$

Here c_a to c_d are numerical constants (possibly different from that in the above theorem).

We will omit the proof as it is almost identical to that of corollary 2.2.

2.2 Useful Technical Lemmas and Proof Ideas for Orthogonal Dictionaries

The proof of Theorem 2.1 is conceptually straightforward: one shows that $\mathbb{E}_{X_0}[g(w; X_0)]$ has the claimed properties, and then proves that each of the quantities of interest concentrates uniformly about its expectation. The detailed calculations are nontrivial.

The next three propositions show that in the expected function landscape, we see successively strongly convex region, nonzero gradient region, and directional negative curvature region when moving away from zero, as depicted in Figure 2 and sketched in Section 1.4.2. Note that in this case

$$\mathbb{E}_{X_0}[g(q; X_0)] = \mathbb{E}_{x \sim_{i.i.d.} \mathrm{BG}(\theta)}[h_\mu(q(w)^* x)].$$

Proposition 2.5 *There exists a positive constant c, such that for every $\theta \in \left(0, \frac{1}{2}\right)$ and any $R_h \in \left(0, \sqrt{\frac{4n-1}{4n}}\right)$, if $\mu \leq c \min \left\{\theta R_h^2 n^{-1}, R_h n^{-5/4}\right\}$, it holds for every $\boldsymbol{w}$ satisfying $R_h \leq \|\boldsymbol{w}\| \leq \sqrt{\frac{4n-1}{4n}}$ that*

$$\frac{\boldsymbol{w}^* \nabla_{\boldsymbol{w}}^2 \mathbb{E}\left[h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right]\boldsymbol{w}}{\|\boldsymbol{w}\|^2} \leq -\frac{\theta}{2\sqrt{2\pi}}.$$

Proof See Section 7.1.1 on Page 47. ∎

Proposition 2.6 *For every $\theta \in \left(0, \frac{1}{2}\right)$ and every $\mu \leq 9/50$, it holds for every $\boldsymbol{w}$ satisfying $r_g \leq \|\boldsymbol{w}\| \leq R_g$, where $r_g = \frac{\mu}{6\sqrt{2}}$ and $R_g = \frac{1-\theta}{10\sqrt{5}}$, that*

$$\frac{\boldsymbol{w}^* \nabla_{\boldsymbol{w}} \mathbb{E}\left[h_\mu(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x})\right]}{\|\boldsymbol{w}\|} \geq \frac{\theta}{20\sqrt{2\pi}}.$$

Proof See Section 7.1.2 on Page 52. ∎

Proposition 2.7 *For every $\theta \in \left(0, \frac{1}{2}\right)$, and every $\mu \leq \frac{1}{20\sqrt{n}}$, it holds for every $\boldsymbol{w}$ satisfying $\|\boldsymbol{w}\| \leq \frac{\mu}{4\sqrt{2}}$ that*

$$\mathbb{E}\left[\nabla_{\boldsymbol{w}}^2 h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right] \succeq \frac{\theta}{25\sqrt{2\pi}\mu}\boldsymbol{I}.$$

Proof See Section 7.1.3 on Page 54. ∎

To prove that the above hold qualitatively for finite p, i.e., the function $g\left(\boldsymbol{w}; \boldsymbol{X}_0\right)$, we will need first prove that for a fixed $\boldsymbol{w}$ each of the quantity of interest concentrate about their expectation w.h.p., and the function is nice enough (Lipschitz) such that we can extend the results to all $\boldsymbol{w}$ via a discretization argument. The next three propositions provide the desired pointwise concentration results.

Proposition 2.8 *Suppose $0 < \mu \leq \frac{1}{\sqrt{n}}$. For every $\boldsymbol{w} \in \Gamma$, it holds that for any $t > 0$,*

$$\mathbb{P}\left[\left|\frac{\boldsymbol{w}^* \nabla^2 g(\boldsymbol{w}; \boldsymbol{X}_0)\boldsymbol{w}}{\|\boldsymbol{w}\|^2} - \mathbb{E}\left[\frac{\boldsymbol{w}^* \nabla^2 g(\boldsymbol{w}; \boldsymbol{X}_0)\boldsymbol{w}}{\|\boldsymbol{w}\|^2}\right]\right| \geq t\right] \leq 4\exp\left(-\frac{p\mu^2 t^2}{512n^2 + 32n\mu t}\right).$$

Proof See Page 58 under Section 7.1.4. ∎

Proposition 2.9 *For every $\boldsymbol{w} \in \Gamma$, it holds that for any $t > 0$,*

$$\mathbb{P}\left[\left|\frac{\boldsymbol{w}^* \nabla g(\boldsymbol{w}; \boldsymbol{X}_0)}{\|\boldsymbol{w}\|} - \mathbb{E}\left[\frac{\boldsymbol{w}^* \nabla g(\boldsymbol{w}; \boldsymbol{X}_0)}{\|\boldsymbol{w}\|}\right]\right| \geq t\right] \leq 2\exp\left(-\frac{pt^2}{8n + 4\sqrt{n}t}\right).$$

Proof See Page 59 under Section 7.1.4. ∎

19

Proposition 2.10 *Suppose $0 < \mu \leq \frac{1}{\sqrt{n}}$. For every $\boldsymbol{w} \in \Gamma \cap \{\boldsymbol{w} : \|\boldsymbol{w}\| \leq 1/4\}$, it holds that for any $t > 0$,*

$$\mathbb{P}\left[\left\|\nabla^2 g(\boldsymbol{w}; \boldsymbol{X}_0) - \mathbb{E}\left[\nabla^2 g(\boldsymbol{w}; \boldsymbol{X}_0)\right]\right\| \geq t\right] \leq 4n \exp\left(-\frac{p\mu^2 t^2}{512n^2 + 32\mu n t}\right).$$

Proof See Page 60 under Section 7.1.4. ■

The next three propositions provide the desired Lipschitz results.

Proposition 2.11 (Hessian Lipschitz) *Fix any $r_\cap \in (0, 1)$. Over the set $\Gamma \cap \{\boldsymbol{w} : \|\boldsymbol{w}\| \geq r_\cap\}$, $\frac{\boldsymbol{w}^* \nabla^2 g(\boldsymbol{w}; \boldsymbol{X}_0) \boldsymbol{w}}{\|\boldsymbol{w}\|^2}$ is $L_\cap$-Lipschitz with*

$$L_\cap \leq \frac{16n^3}{\mu^2}\|\boldsymbol{X}_0\|_\infty^3 + \frac{8n^{3/2}}{\mu r_\cap}\|\boldsymbol{X}_0\|_\infty^2 + \frac{48n^{5/2}}{\mu}\|\boldsymbol{X}_0\|_\infty^2 + 96n^{5/2}\|\boldsymbol{X}_0\|_\infty.$$

Proof See Page 64 under Section 7.1.5. ■

Proposition 2.12 (Gradient Lipschitz) *Fix any $r_g \in (0, 1)$. Over the set $\Gamma \cap \{\boldsymbol{w} : \|\boldsymbol{w}\| \geq r_g\}$, $\frac{\boldsymbol{w}^* \nabla g(\boldsymbol{w}; \boldsymbol{X}_0)}{\|\boldsymbol{w}\|}$ is L_g-Lipschitz with*

$$L_g \leq \frac{2\sqrt{n}\|\boldsymbol{X}_0\|_\infty}{r_g} + 8n^{3/2}\|\boldsymbol{X}_0\|_\infty + \frac{4n^2}{\mu}\|\boldsymbol{X}_0\|_\infty^2.$$

Proof See Page 65 under Section 7.1.5. ■

Proposition 2.13 (Lipschitz for Hessian around zero) *Fix any $r_\cup \in \left(0, \frac{1}{2}\right)$. Over the set $\Gamma \cap \{\boldsymbol{w} : \|\boldsymbol{w}\| \leq r_\cup\}$, $\nabla^2 g(\boldsymbol{w}; \boldsymbol{X}_0)$ is $L_\cup$-Lipschitz with*

$$L_\cup \leq \frac{4n^2}{\mu^2}\|\boldsymbol{X}_0\|_\infty^3 + \frac{4n}{\mu}\|\boldsymbol{X}_0\|_\infty^2 + \frac{8\sqrt{2}\sqrt{n}}{\mu}\|\boldsymbol{X}_0\|_\infty^2 + 8\|\boldsymbol{X}_0\|_\infty.$$

Proof See Page 65 under Section 7.1.5. ■

Integrating the above pieces, Section 7.2 provides a complete proof of Theorem 2.1.

2.3 Extending to Complete Dictionaries

As hinted in Section 1.4.2, instead of proving things from scratch, we build on the results we have obtained for orthogonal dictionaries. In particular, we will work with the preconditioned data matrix

$$\overline{\boldsymbol{Y}} \doteq \left(\frac{1}{p\theta}\boldsymbol{Y}\boldsymbol{Y}^*\right)^{-1/2}\boldsymbol{Y} \tag{2.14}$$

and show that the function landscape $f\left(\boldsymbol{q}; \overline{\boldsymbol{Y}}\right)$ looks qualitatively like that of orthogonal dictionaries (up to a global rotation), provided that p is large enough.

The next lemma shows $\overline{\boldsymbol{Y}}$ can be treated as being generated from an orthobasis with the same BG coefficients, plus small noise.

Lemma 2.14 *For any $\theta \in (0, 1/2)$, suppose $\boldsymbol{A}_0$ is complete with condition number $\kappa(\boldsymbol{A}_0)$ and $\boldsymbol{X}_0 \sim_{i.i.d.}$ BG (θ). Provided $p \geq C\kappa^4(\boldsymbol{A}_0)\theta n^2 \log(n\theta\kappa(\boldsymbol{A}_0))$, one can write $\overline{\boldsymbol{Y}}$ as defined in (2.14) as*

$$\overline{\boldsymbol{Y}} = \boldsymbol{U}\boldsymbol{V}^*\boldsymbol{X}_0 + \boldsymbol{\Xi}\boldsymbol{X}_0,$$

for a certain $\boldsymbol{\Xi}$ obeying $\|\boldsymbol{\Xi}\| \leq 20\kappa^4(\boldsymbol{A})\sqrt{\frac{\theta n \log p}{p}}$, with probability at least $1 - p^{-8}$. Here $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = $ SVD $(\boldsymbol{A}_0)$, and C is a positive numerical constant.

Proof See Page 69 under Section 7.3. ∎

Notice that $\boldsymbol{U}\boldsymbol{V}^*$ above is orthogonal, and that landscape of $f(\boldsymbol{q}; \overline{\boldsymbol{Y}})$ is simply a rotated version of that of $f(\boldsymbol{q}; \boldsymbol{V}\boldsymbol{U}^*\overline{\boldsymbol{Y}})$, or using the notation in the above lemma, that of $f(\boldsymbol{q}; \boldsymbol{X}_0 + \boldsymbol{V}\boldsymbol{U}^*\boldsymbol{\Xi}\boldsymbol{X}_0) = f(\boldsymbol{q}; \boldsymbol{X}_0 + \widetilde{\boldsymbol{\Xi}}\boldsymbol{X}_0)$ assuming $\widetilde{\boldsymbol{\Xi}} \doteq \boldsymbol{V}\boldsymbol{U}^*\boldsymbol{\Xi}$. So similar to the orthogonal case, it is enough to consider this "canonical" case, and its "canonical" reparametrization:

$$g\left(\boldsymbol{w}; \boldsymbol{X}_0 + \widetilde{\boldsymbol{\Xi}}\boldsymbol{X}_0\right) = \frac{1}{p}\sum_{k=1}^{p} h_\mu\left(\boldsymbol{q}^*(\boldsymbol{w})(\boldsymbol{x}_0)_k + \boldsymbol{q}^*(\boldsymbol{w})\widetilde{\boldsymbol{\Xi}}(\boldsymbol{x}_0)_k\right).$$

The following lemma provides quantitative comparison between the gradient and Hessian of $g\left(\boldsymbol{w}; \boldsymbol{X}_0 + \widetilde{\boldsymbol{\Xi}}\boldsymbol{X}_0\right)$ and that of $g(\boldsymbol{w}; \boldsymbol{X}_0)$.

Lemma 2.15 *There exist positive constants C_a and C_b, such that for all $\boldsymbol{w} \in \Gamma$,*

$$\left\|\nabla_{\boldsymbol{w}}g(\boldsymbol{w}; \boldsymbol{X}_0 + \widetilde{\boldsymbol{\Xi}}\boldsymbol{X}_0) - \nabla_{\boldsymbol{w}}g(\boldsymbol{w}; \boldsymbol{X}_0)\right\| \leq C_a\frac{n}{\mu}\log(np)\|\widetilde{\boldsymbol{\Xi}}\|,$$

$$\left\|\nabla_{\boldsymbol{w}}^2 g(\boldsymbol{w}; \boldsymbol{X}_0 + \widetilde{\boldsymbol{\Xi}}\boldsymbol{X}_0) - \nabla_{\boldsymbol{w}}^2 g(\boldsymbol{w}; \boldsymbol{X}_0)\right\| \leq C_b\max\left\{\frac{n^{3/2}}{\mu^2}, \frac{n^2}{\mu}\right\}\log^{3/2}(np)\|\widetilde{\boldsymbol{\Xi}}\|$$

with probability at least $1 - \theta(np)^{-7} - \exp(-0.3\theta np)$.

Proof See Page 70 under Section 7.3. ∎

Combining the above two lemmas, it is easy to see when p is large enough, $\|\widetilde{\boldsymbol{\Xi}}\| = \|\boldsymbol{\Xi}\|$ is then small enough (Lemma 2.14), and hence the changes to the gradient and Hessian caused by the perturbation are small. This gives the results presented in Theorem 2.3; see Section 7.3 for the detailed proof. In particular, for the p chosen in Theorem 2.3, it holds that

$$\left\|\widetilde{\boldsymbol{\Xi}}\right\| \leq cc_\star\theta\left(\max\left\{\frac{n^{3/2}}{\mu^2}, \frac{n^2}{\mu}\right\}\log^{3/2}(np)\right)^{-1} \tag{2.15}$$

for a certain constant c which can be made arbitrarily small by making the constant C in p large.

3 Finding One Local Minimizer via the Riemannian Trust-Region Method

The above geometric results show every local minimizer of $f(\boldsymbol{q}; \widehat{\boldsymbol{Y}})$ over $\mathbb{S}^{n-1}$ approximately recovers one row of $\boldsymbol{X}_0$. So the crucial problem left now is how to efficiently obtain one of the local minimizers.

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 efficiently produces an accurate approximation[25] 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 [AMS09] for relevant background and details.

3.1 The Riemannian Trust-Region Algorithm over the Sphere

We are interested to seek one local minimizer of the problem

$$\text{minimize} \quad f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}) \doteq \frac{1}{p} \sum_{k=1}^{p} h_\mu(\boldsymbol{q}^* \widehat{\boldsymbol{y}}_i) \quad \text{subject to} \quad \boldsymbol{q} \in \mathbb{S}^{n-1}. \tag{3.1}$$

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

Here, we are interested in the restriction of f to the unit sphere $\mathbb{S}^{n-1}$. Instead of directly approximating the function in $\mathbb{R}^n$, we form quadratic approximations of f in the tangent space of $\mathbb{S}^{n-1}$. Recall that the tangent space of a sphere at a point $\boldsymbol{q} \in \mathbb{S}^{n-1}$ is $T_{\boldsymbol{q}}\mathbb{S}^{n-1} = \{\boldsymbol{\delta} \in \mathbb{R}^n \mid \boldsymbol{q}^* \boldsymbol{\delta} = 0\}$, i.e., the set of vectors that are orthogonal to $\boldsymbol{q}$. Consider $\boldsymbol{\delta} \in T_{\boldsymbol{q}}\mathbb{S}^{n-1}$ with $\|\boldsymbol{\delta}\| = 1$. The map $\gamma(t) : t \mapsto \boldsymbol{q} \cos t + \boldsymbol{\delta} \sin t$ defines a smooth curve on the sphere that satisfies $\gamma(0) = \boldsymbol{q}$ and $\dot{\gamma}(0) = \boldsymbol{\delta}$. The function $f \circ \gamma(t)$ obviously is smooth and we expect Taylor expansion around 0 a good approximation of the function, at least in the vicinity of 0. Taylor's theorem gives

$$f \circ \gamma(t) = f(\boldsymbol{q}) + t \langle \nabla f(\boldsymbol{q}), \boldsymbol{\delta} \rangle + \frac{t^2}{2} \left(\boldsymbol{\delta}^* \nabla^2 f(\boldsymbol{q}) \boldsymbol{\delta} - \langle \nabla f(\boldsymbol{q}), \boldsymbol{q} \rangle \right) + O(t^3).$$

We therefore form the "quadratic" approximation $\widehat{f}(\boldsymbol{\delta}; \boldsymbol{q}) : T_{\boldsymbol{q}}\mathbb{S}^{n-1} \mapsto \mathbb{R}$ as

$$\widehat{f}(\boldsymbol{\delta}; \boldsymbol{q}, \widehat{\boldsymbol{Y}}) \doteq f(\boldsymbol{q}) + \left\langle \nabla f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}), \boldsymbol{\delta} \right\rangle + \frac{1}{2} \boldsymbol{\delta}^* \left(\nabla^2 f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}) - \left\langle \nabla f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}), \boldsymbol{q} \right\rangle \boldsymbol{I} \right) \boldsymbol{\delta}. \tag{3.2}$$

Given the previous iterate $\boldsymbol{q}^{(k-1)}$, the TRM produces the next iterate by generating a solution $\widehat{\boldsymbol{\delta}}$ to

$$\min_{\boldsymbol{\delta} \in T_{\boldsymbol{q}^{(k-1)}}\mathbb{S}^{n-1}, \|\boldsymbol{\delta}\| \le \Delta} \widehat{f}(\boldsymbol{\delta}; \boldsymbol{q}^{(k-1)}), \tag{3.3}$$

and then "pull" the solution $\widehat{\boldsymbol{\delta}}$ from $T_{\boldsymbol{q}}\mathbb{S}^{n-1}$ back to $\mathbb{S}^{n-1}$. Moreover, for any vector $\boldsymbol{\delta} \in T_{\boldsymbol{q}}\mathbb{S}^{n-1}$, the exponential map $\exp_{\boldsymbol{q}}(\boldsymbol{\delta}) : T_{\boldsymbol{q}}\mathbb{S}^{n-1} \mapsto \mathbb{S}^{n-1}$ is

$$\exp_{\boldsymbol{q}}(\boldsymbol{\delta}) = \boldsymbol{q} \cos \|\boldsymbol{\delta}\| + \frac{\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|} \sin \|\boldsymbol{\delta}\|.$$

---

[25]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 ε, and polynomial in other problem parameters.

If we choose the exponential map to pull back the movement $\widehat{\boldsymbol{\delta}}$[26], the next iterate then reads

$$\boldsymbol{q}^{(k)} = \boldsymbol{q}^{(k-1)} \cos \|\widehat{\boldsymbol{\delta}}\| + \frac{\widehat{\boldsymbol{\delta}}}{\|\widehat{\boldsymbol{\delta}}\|} \sin \|\widehat{\boldsymbol{\delta}}\|. \tag{3.4}$$

We have motivated (3.2) and hence the algorithm in an intuitive way from the Taylor approximation to the function f over $\mathbb{S}^{n-1}$. To understand its properties, it is useful to interpret it as a *Riemannian trust-region method* over the manifold $\mathbb{S}^{n-1}$. The class of algorithm is discussed in detail in the monograph [AMS09]. In particular, the quadratic approximation (3.2) can be obtained by noting that the function $f \circ \exp_{\boldsymbol{q}}(\boldsymbol{\delta}; \widehat{\boldsymbol{Y}}) : T_{\boldsymbol{q}} \mathbb{S}^{n-1} \mapsto \mathbb{R}$ obeys

$$f \circ \exp_{\boldsymbol{q}}(\boldsymbol{\delta}; \widehat{\boldsymbol{Y}}) = f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}) + \left\langle \boldsymbol{\delta}, \operatorname{grad} f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}) \right\rangle + \frac{1}{2} \boldsymbol{\delta}^* \operatorname{Hess} f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}) \boldsymbol{\delta} + O(\|\boldsymbol{\delta}\|^3),$$

where $\operatorname{grad} f(\boldsymbol{q}; \widehat{\boldsymbol{Y}})$ and $\operatorname{Hess} f(\boldsymbol{q}; \widehat{\boldsymbol{Y}})$ are the Riemannian gradient and Riemannian Hessian [AMS09] respectively, defined as

$$\operatorname{grad} f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}) \doteq \mathcal{P}_{T_{\boldsymbol{q}} \mathbb{S}^{n-1}} \nabla f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}),$$
$$\operatorname{Hess} f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}) \doteq \mathcal{P}_{T_{\boldsymbol{q}} \mathbb{S}^{n-1}} \left(\nabla^2 f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}) - \left\langle \nabla f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}), \boldsymbol{q} \right\rangle \boldsymbol{I} \right) \mathcal{P}_{T_{\boldsymbol{q}} \mathbb{S}^{n-1}},$$

with $\mathcal{P}_{T_{\boldsymbol{q}} \mathbb{S}^{n-1}} \doteq \boldsymbol{I} - \boldsymbol{q} \boldsymbol{q}^*$ the orthoprojector onto the tangent space $T_{\boldsymbol{q}} \mathbb{S}^{n-1}$. We will use these standard notions in analysis of the algorithm.

To solve the subproblem (3.3) numerically, we can take any matrix $\boldsymbol{U} \in \mathbb{R}^{n \times (n-1)}$ whose columns form an orthonormal basis for $T_{\boldsymbol{q}^{(k-1)}} \mathbb{S}^{n-1}$, and produce a solution $\widehat{\boldsymbol{\xi}}$ to

$$\min_{\|\boldsymbol{\xi}\| \leq \Delta} \quad \widehat{f}(\boldsymbol{U}\boldsymbol{\xi}; \boldsymbol{q}^{(k-1)}), \tag{3.5}$$

where by (3.2),

$$\widehat{f}(\boldsymbol{U}\boldsymbol{\xi}; \boldsymbol{q}^{(k-1)}) = f(\boldsymbol{q}) + \left\langle \boldsymbol{U}^* \nabla f(\boldsymbol{q}^{(k-1)}), \boldsymbol{\xi} \right\rangle +$$
$$\frac{1}{2} \boldsymbol{\xi}^* \left(\boldsymbol{U}^* \nabla^2 f(\boldsymbol{q}^{(k-1)}; \widehat{\boldsymbol{Y}}) \boldsymbol{U} - \left\langle \nabla f(\boldsymbol{q}^{(k-1)}; \widehat{\boldsymbol{Y}}), \boldsymbol{q}^{(k-1)} \right\rangle \boldsymbol{I}_{n-1} \right) \boldsymbol{\xi}.$$

Solution to (3.3) can then be recovered as $\widehat{\boldsymbol{\delta}} = \boldsymbol{U}\widehat{\boldsymbol{\xi}}$. The problem (3.5) is an instance of the classic *trust region subproblem*, i.e., minimizing a quadratic function subject to a single quadratic constraint, which can be solved in polynomial time, either by root finding methods [MS83, CGT00b] or by semidefinite programming (SDP) [RW97, YZ03, FW04, HK14]. As the root finding methods numerically suffer from the so-called "hard case" [MS83], we deploy the SDP approach here. We introduce

$$\tilde{\boldsymbol{\xi}} = [\boldsymbol{\xi}^*, 1]^*, \ \boldsymbol{\Theta} = \tilde{\boldsymbol{\xi}} \tilde{\boldsymbol{\xi}}^*, \ \boldsymbol{M} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{b} \\ \boldsymbol{b}^* & 0 \end{bmatrix}, \tag{3.6}$$

where $\boldsymbol{A} = \boldsymbol{U}^*(\nabla^2 f(\boldsymbol{q}^{(k-1)}; \widehat{\boldsymbol{Y}}) - \left\langle \nabla f(\boldsymbol{q}^{(k-1)}; \widehat{\boldsymbol{Y}}), \boldsymbol{q}^{(k-1)} \right\rangle \boldsymbol{I})\boldsymbol{U}$ and $\boldsymbol{b} = \boldsymbol{U}^* \nabla f(\boldsymbol{q}^{(k-1)}; \widehat{\boldsymbol{Y}})$. The resulting SDP to solve is

$$\text{minimize}_{\boldsymbol{\Theta}} \ \langle \boldsymbol{M}, \boldsymbol{\Theta} \rangle, \ \text{subject to } \operatorname{tr}(\boldsymbol{\Theta}) \leq \Delta^2 + 1, \ \langle \boldsymbol{E}_{n+1}, \boldsymbol{\Theta} \rangle = 1, \ \boldsymbol{\Theta} \succeq \boldsymbol{0}, \tag{3.7}$$

---

[26]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 Chapter 4 of [AMS09].

where $\boldsymbol{E}_{n+1} = \boldsymbol{e}_{n+1}\boldsymbol{e}_{n+1}^*$. Once the problem (3.7) is solved to its optimal $\boldsymbol{\Theta}_\star$, one can provably recover the optimal solution $\boldsymbol{\xi}_\star$ of (3.5) by computing the SVD of $\boldsymbol{\Theta}_\star = \widetilde{\boldsymbol{U}}\boldsymbol{\Sigma}\widetilde{\boldsymbol{V}}^*$, and extract as a subvector by the first $n-1$ coordinates of the principal eigenvector $\widetilde{\boldsymbol{u}}_1$ (see Appendix B of [BV04]).

The choice of trust region size Δ is important both for the convergence theory and practical effectiveness of TRMs. Following standard recommendations (see, e.g., Chapter 4 of [NW06]), we use a backtracking approach which modifies Δ from iteration to iteration based on the accuracy of the approximation $\widehat{f}$. The whole algorithmic procedure is described as pseudocode as Algorithm 1. In our numerical implementation, we randomly initialize $\boldsymbol{q}^{(0)}$ and set $\Delta^{(0)} = 0.1$, $\eta_{vs} =$

Algorithm 1 Riemannian TRM Algorithm for Finding One Local Minimizer

---

Input: Data matrix $\boldsymbol{Y} \in \mathbb{R}^{n \times p}$, smoothing parameter μ and parameters η_{vs}, η_s, γ_i, γ_d, $\Delta_{\max}$, $\Delta_{\min}$
Output: $\widehat{\boldsymbol{q}} \in \mathbb{S}^{n-1}$

1: Initialize $\boldsymbol{q}^{(0)} \in \mathbb{S}^{n-1}$, $\Delta^{(0)}$ and $k = 1$,
2: **while** not converged **do**
3:     Set $\boldsymbol{U} \in \mathbb{R}^{n \times (n-1)}$ to be an orthonormal basis for $T_{\boldsymbol{q}^{(k-1)}}\mathbb{S}^{n-1}$
4:     Solve the trust region subproblem

$$\widehat{\boldsymbol{\xi}} = \underset{\|\boldsymbol{\xi}\| \leq \Delta^{(k-1)}}{\arg\min} \widehat{f}(\boldsymbol{U}\boldsymbol{\xi}; \boldsymbol{q}^{(k-1)}, \widehat{\boldsymbol{Y}})$$

5:     Set

$$\widehat{\boldsymbol{\delta}} \leftarrow \boldsymbol{U}\widehat{\boldsymbol{\xi}}, \quad \widehat{\boldsymbol{q}} \leftarrow \boldsymbol{q}^{(k-1)} \cos\|\widehat{\boldsymbol{\delta}}\| + \frac{\widehat{\boldsymbol{\delta}}}{\|\widehat{\boldsymbol{\delta}}\|}\sin\|\widehat{\boldsymbol{\delta}}\|.$$

6:     Set

$$\rho_k \leftarrow \frac{f(\boldsymbol{q}^{(k-1)}; \widehat{\boldsymbol{Y}}) - f(\widehat{\boldsymbol{q}}; \widehat{\boldsymbol{Y}})}{f(\boldsymbol{q}^{(k-1)}; \widehat{\boldsymbol{Y}}) - \widehat{f}(\widehat{\boldsymbol{\delta}}; \boldsymbol{q}^{(k-1)}, \widehat{\boldsymbol{Y}})}$$

7:     **if** $\rho_k \geq \eta_{vs}$ **then**
8:         Set $\boldsymbol{q}^{(k)} \leftarrow \widehat{\boldsymbol{q}}$ and $\Delta^{(k)} \leftarrow \min\left(\gamma_i\Delta^{(k-1)}, \Delta_{\max}\right)$.   ▷ very successful
9:     **else if** $\rho_k \geq \eta_s$ **then**
10:        Set $\boldsymbol{q}^{(k)} \leftarrow \widehat{\boldsymbol{q}}$ and $\Delta^{(k)} \leftarrow \Delta^{(k-1)}$.   ▷ successful
11:     **else**
12:        Set $\boldsymbol{q}^{(k)} \leftarrow \boldsymbol{q}^{(k-1)}$ and $\Delta^{(k)} \leftarrow \max\left(\gamma_d\Delta^{(k-1)}, \Delta_{\min}\right)$.   ▷ unsuccessful
13:     **end if**
14:     Set $k = k + 1$.
15: **end while**

0.9, $\eta_s = 0.1$, $\gamma_d = 1/2$, $\gamma_i = 2$, $\Delta_{\max} = 1$ and $\Delta_{\min} = 10^{-16}$, and the algorithm is stopped when $\left(f(\widehat{\boldsymbol{q}}) - f(\boldsymbol{q}^{(k-1)})\right)/\|\widehat{\boldsymbol{\delta}}\| \leq 10^{-6}$.

3.2 Main Convergence Results

By using general results on the Riemannian TRM (see, e.g., Chapter 7 of [AMS09]), it is not difficult to prove that the iterates $\boldsymbol{q}^{(k)}$ produced by Algorithm 1 converge to a critical point of the objective $f(\boldsymbol{q})$ over $\mathbb{S}^{n-1}$. In this section, we show that under our probabilistic assumptions, this claim can be 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. The arguments described in Section 2 show that with high probability every local minimizer of f produces a close approximation of one row of $\boldsymbol{X}_0$. Taken together, this implies that the algorithm efficiently produces a close approximation to one row of $\boldsymbol{X}_0$.

Our next two theorems summarize the convergence results for orthogonal and complete dictionaries, respectively.

Theorem 3.1 (TRM convergence - orthogonal dictionary) *Suppose the dictionary $\boldsymbol{A}_0$ is orthogonal. Then there exists a positive constant C, such that for all $\theta \in (0, 1/2)$, and $\mu < \min\{c_a\theta n^{-1}, c_b n^{-5/4}\}$, whenever $\exp(n) \geq p \geq Cn^3 \log \frac{n}{\mu\theta}/(\mu^2\theta^2)$, with probability at least $1 - 8n^2p^{-10} - \theta(np)^{-7} - \exp(-0.3\theta np) - p^{-10} - c_c \exp(-c_d p\mu^2\theta^2/n^2)$, the Riemannian trust-region algorithm with input data matrix $\widehat{\boldsymbol{Y}} = \boldsymbol{Y}$, any initialization $\boldsymbol{q}^{(0)}$ on the sphere, and a step size satisfying*

$$\Delta \leq \min\left\{\frac{c_e c_\star \theta\mu^2}{n^{5/2}\log^{3/2}(np)}, \frac{c_f c_\sharp^3 \theta^3\mu}{n^{7/2}\log^{7/2}(np)}\right\}. \tag{3.8}$$

returns a solution $\widehat{\boldsymbol{q}} \in \mathbb{S}^{n-1}$ which is ε near to one of the local minimizers $\boldsymbol{q}_\star$ (i.e., $\|\widehat{\boldsymbol{q}} - \boldsymbol{q}_\star\| \leq \varepsilon$) in

$$\max\left\{\frac{c_g n^6 \log^3(np)}{c_\star^3\theta^3\mu^4}, \frac{c_h n}{c_\sharp^2\theta^2\Delta^2}\right\}\left(f(\boldsymbol{q}^{(0)}) - f(\boldsymbol{q}_\star)\right) + \log\log\frac{c_i c_\star \theta\mu}{\varepsilon n^{3/2}\log^{3/2}(np)} \tag{3.9}$$

iterations. Here $c_\star$, $c_\sharp$ as defined in Theorem 2.1 and Lemma 3.9 respectively ($c_\star$ and $c_\sharp$ can be set to the same constant value), and c_a, c_b are the same numerical constants as defined in Theorem 2.1, c_c to c_i are other positive numerical constants.

Theorem 3.2 (TRM convergence - complete dictionary) *Suppose the dictionary $\boldsymbol{A}_0$ is complete with condition number $\kappa(\boldsymbol{A}_0)$. There exists a positive constant C, such that for all $\theta \in (0, 1/2)$, and $\mu < \min\{c_a\theta n^{-1}, c_b n^{-5/4}\}$, whenever $\exp(n) \geq p \geq \frac{C}{c_\star^2\theta}\max\left\{\frac{n^4}{\mu^4}, \frac{n^5}{\mu^2}\right\}\kappa^8(\boldsymbol{A}_0)\log^4\left(\frac{\kappa(\boldsymbol{A}_0)n}{\mu\theta}\right)$, with probability at least $1 - 8n^2p^{-10} - \theta(np)^{-7} - \exp(-0.3\theta np) - 2p^{-8} - c_c \exp(-c_d p\mu^2\theta^2/n^2)$, the Riemannian trust-region algorithm with input data matrix $\overline{\boldsymbol{Y}} \doteq \sqrt{p\theta}(\boldsymbol{YY}^*)^{-1/2}\boldsymbol{Y}$ where $\boldsymbol{U\Sigma V}^* = \mathtt{SVD}(\boldsymbol{A}_0)$, any initialization $\boldsymbol{q}^{(0)}$ on the sphere and a step size satisfying*

$$\Delta \leq \min\left\{\frac{c_e c_\star \theta\mu^2}{n^{5/2}\log^{3/2}(np)}, \frac{c_f c_\sharp^3 \theta^3\mu}{n^{7/2}\log^{7/2}(np)}\right\}. \tag{3.10}$$

returns a solution $\widehat{\boldsymbol{q}} \in \mathbb{S}^{n-1}$ which is ε near to one of the local minimizers $\boldsymbol{q}_\star$ (i.e., $\|\widehat{\boldsymbol{q}} - \boldsymbol{q}_\star\| \leq \varepsilon$) in

$$\max\left\{\frac{c_g n^6 \log^3(np)}{c_\star^3\theta^3\mu^4}, \frac{c_h n}{c_\sharp^2\theta^2\Delta^2}\right\}\left(f(\boldsymbol{q}^{(0)}) - f(\boldsymbol{q}_\star)\right) + \log\log\frac{c_i c_\star \theta\mu}{\varepsilon n^{3/2}\log^{3/2}(np)} \tag{3.11}$$

iterations. Here $c_\star$, $c_\sharp$ as defined in Theorem 2.1 and Lemma 3.9 respectively ($c_\star$ and $c_\sharp$ can be set to the same constant value), and c_a, c_b are the same numerical constants as defined in Theorem 2.1, c_c to c_i are other positive numerical constants.

25

Our convergence result shows that for any target accuracy $\varepsilon > 0$ the algorithm terminates within polynomially many steps. Our estimate of the number of steps is pessimistic: our analysis has assumed a fixed step size Δ and the running time is relatively large degree polynomial in p and n, while on typical numerical examples (e.g., $\mu = 10^{-2}$, $n \sim 100$, and $\varepsilon = O(\mu)$), the algorithm with adaptive step size as described in Algorithm 1 produces an accurate solution in relatively few (20-50) iterations. Nevertheless, 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 [MK87] show that in general it is NP-hard to find a local minimum of a nonconvex function.

3.3 Useful Technical Results and Proof Ideas for Orthogonal Dictionaries

The reason that our algorithm is successful derives from the geometry depicted in Figure 2 and formalized in Theorem 2.1. Basically, the sphere $\mathbb{S}^{n-1}$ 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 (an eigenvalue of the Hessian which is bounded below zero). The Riemannian TRM aglorithm is capable of exploiting these quantities to reduce the objective value by at least a constant in each iteration. The total number of iterations spent away from the vicinity of the local minimizers can be bounded by comparing this constant to the initial objective value. Our proofs follow exactly this line and make the various quantities precise.

3.3.1 Basic Facts about the Sphere

For any point $q \in \mathbb{S}^{n-1}$, the tangent space $T_q \mathbb{S}^{n-1}$ and the orthoprojector $\mathcal{P}_{T_q \mathbb{S}^{n-1}}$ onto $T_q \mathbb{S}^{n-1}$ are given by

$$T_q \mathbb{S}^{n-1} = \left\{ \boldsymbol{\delta} \in \mathbb{R}^n \mid q^* \boldsymbol{\delta} = 0 \right\},$$
$$\mathcal{P}_{T_q \mathbb{S}^{n-1}} = (\boldsymbol{I} - q q^*) = \boldsymbol{U} \boldsymbol{U}^*,$$

where $\boldsymbol{U} \in \mathbb{R}^{n \times (n-1)}$ is an arbitrary orthonormal basis for $T_q \mathbb{S}^{n-1}$ (note that the orthoprojector is independent of the basis $\boldsymbol{U}$ we choose). Moreover, for any $\boldsymbol{\delta} \in T_q \mathbb{S}^{n-1}$, the exponential map $\exp_q(\boldsymbol{\delta}) : T_q \mathbb{S}^{n-1} \mapsto \mathbb{S}^{n-1}$ is given by

$$\exp_q(\boldsymbol{\delta}) = q \cos \|\boldsymbol{\delta}\| + \frac{\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|} \sin \|\boldsymbol{\delta}\| .$$

Let $\nabla f(q)$ and $\nabla^2 f(q)$ denote the usual (Euclidean) gradient and Hessian of f w.r.t. q in $\mathbb{R}^n$. For our specific f defined in (3.1), it is easy to check that

$$\nabla f\left(q; \widehat{\boldsymbol{Y}}\right) = \frac{1}{p} \sum_{k=1}^{p} \tanh\left(\frac{q^* \widehat{\boldsymbol{y}}_k}{\mu}\right) \widehat{\boldsymbol{y}}_k, \tag{3.12}$$

$$\nabla^2 f\left(q; \widehat{\boldsymbol{Y}}\right) = \frac{1}{p} \sum_{k=1}^{p} \frac{1}{\mu} \left[1 - \tanh^2\left(\frac{q^* \widehat{\boldsymbol{y}}_k}{\mu}\right) \right] \widehat{\boldsymbol{y}}_k \widehat{\boldsymbol{y}}_k^*. \tag{3.13}$$

Since $\mathbb{S}^{n-1}$ is an embedded submanifold of $\mathbb{R}^n$, the Riemannian gradient and Riemannian Hessian defined on $T_{\boldsymbol{q}}\mathbb{S}^{n-1}$ are given by

$$\operatorname{grad} f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}) = \mathcal{P}_{T_{\boldsymbol{q}}\mathbb{S}^{n-1}} \nabla f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}), \tag{3.14}$$

$$\operatorname{Hess} f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}) = \mathcal{P}_{T_{\boldsymbol{q}}\mathbb{S}^{n-1}} \left(\nabla^2 f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}) - \left\langle \nabla f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}), \boldsymbol{q} \right\rangle \boldsymbol{I} \right) \mathcal{P}_{T_{\boldsymbol{q}}\mathbb{S}^{n-1}}; \tag{3.15}$$

so the second-order Taylor approximation for the function f is

$$\widehat{f}\left(\boldsymbol{\delta}; \boldsymbol{q}, \widehat{\boldsymbol{Y}}\right) = f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}) + \left\langle \boldsymbol{\delta}, \operatorname{grad} f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}) \right\rangle + \frac{1}{2}\boldsymbol{\delta}^* \operatorname{Hess} f(\boldsymbol{q}; \widehat{\boldsymbol{Y}})\boldsymbol{\delta}, \qquad \forall \, \boldsymbol{\delta} \in T_{\boldsymbol{q}}\mathbb{S}^{n-1}.$$

The first order necessary condition for *unconstrained* minimization of function $\widehat{f}$ over $T_{\boldsymbol{q}}\mathbb{S}^{n-1}$ is

$$\operatorname{grad} f(\boldsymbol{q}; \widehat{\boldsymbol{Y}}) + \operatorname{Hess} f(\boldsymbol{q}; \widehat{\boldsymbol{Y}})\boldsymbol{\delta}_\star = \boldsymbol{0}; \tag{3.16}$$

if $\operatorname{Hess} f(\boldsymbol{q})$ is positive semidefinite and has full rank $n - 1$ (hence "nondegenerate"[27]), the unique solution $\boldsymbol{\delta}_\star$ is

$$\boldsymbol{\delta}_\star = -\boldsymbol{U} \left(\boldsymbol{U}^* \left[\operatorname{Hess} f(\boldsymbol{q})\right] \boldsymbol{U}\right)^{-1} \boldsymbol{U}^* \operatorname{grad} f(\boldsymbol{q}),$$

which is also invariant to the choice of basis $\boldsymbol{U}$. Given a tangent vector $\boldsymbol{\delta} \in T_{\boldsymbol{q}}\mathbb{S}^{n-1}$, let $\gamma(t) \doteq \exp_{\boldsymbol{q}}(t\boldsymbol{\delta})$ denote a geodesic curve on $\mathbb{S}^{n-1}$. Following the notation of [AMS09], let

$$\mathcal{P}_\gamma^{\tau \leftarrow 0} : T_{\boldsymbol{q}}\mathbb{S}^{n-1} \to T_{\gamma(\tau)}\mathbb{S}^{n-1}$$

denotes the parallel translation operator, which translates the tangent vector $\boldsymbol{\delta}$ at $\boldsymbol{q} = \gamma(0)$ to a tangent vector at $\gamma(\tau)$, in a "parallel" manner. In the sequel, we identify $\mathcal{P}_\gamma^{\tau \leftarrow 0}$ with the following $n \times n$ matrix, whose restriction to $T_{\boldsymbol{q}}\mathbb{S}^{n-1}$ is the parallel translation operator (the detailed derivation can be found in Chapter 8.1 of [AMS09]):

$$
\begin{aligned}
\mathcal{P}_\gamma^{\tau \leftarrow 0} &= \left(\boldsymbol{I} - \frac{\boldsymbol{\delta}\boldsymbol{\delta}^*}{\|\boldsymbol{\delta}\|^2}\right) - \boldsymbol{q} \sin\left(\tau \|\boldsymbol{\delta}\|\right) \frac{\boldsymbol{\delta}^*}{\|\boldsymbol{\delta}\|} + \frac{\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|} \cos\left(\tau \|\boldsymbol{\delta}\|\right) \frac{\boldsymbol{\delta}^*}{\|\boldsymbol{\delta}\|} \\
&= \boldsymbol{I} + (\cos(\tau \|\boldsymbol{\delta}\|) - 1) \frac{\boldsymbol{\delta}\boldsymbol{\delta}^*}{\|\boldsymbol{\delta}\|^2} - \sin\left(\tau \|\boldsymbol{\delta}\|\right) \frac{\boldsymbol{q}\boldsymbol{\delta}^*}{\|\boldsymbol{\delta}\|}.
\end{aligned}
\tag{3.17}
$$

Similarly, following the notation of [AMS09], we denote the inverse of this matrix by $\mathcal{P}_\gamma^{0 \leftarrow \tau}$, where its restriction to $T_{\gamma(\tau)}\mathbb{S}^{n-1}$ is the inverse of the parallel translation operator $\mathcal{P}_\gamma^{\tau \leftarrow 0}$.

3.3.2 Key Steps towards the Proof

Note that for any orthogonal $\boldsymbol{A}_0$, $f(\boldsymbol{q}; \boldsymbol{A}_0\boldsymbol{X}_0) = f(\boldsymbol{A}_0^*\boldsymbol{q}; \boldsymbol{X}_0)$. In words, this is the above established fact that the function landscape of $f(\boldsymbol{q}; \boldsymbol{A}_0\boldsymbol{X}_0)$ is a rotated version of that of $f(\boldsymbol{q}; \boldsymbol{X}_0)$. Thus, any local minimizer $\boldsymbol{q}_\star$ of $f(\boldsymbol{q}; \boldsymbol{X}_0)$ is rotated to $\boldsymbol{A}_0\boldsymbol{q}_\star$, one minimizer of $f(\boldsymbol{q}; \boldsymbol{A}_0\boldsymbol{X}_0)$. Also if our algorithm generates iteration sequence $\boldsymbol{q}_0, \boldsymbol{q}_1, \boldsymbol{q}_2, \ldots$ for $f(\boldsymbol{q}; \boldsymbol{X}_0)$ upon initialization $\boldsymbol{q}_0$, it will generate the

---

[27]Note that the $n \times n$ matrix $\operatorname{Hess} f(\boldsymbol{q}; \widehat{\boldsymbol{Y}})$ has rank at most $n - 1$, as the nonzero $\boldsymbol{q}$ obviously is in its null space. When $\operatorname{Hess} f(\boldsymbol{q}; \widehat{\boldsymbol{Y}})$ 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.

iteration sequence $\boldsymbol{A}_0\boldsymbol{q}_0, \boldsymbol{A}_0\boldsymbol{q}_1, \boldsymbol{A}_0\boldsymbol{q}_2, \ldots$ for $f\left(\boldsymbol{q}; \boldsymbol{A}_0\boldsymbol{X}_0\right)$. So w.l.o.g. it is adequate that we prove the convergence results for the case $\boldsymbol{A}_0 = \boldsymbol{I}$. So in this section (Section 3.3), we write $f(\boldsymbol{q})$ to mean $f(\boldsymbol{q}; \boldsymbol{X}_0)$.

We partition the sphere into three regions, for which we label as R_{I}, R_{II}, R_{III}, corresponding to the strongly convex, nonzero gradient, and negative curvature regions, respectively (see Theorem 2.1). That is, R_{I} consists of a union of $2n$ spherical caps of radius $\frac{\mu}{4\sqrt{2}}$, each centered around a signed standard basis vector $\pm\boldsymbol{e}_i$. R_{II} consist of the set difference of a union of $2n$ spherical caps of radius $\frac{1}{20\sqrt{5}}$, centered around the standard basis vectors $\pm\boldsymbol{e}_i$, and R_{I}. Finally, R_{III} covers the rest of the sphere. We say a trust-region step takes an R_{I} step if the current iterate is in R_{I}; similarly for R_{II} and R_{III} steps. Since we use the geometric structures derived in Theorem 2.1 and Corollary 2.2, the conditions

$$\theta \in (0, 1/2), \quad \mu < \min\left\{c_a \theta n^{-1}, c_b n^{-5/4}\right\}, \quad p \geq \frac{C}{\mu^2\theta^2}n^3 \log\frac{n}{\mu\theta} \tag{3.18}$$

are always in force.

At each step k of the algorithm, suppose $\boldsymbol{\delta}^{(k)}$ is the minimizer of the trust-region subproblem (3.3). We call the step "*constrained*" if $\left\|\boldsymbol{\delta}^{(k)}\right\| = \Delta$ (the minimizer lies on the boundary and hence the constraint is active), and call it "*unconstrained*" if $\|\boldsymbol{\delta}^{(k)}\| < \Delta$ (the minimizer lies in the relative interior and hence the constraint is not in force). Thus, in the unconstrained case the optimality condition is (3.16).

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

Lemma 3.3 *We have the following estimates about ∇f and $\nabla^2 f$:*

$$\sup_{\boldsymbol{q}\in\mathbb{S}^{n-1}} \left\|\nabla f\left(\boldsymbol{q}\right)\right\| \doteq M_\nabla \leq \sqrt{n}\left\|\boldsymbol{X}_0\right\|_\infty,$$

$$\sup_{\boldsymbol{q}\in\mathbb{S}^{n-1}} \left\|\nabla^2 f\left(\boldsymbol{q}\right)\right\| \doteq M_{\nabla^2} \leq \frac{n}{\mu}\left\|\boldsymbol{X}_0\right\|_\infty^2,$$

$$\sup_{\boldsymbol{q},\boldsymbol{q}'\in\mathbb{S}^{n-1},\boldsymbol{q}\neq\boldsymbol{q}'} \frac{\left\|\nabla f\left(\boldsymbol{q}\right) - \nabla f\left(\boldsymbol{q}'\right)\right\|}{\left\|\boldsymbol{q} - \boldsymbol{q}'\right\|} \doteq L_\nabla \leq \frac{n}{\mu}\left\|\boldsymbol{X}_0\right\|_\infty^2,$$

$$\sup_{\boldsymbol{q},\boldsymbol{q}'\in\mathbb{S}^{n-1},\boldsymbol{q}\neq\boldsymbol{q}'} \frac{\left\|\nabla^2 f\left(\boldsymbol{q}\right) - \nabla^2 f\left(\boldsymbol{q}'\right)\right\|}{\left\|\boldsymbol{q} - \boldsymbol{q}'\right\|} \doteq L_{\nabla^2} \leq \frac{2}{\mu^2}n^{3/2}\left\|\boldsymbol{X}_0\right\|_\infty^3.$$

Proof See Page 72 under Section 8. ∎

Our next lemma says if the trust-region step size Δ is small enough, one Riemannian trust-region step reduces the objective value by a certain amount when there is any descent direction.

Lemma 3.4 *Suppose that the trust region size $\Delta \leq 1$, and there exists a tangent vector $\boldsymbol{\delta} \in T_{\boldsymbol{q}}\mathbb{S}^{n-1}$ with $\|\boldsymbol{\delta}\| \leq \Delta$, such that*

$$f(\exp_{\boldsymbol{q}}(\boldsymbol{\delta})) \leq f(\boldsymbol{q}) - s$$

for some positive scalar $s \in \mathbb{R}$. Then the trust region subproblem produces a point $\boldsymbol{\delta}_\star$ with

$$f(\exp_{\boldsymbol{q}}(\boldsymbol{\delta}_\star)) \leq f(\boldsymbol{q}) - s + \frac{1}{3}\eta_f\Delta^3,$$

where $\eta_f \doteq M_\nabla + 2M_{\nabla^2} + L_\nabla + L_{\nabla^2}$ and M_∇, M_{∇^2}, L_∇, L_{∇^2} are the quantities defined in Lemma 3.3.

28

Proof See Page 72 under Section 8. ∎

To show decrease in objective value for R_{II} and R_{III}, now it is enough to exhibit a descent direction for each point in these regions. The next two lemmas help us almost accomplish the goal. For convenience again we choose to state the results for the "canonical" section that is in the vicinity of e_n and the projection map $q(w) = [w; (1 - \|w\|^2)^{1/2}]$, with the idea that similar statements hold for other symmetric sections.

Lemma 3.5 *Suppose that the trust region size $\Delta \leq 1$, $w^* \nabla g(w) / \|w\| \geq \beta_g$ for some scalar β_g, and that $w^* \nabla g(w) / \|w\|$ is L_g-Lipschitz on an open ball $\mathcal{B}\left(w, \frac{3\Delta}{2\pi \sqrt{n}}\right)$ centered at w. Then there exists a tangent vector $\boldsymbol{\delta} \in T_q \mathbb{S}^{n-1}$ with $\|\boldsymbol{\delta}\| \leq \Delta$, such that*

$$f(\exp_q(\boldsymbol{\delta})) \leq f(q) - \min\left\{ \frac{\beta_g^2}{2L_g}, \frac{3\beta_g \Delta}{4\pi \sqrt{n}} \right\}.$$

Proof See Page 74 under Section 8. ∎

Lemma 3.6 *Suppose that the trust-region size $\Delta \leq 1$, $w^* \nabla^2 g(w) w / \|w\|^2 \leq -\beta_\frown$, for some $\beta_\frown$, and that $w^* \nabla^2 g(w) w / \|w\|^2$ is $L_\frown$ Lipschitz on the open ball $\mathcal{B}\left(w, \frac{3\Delta}{2\pi \sqrt{n}}\right)$ centered at w. Then there exists a tangent vector $\boldsymbol{\delta} \in T_q \mathbb{S}^{n-1}$ with $\|\boldsymbol{\delta}\| \leq \Delta$, such that*

$$f(\exp_q(\boldsymbol{\delta})) \leq f(q) - \min\left\{ \frac{2\beta_\frown^3}{3L_\frown^2}, \frac{3\Delta^2 \beta_\frown}{8\pi^2 n} \right\}.$$

Proof See Page 74 under Section 8. ∎

One can take $\beta_g = \beta_\frown = c_\star \theta$ as shown in Theorem 2.1, and take the Lipschitz results in Section 2.2 (note that $\|X_0\|_\infty \leq 4 \log^{1/2}(np)$ w.h.p. by Lemma 7.11), repeat the argument for other $2n - 1$ symmetric regions, and conclude that w.h.p. the objective value decreases by at least a constant amount. The next proposition summarizes the results.

Proposition 3.7 *Assume (3.18). In regions R_{II} and R_{III}, each trust-region step reduces the objective value by at least*

$$d_{\mathrm{II}} = \frac{1}{2} \min\left(\frac{c_\star^2 c_a \theta^2 \mu}{n^2 \log(np)}, \frac{3\Delta c_\star \theta}{4\pi \sqrt{n}} \right), \quad \text{and} \quad d_{\mathrm{III}} = \frac{1}{2} \min\left(\frac{c_\star^3 c_b \theta^3 \mu^4}{n^6 \log^3(np)}, \frac{3\Delta^2 c_\star \theta}{8\pi^2 n} \right) \quad (3.19)$$

respectively, provided that

$$\Delta < \frac{c_c c_\star \theta \mu^2}{n^{5/2} \log^{3/2}(np)}, \quad (3.20)$$

where c_a to c_c are positive numerical constants, and $c_\star$ is as defined in Theorem 2.1.

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 $q^{(k)}$ is in the region R_{II}, by Theorem 2.1, w.h.p., we have $w^* g(w) / \|w\| \geq c_\star \theta$ for the constant $c_\star$. By Proposition 2.12 and Lemma 7.11, w.h.p., $w^* g(w) / \|w\|$

29

is $C_2 n^2 \log(np)/\mu$-Lipschitz. Therefore, By Lemma 3.4 and Lemma 3.5, a trust-region step decreases the objective value by at least

$$d_{\mathrm{II}} \doteq \min\left(\frac{c_\star^2\theta^2\mu}{2C_2 n^2 \log(np)}, \frac{3c_\star\theta\Delta}{4\pi\sqrt{n}}\right) - \frac{c_0 n^{3/2}\log^{3/2}(np)}{3\mu^2}\Delta^3.$$

Similarly, if $q^{(k)}$ is in the region R_{III}, by Proposition 2.11, Theorem 2.1 and Lemma 7.11, w.h.p., $w^*\nabla^2 g(w)\,w/\|w\|^2$ is $C_3 n^3\log^{3/2}(np)/\mu^2$-Lipschitz and upper bounded by $-c_\star\theta$. By Lemma 3.4 and Lemma 3.6, a trust-region step decreases the objective value by at least

$$d_{\mathrm{III}} \doteq \min\left(\frac{2c_\star^3\theta^3\mu^4}{3C_3^2 n^6\log^3(np)}, \frac{3\Delta^2 c_\star\theta}{8\pi^2 n}\right) - \frac{c_0 n^{3/2}\log^{3/2}(np)}{3\mu^2}\Delta^3.$$

It can be easily verified that when Δ obeys (3.19), (3.20) holds. ∎

The analysis for R_{I} is slightly trickier. In this region, near each local minimizer, the objective function is strongly convex. So we still expect each trust-region step decreases the objective value. On the other hand, it is very unlikely that we can provide a universal lower bound for the amount of decrease - as the iteration sequence approaches one local minimizer, the movement is expected to be diminishing. Nevertheless, close to the minimizer the trust-region algorithm takes "unconstrained" steps. For constrained R_{I} steps, we will again show reduction in objective value by at least a fixed amount; for unconstrained step, we will show the distance between the iterate and the nearest local minimizer drops down rapidly.

The next lemma concerns the function value reduction for constrained R_{I} steps.

Lemma 3.8 *Suppose the trust-region size $\Delta \le 1$, and that at a given iterate k, $\mathrm{Hess}\, f\left(q^{(k)}\right) \succeq m_H \mathcal{P}_{T_{q^{(k)}}\mathbb{S}^{n-1}}$, and $\left\|\mathrm{Hess}\, f\left(q^{(k)}\right)\right\| \le M_H$. Further assume the optimal solution $\delta_\star \in T_{q^{(k)}}\mathbb{S}^{n-1}$ to the trust-region sub-problem (3.3) satisfies $\|\delta_\star\| = \Delta$, i.e., the norm constraint is active. Then there exists a tangent vector $\delta \in T_{q^{(k)}}\mathbb{S}^{n-1}$ with $\|\delta\| \le \Delta$, such that*

$$f(\exp_{q^{(k)}}(\delta)) \;\le\; f\left(q^{(k)}\right) - \frac{m_H^2\Delta^2}{M_H} + \frac{1}{6}\eta_f\Delta^3,$$

where η_f is defined the same as Lemma 3.4.

Proof See Page 75 under Section 8. ∎

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 3.9 *There exist positive constants C and $c_\sharp$, such that for all $\theta \in (0, 1/2)$ and $\mu < \min\{\theta, 1/10\}$, whenever $p \ge Cn^3\log\frac{n}{\theta\mu}/(\mu\theta^2)$, it holds with probability at least $1 - \theta(np)^{-7} - \exp(-0.3\theta np) - p^{-10}$ that for all q with $\|w(q)\| \le \frac{\mu}{4\sqrt{2}}$,*

$$\mathrm{Hess}\, f(q) \succeq c_\sharp\frac{\theta}{\mu}\mathcal{P}_{T_q\mathbb{S}^{n-1}}.$$

Proof See Page 78 under Section 8. ■

We know that $\|\boldsymbol{X}_0\|_\infty \leq 4\log^{1/2}(np)$ w.h.p., and hence by the definition of Riemannian Hessian and Lemma 3.3,

$$M_H \doteq \|\mathrm{Hess}\, f(\boldsymbol{q})\| \leq \left\|\nabla^2 f(\boldsymbol{q})\right\| + \|\nabla f(\boldsymbol{q})\| \leq M_{\nabla^2} + M_\nabla \leq \frac{2n}{\mu}\|\boldsymbol{X}_0\|_\infty^2 \leq \frac{16n}{\mu}\log(np),$$

Combining this estimate and Lemma 3.9, and Lemma 3.4, we obtain a concrete lower bound for the reduction of objective value for each constrained R_I step.

Proposition 3.10 *Assume* (3.18). *Each constrained R_I trust-region step (i.e., $\|\boldsymbol{\delta}\| = \Delta$) reduces the objective value by at least*

$$d_\mathrm{I} = \frac{cc_\star^2\theta^2}{\mu n\log(np)}\Delta^2, \tag{3.21}$$

provided

$$\Delta \leq \frac{c'c_\sharp^2\theta^2\mu}{n^{5/2}\log^{5/2}(np)}. \tag{3.22}$$

The constant $c_\sharp$ is as defined in Lemma 3.9 and c, c' are a positive numerical constants.

Proof We only consider the symmetric section in the vicinity of $\boldsymbol{e}_n$ and the claims carry on to others by symmetry. We have that w.h.p.

$$\|\mathrm{Hess}\, f(\boldsymbol{q})\| \leq \frac{16n}{\mu}\log(np), \quad \text{and} \quad \mathrm{Hess}\, f(\boldsymbol{q}) \succeq c_\sharp\frac{\theta}{\mu}\mathcal{P}_{T_{\boldsymbol{q}}\mathbb{S}^{n-1}},$$

where $c_\sharp$ is as defined in Lemma 3.9. Combining these estimates with Lemma 3.4 and Lemma 3.8, one trust-region step will find next iterate $\boldsymbol{q}^{(k+1)}$ that decreases the objective value by at least

$$d_\mathrm{I} \doteq \frac{c_\sharp^2\theta^2/\mu^2}{2n\log(np)/\mu}\Delta^2 - \frac{c_0 n^{3/2}\log^{3/2}(np)}{\mu^2}\Delta^3.$$

Finally, by the condition on Δ in (3.22) and the assumed conditions (3.18), we obtain

$$d_\mathrm{I} \geq \frac{c_\sharp^2\theta^2}{2\mu n\log(np)}\Delta^2 - \frac{c_0 n^{3/2}\log^{3/2}(np)}{\mu^2}\Delta^3 \geq \frac{c_\sharp^2\theta^2}{4\mu n\log(np)}\Delta^2,$$

as desired. ■

By the proof strategy for R_I we sketched before Lemma 3.8, we expect the iteration sequence ultimately always takes unconstrained steps when it moves very near to a local minimizer. We will show that the following is true: when Δ is small enough, once the iteration sequence starts to take unconstrained R_I step, it will take consecutive unconstrained R_I steps afterwards. It takes two steps to show this: (1) upon an unconstrained R_I step, the next iterate will stay in R_I. It is obvious we can make $\Delta \in O(1)$ to ensure the next iterate stays in $R_\mathrm{I} \cup R_\mathrm{II}$. To strengthen the result, we use the gradient information. From Theorem 2.1, we expect the magnitudes of the gradients in R_II to be lower bounded; on the other hand, in R_I where points are near local minimizers, continuity

31

argument implies that the magnitudes of gradients should be upper bounded. We will show that when Δ is small enough, there is a gap between these two bounds, implying the next iterate stays in R_{I}; (2) when Δ 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_{II}.

Lemma 3.11 *For all q satisfying $\frac{\mu}{4\sqrt{2}} \leq \|w(q)\| \leq \frac{1}{20\sqrt{5}}$, it holds that*

$$\|\operatorname{grad} f(q)\| \geq \frac{9}{10} \frac{w^* \nabla g(w)}{\|w\|}.$$

Proof See Page 82 under Section 8. ∎

Assuming (3.18), Theorem 2.1 gives that w.h.p. $w^* \nabla g(w)/\|w\| \geq c_\star \theta$. Thus, w.h.p, $\|\operatorname{grad} f(q)\| \geq 9c_\star\theta/10$ for all $q \in R_{\text{II}}$. The next lemma compares the magnitudes of gradients before and after taking one unconstrained R_{I} 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 3.12 *Suppose the trust-region size $\Delta \leq 1$, and at a given iterate k, $\operatorname{Hess} f\left(q^{(k)}\right) \succeq m_H \mathcal{P}_{T_{q^{(k)}} \mathbb{S}^{n-1}}$, and that the unique minimizer $\delta_\star \in T_{q^{(k)}} \mathbb{S}^{n-1}$ to the trust region subproblem (3.3) satisfies $\|\delta_\star\| < \Delta$ (i.e., the constraint is inactive). Then, for $q^{(k+1)} = \exp_{q^{(k)}}(\delta_\star)$, we have*

$$\|\operatorname{grad} f(q^{(k+1)})\| \leq \frac{L_H}{2m_H^2} \|\operatorname{grad} f(q^{(k)})\|^2,$$

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

Proof See Page 82 under Section 8. ∎

We can now bound the Riemannian gradient of the next iterate as

$$\begin{aligned}
\|\operatorname{grad} f(q^{(k+1)})\| &\leq \frac{L_H}{2m_H^2} \|\operatorname{grad} f(q^{(k)})\|^2 \\
&\leq \frac{L_H}{2m_H^2} \|[U^* \operatorname{Hess} f(q^{(k)}) U][U^* \operatorname{Hess} f(q^{(k)}) U]^{-1} \operatorname{grad} f(q^{(k)})\|^2 \\
&\leq \frac{L_H}{2m_H^2} \left\|\operatorname{Hess} f(q^{(k)})\right\|^2 \Delta^2 = \frac{L_H M_H^2}{2m_H^2} \Delta^2.
\end{aligned}$$

Obviously, one can make the upper bound small by tuning down Δ. Combining the above lower bound for $\|\operatorname{grad} f(q)\|$ for $q \in R_{\text{II}}$, one can conclude that when Δ is small, the next iterate $q^{(k+1)}$ stays in R_{I}. Another application of the optimality condition (3.16) gives conditions on Δ that guarantees the next trust-region step is also unconstrained. Detailed argument can be found in proof of the following proposition.

Proposition 3.13 *Assume (3.18). W.h.p, once the trust-region algorithm takes an unconstrained R_{I} step (i.e., $\|\delta\| < \Delta$), it always takes unconstrained R_{I} steps, provided that*

$$\Delta \leq \frac{c c_\sharp^3 \theta^3 \mu}{n^{7/2} \log^{7/2}(np)}, \tag{3.23}$$

Here c is a positive numerical constant, and $c_\sharp$ is as defined in Lemma 3.9.

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_I step. Then

$$\|\boldsymbol{w}(\boldsymbol{q}^{(k+1)}) - \boldsymbol{w}(\boldsymbol{q}^{(k)})\| \leq \|\boldsymbol{q}^{(k+1)} - \boldsymbol{q}^{(k)}\| = \|\exp_{\boldsymbol{q}^{(k)}(\boldsymbol{\delta})} - \boldsymbol{q}^{(k)}\|$$
$$= \sqrt{2 - 2\cos\|\boldsymbol{\delta}\|} = 2\sin(\|\boldsymbol{\delta}\|/2) \leq \|\boldsymbol{\delta}\| < \Delta.$$

Thus, if $\Delta \leq \frac{1}{20\sqrt{5}} - \frac{\mu}{4\sqrt{2}}$, $\boldsymbol{q}^{(k+1)}$ will be in $R_I \cup R_{II}$. Next, we show that if Δ is sufficiently small, $\boldsymbol{q}^{(k+1)}$ will be indeed in R_I. By Lemma 3.12,

$$\left\|\text{grad } f\left(\boldsymbol{q}^{(k+1)}\right)\right\| \leq \frac{L_H}{2m_H^2}\left\|\text{grad } f\left(\boldsymbol{q}^{(k)}\right)\right\|^2$$
$$\leq \frac{L_H M_H^2}{2m_H^2}\left\|\left[\boldsymbol{U}^* \text{ Hess } f\left(\boldsymbol{q}^{(k)}\right)\boldsymbol{U}\right]^{-1}\boldsymbol{U}^* \text{ grad } f\left(\boldsymbol{q}^{(k)}\right)\right\|^2 \leq \frac{L_H M_H^2}{2m_H^2}\Delta^2, \quad (3.24)$$

where we have used the fact that

$$\left\|\boldsymbol{\delta}^{(k)}\right\| = \left\|\left[\boldsymbol{U}^* \text{ Hess } f\left(\boldsymbol{q}^{(k)}\right)\boldsymbol{U}\right]^{-1}\boldsymbol{U}^* \text{ grad } f\left(\boldsymbol{q}^{(k)}\right)\right\| < \Delta,$$

as the step is unconstrained. On the other hand, by Theorem 2.1 and Lemma 3.11, w.h.p.

$$\|\text{grad } f\left(\boldsymbol{q}\right)\| \geq \beta_{\text{grad}} \doteq \frac{9}{10}c_\star\theta, \quad \forall\, \boldsymbol{q} \in R_{II}. \tag{3.25}$$

Hence, provided

$$\Delta < \frac{m_H}{M_H}\sqrt{\frac{2\beta_{\text{grad}}}{L_H}}, \tag{3.26}$$

we have $\boldsymbol{q}^{(k+1)} \in R_I$.

We next show that when Δ is small enough, the next step is also unconstrained. Straight forward calculations give

$$\left\|\boldsymbol{U}\left[\boldsymbol{U}^* \text{ Hess } f\left(\boldsymbol{q}^{(k+1)}\right)\boldsymbol{U}\right]^{-1}\boldsymbol{U}^* \text{ grad } f\left(\boldsymbol{q}^{(k+1)}\right)\right\| \leq \frac{L_H M_H^2}{2m_H^3}\Delta^2.$$

Hence, provided that

$$\Delta < \frac{2m_H^3}{L_H M_H^2}, \tag{3.27}$$

we will have

$$\left\|\boldsymbol{U}\left[\boldsymbol{U}^* \text{ Hess } f\left(\boldsymbol{q}^{(k+1)}\right)\boldsymbol{U}\right]^{-1}\boldsymbol{U}^* \text{ grad } f\left(\boldsymbol{q}^{(k+1)}\right)\right\| < \Delta;$$

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 3.12 and Lemma 7.11, we have

$$L_H = C_1 n^{3/2}\log^{3/2}(np)/\mu^2, \tag{3.28}$$

33

w.h.p. for some numerical constant C_1. Combining this and our previous estimates of m_H, M_H, we conclude whenever

$$\Delta \leq \min\left\{\frac{1}{20\sqrt{5}} - \frac{\mu}{4\sqrt{2}}, \frac{c_1\mu c_\sharp c_\star^{1/2}\theta^{3/2}}{n^{7/4}\log^{7/4}(np)}, \frac{c_2\mu c_\sharp^3\theta^3}{n^{7/2}\log^{7/2}(np)}\right\}.$$

for some positive numerical constants c_1 and c_2, w.h.p. our next trust-region step is also an unconstrained R_{I} step. Noting that $c_\star$ and $c_\sharp$ can be made the same by our definition, we make the claimed simplification on Δ. This completes the proof. ∎

Finally, we want to show that ultimate unconstrained R_{I} iterates actually converges to one nearby local minimizer rapidly. Lemma 3.12 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 3.14 *Let $q_\star \in \mathbb{S}^{n-1}$ such that $\mathrm{grad}\, f(q_\star) = 0$, and $\delta \in T_{q_\star}\mathbb{S}^{n-1}$. Consider a geodesic $\gamma(t) = \exp_{q_\star}(t\delta)$, and suppose that on $[0,\tau]$, $\mathrm{Hess}\, f(\gamma(t)) \succeq m_H \mathcal{P}_{T_{\gamma(t)}\mathbb{S}^{n-1}}$. Then*

$$\|\mathrm{grad}\, f(\gamma(\tau))\| \geq m_H\tau\|\delta\|.$$

Proof See Page 83 under Section 8. ∎

To see this relates the magnitude of gradient to the distance away from the critical point, w.l.o.g., one can assume $\tau = 1$ and consider the point $q = \exp_{q_\star}(\delta)$. Then

$$\|q_\star - q\| = \|\exp_{q_\star}(\delta) - q\| = \sqrt{2 - 2\cos\|\delta\|} = 2\sin(\|\delta\|/2) \leq \|\delta\| \leq \|\mathrm{grad}\, f(q)\|/m_H,$$

where at the last inequality above we have used Lemma 3.14. Hence, combining this observation with Lemma 3.12, we can derive the asymptotic sequence convergence result as follows.

Proposition 3.15 *Assume (3.18) and the conditions in Lemma 3.13. Let $q^{(k_0)} \in R_{\mathrm{I}}$ and the k_0-th step the first unconstrained R_{I} step and $q_\star$ be the unique local minimizer of f over one connected component of R_{I} that contains $q^{(k_0)}$. Then w.h.p., for any positive integer $k' \geq 1$,*

$$\left\|q^{(k_0+k')} - q_\star\right\| \leq \frac{cc_\sharp\theta\mu}{n^{3/2}\log^{3/2}(np)}2^{-2^{k'}}, \tag{3.29}$$

provided that

$$\Delta \leq \frac{c'c_\sharp^2\theta^2\mu}{n^{5/2}\log^{5/2}(np)}. \tag{3.30}$$

Here $c_\sharp$ is as defined in Lemma 3.9 that can be made equal to $c_{s\star}$ as defined in Theorem 2.1, and c, c' are positive numerical constants.

Proof By the geometric characterization in Theorem 2.1 and corollary 2.2, f has $2n$ separated local minimizers, each located in R_{I} and within distance $\sqrt{2}\mu/16$ of one of the $2n$ signed basis vectors $\{\pm e_i\}_{i\in[n]}$. Moreover, it is obvious when $\mu \leq 1$, R_{I} consists of $2n$ disjoint connected components. We only consider the symmetric component in the vicinity of e_n and the claims carry on to others by symmetry.

34

Suppose that k_0 is the index of the first unconstrained iterate in region R_{I}, i.e., $\boldsymbol{q}^{(k_0)} \in R_{\mathrm{I}}$. By Lemma 3.12, for any integer $k' \geq 1$, we have

$$\left\| \operatorname{grad} f\left(\boldsymbol{q}^{(k_0+k')}\right) \right\| \leq \frac{2m_H^2}{L_H} \left(\frac{L_H}{2m_H^2} \left\| \operatorname{grad} f\left(\boldsymbol{q}^{(k_0)}\right) \right\| \right)^{2^{k'}}. \tag{3.31}$$

where L_H is as defined in Lemma 3.12, m_H as the strong convexity parameter for R_{I} defined above.

Now suppose $\boldsymbol{q}_\star$ is the unique local minimizer of f, lies in the same R_{I} component that $q^{(k_0)}$ is located. Let $\gamma_{k'}(t) = \exp_{\boldsymbol{q}_\star}(t\boldsymbol{\delta})$ to be the unique geodesic that connects $\boldsymbol{q}_\star$ and $\boldsymbol{q}^{(k_0+k')}$ with $\gamma_{k'}(0) = \boldsymbol{q}_\star$ and $\gamma_{k'}(1) = \boldsymbol{q}^{(k_0+k')}$. We have

$$\left\| \boldsymbol{q}^{(k_0+k')} - \boldsymbol{q}_\star \right\| \leq \left\| \exp_{\boldsymbol{q}_\star}(\boldsymbol{\delta}) - \boldsymbol{q}_\star \right\| = \sqrt{2 - 2\cos\|\boldsymbol{\delta}\|} = 2\sin(\|\boldsymbol{\delta}\|/2)$$

$$\leq \|\boldsymbol{\delta}\| \leq \frac{1}{m_H} \left\| \operatorname{grad} f\left(\boldsymbol{q}^{(k_0+k')}\right) \right\| \leq \frac{2m_H}{L_H} \left(\frac{L_H}{2m_H^2} \left\| \operatorname{grad} f\left(\boldsymbol{q}^{(k_0)}\right) \right\| \right)^{2^{k'}},$$

where at the second line we have repeatedly applied Lemma 3.14.

By the optimality condition (3.16) and the fact that $\left\| \boldsymbol{\delta}^{(k_0)} \right\| < \Delta$, we have

$$\frac{L_H}{2m_H^2} \left\| \operatorname{grad} f\left(\boldsymbol{q}^{(k_0)}\right) \right\| \leq \frac{L_H}{2m_H^2} M_H \left\| \left[\boldsymbol{U}^* \operatorname{Hess} f\left(\boldsymbol{q}^{(k_0)}\right) \boldsymbol{U} \right]^{-1} \boldsymbol{U}^* \operatorname{grad} f\left(\boldsymbol{q}^{(k_0)}\right) \right\| \leq \frac{L_H M_H}{2m_H^2} \Delta.$$

Thus, provided

$$\Delta < \frac{m_H^2}{L_H M_H}, \tag{3.32}$$

we can combine the above results and obtain

$$\left\| \boldsymbol{q}^{(k_0+k')} - \boldsymbol{q}_\star \right\| \leq \frac{2m_H}{L_H} 2^{-2^{k'}}.$$

Based on the previous estimates for m_H, M_H and L_H, we obtain that w.h.p.,

$$\left\| \boldsymbol{q}^{(k_0+k')} - \boldsymbol{q}_\star \right\| \leq \frac{c_1 c_\sharp \theta \mu}{n^{3/2} \log^{3/2}(np)} 2^{-2^{k'}}.$$

Moreover, by (3.32), w.h.p., it is sufficient to have the trust region size

$$\Delta \leq \frac{c_2 c_\sharp^2 \theta^2 \mu}{n^{5/2} \log^{5/2}(np)}.$$

Thus, we complete the proof. ∎

Now we are ready to piece together the above technical proposition to prove Theorem 3.1.
Proof [of Theorem 3.1] Assuming (3.18) and in addition that

$$\Delta < \min \left\{ \frac{c_1 c_\star \theta \mu^2}{n^{5/2} \log^{3/2}(np)}, \frac{c_2 c_\sharp^3 \theta^3 \mu}{n^{7/2} \log^{7/2}(np)} \right\}$$

for small enough numerical constants c_1 and c_2 and $c_\star$, $c_\sharp$ as defined in Theorem 2.1 and Lemma 3.9 respectively ($c_\star$ and $c_\sharp$ can be set to the same constant value), it can be verified that the conditions of all the above propositions are satisfied. Since each of the local minimizers is contained in the relative interior of one connected component of R_{I} (comparing distance of local minimizers to their respective signed basis vector, as stated in Corollary 2.2, with size of each connected R_{I} component yields this), we can define a threshold value

$$\zeta \doteq \min \left\{ \min_{\boldsymbol{q} \in \overline{R_{\mathrm{II}} \cup R_{\mathrm{III}}}} f(\boldsymbol{q}), \max_{\boldsymbol{q} \in \overline{R_{\mathrm{I}}}} f(\boldsymbol{q}) \right\}$$

where overline ¯ here denotes set closure. Obviously ζ is well-defined as the function f is continuous, and both sets $\overline{R_{\mathrm{II}} \cup R_{\mathrm{III}}}$ and $\overline{R_{\mathrm{I}}}$ are compact. Also for any of the local minimizers, say $\boldsymbol{q}_\star$, it holds that $\zeta > f(\boldsymbol{q}_\star)$.

By the four propositions above, 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 one local minimizer quadratically. Hence, regardless 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, *in a finite number of steps, the function value must drops below ζ and all future iterates stay in R_{I}*. Indeed, if the function value never drops below ζ, by continuity the whole sequence must be of entirely Type A - whereby either the finite-length sequence converges to one local minimizer, or every iterate of the infinite sequence steadily decreases the objective value by at least a *fixed* amount - in either case, the objective value should ever drop below ζ in finitely many steps; hence contradiction arises. Once the function value drops below ζ, type A future steps decreases the objective value further down below ζ - by definition of ζ, these iterates stay within R_{I}, and type B future steps, aka unconstrained R_{I} steps obviously keep all subsequent iterates in R_{I}.

There are three possibilities after the objective value drop below ζ and all future iterates stay in R_{I}. Assume $\boldsymbol{q}_\star$ is the unique local minimizer in the same connected component of R_{I} as the current iterate: (1) the sequence always take constrained R_{I} steps and hits $\boldsymbol{q}_\star$ exactly in finitely many steps; (2) the sequence takes constrained R_{I} steps until reaching certain point $\boldsymbol{q}' \in R_{\mathrm{I}}$ such that $f(\boldsymbol{q}') < f(\boldsymbol{q}_\star) + d_{\mathrm{I}}$, where d_{I} is as defined in Proposition 3.10. Since each unconstrained R_{I} step must decrease the objective value by at least d_{I}, the next and all future steps must be unconstrained R_{I} steps and the sequence converges to $\boldsymbol{q}_\star$; (3) the sequence starts to take unconstrained R_{I} steps at a certain point $\boldsymbol{q}'' \in R_{\mathrm{I}}$ such that $f(\boldsymbol{q}'') \geq f(\boldsymbol{q}_\star) + d_{\mathrm{I}}$. In any case, the sequence converges to the local minimizer $\boldsymbol{q}_\star$. By Proposition 3.7, Proposition 3.10, and Proposition 3.15, the number of iterations to obtain an ε-near solution to $\boldsymbol{q}_\star$ can be grossly bounded by

$$
\begin{aligned}
\#\text{Iter} \ &\leq \ \frac{f(\boldsymbol{q}^{(0)}) - f(\boldsymbol{q}_\star)}{\min\{d_{\mathrm{I}}, d_{\mathrm{II}}, d_{\mathrm{III}}\}} \ + \ \log\log\left(\frac{c_5 c_\sharp \theta \mu}{\varepsilon n^{3/2} \log^{3/2}(np)}\right) \\
&\leq \ \left[\min\left\{\frac{c_3 c_\star^3 \theta^3 \mu^4}{n^6 \log^3(np)}, \frac{c_4 c_\sharp^2 \theta^2}{n}\Delta^2\right\}\right]^{-1} \left(f(\boldsymbol{q}^{(0)}) - f(\boldsymbol{q}_\star)\right) \ + \ \log\log\left(\frac{c_5 c_\sharp \theta \mu}{\varepsilon n^{3/2} \log^{3/2}(np)}\right),
\end{aligned}
$$

where we have assumed $p \leq \exp(n)$ when comparing the various bounds. Finally, the claimed failure probability comes from a simple union bound with careful bookkeeping. ∎

3.4 Extending to Convergence for Complete Dictionaries

Note that for any complete A_0 with condition number $\kappa(A_0)$, from Lemma 2.14 we know when p is large enough, w.h.p. one can write the preconditioned $\overline{Y}$ as

$$\overline{Y} = UV^*X_0 + \Xi X_0$$

for a certain Ξ with small magnitude, and $U\Sigma V^* = \texttt{SVD}(A_0)$. Since UV^* is orthogonal,

$$f(q; UV^*X_0 + \Xi X_0) = f(VU^*q; X_0 + VU^*\Xi X_0).$$

In words, the function landscape of $f(q; UV^*X_0 + \Xi X_0)$ is a rotated version of that of $f(q; X_0 + VU^*\Xi X_0)$. Thus, any local minimizer $q_\star$ of $f(q; X_0 + VU^*\Xi X_0)$ is rotated to $UV^*q_\star$, 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 $f(q; X_0 + VU^*\Xi X_0)$ upon initialization q_0, it will generate the iteration sequence $UV^*q_0, UV^*q_1, UV^*q_2, \ldots$ for $f(q; UV^*X_0 + \Xi X_0)$. So w.l.o.g. it is adequate that we prove the convergence results for the case $f(q; X_0 + VU^*\Xi X_0)$, corresponding to $A_0 = I$ with perturbation $\widetilde{\Xi} \doteq VU^*\Xi$. So in this section (Section 3.4), we write $f(q; \widetilde{X_0})$ to mean $f(q; X_0 + \widetilde{\Xi}X_0)$.

Theorem 2.3 has shown that when

$$\theta \in \left(0, \frac{1}{2}\right), \ \mu \leq \min\left\{\frac{c_a\theta}{n}, \frac{c_b}{n^{5/4}}\right\}, \ p \geq \frac{C}{c_\star^2\theta}\max\left\{\frac{n^4}{\mu^4}, \frac{n^5}{\mu^2}\right\}\kappa^8(A_0)\log^4\left(\frac{\kappa(A_0)n}{\mu\theta}\right), \quad (3.33)$$

the geometric structure of the landscape is qualitatively unchanged and the $c_\star$ constant can be replaced with $c_\star/2$. Particularly, for this choice of p, Lemma 2.14 implies

$$\|\widetilde{\Xi}\| = \|VU^*\Xi\| \leq \left\|\widetilde{\Xi}\right\| \leq cc_\star\theta\left(\max\left\{\frac{n^{3/2}}{\mu^2}, \frac{n^2}{\mu}\right\}\log^{3/2}(np)\right)^{-1} \quad (3.34)$$

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 $\widetilde{\cdot}$ to denote the corresponding perturbed quantities here.

- Lemma 3.3: Note that

$$\|X_0 + \widetilde{\Xi}X_0\|_\infty \leq \|X_0\|_\infty + \|\widetilde{\Xi}X_0\|_\infty \leq \|X_0\|_\infty + \sqrt{n}\|\widetilde{\Xi}\|\|X_0\|_\infty \leq 3\|X_0\|_\infty/2,$$

  where by (3.34) we have used $\|\widetilde{\Xi}\| \leq 1/(2\sqrt{n})$ to simplify the above result. So we obtain

$$\widetilde{M_\nabla} \leq \frac{3}{2}M_\nabla, \ \widetilde{M_{\nabla^2}} \leq \frac{9}{4}M_{\nabla^2}, \ \widetilde{L_\nabla} \leq \frac{9}{4}L_\nabla, \ \widetilde{L_{\nabla^2}} \leq \frac{27}{8}L_{\nabla^2}.$$

- Lemma 3.4: Now we have

$$\widetilde{\eta}_f \doteq \widetilde{M_\nabla} + 2\widetilde{M_{\nabla^2}} + \widetilde{L_\nabla} + \widetilde{L_{\nabla^2}} \leq 4\eta_f.$$

- Lemma 3.5 and Lemma 3.6 are generic and nothing changes.

- Proposition 3.7: We have now $\boldsymbol{w}^* \boldsymbol{g}(\boldsymbol{w})/\|\boldsymbol{w}\| \geq c_\star \theta/2$ by Theorem 2.3 and w.h.p. $\boldsymbol{w}^* \nabla g(\boldsymbol{w})/\|\boldsymbol{w}\|$ is $C_1 n^2 \log(np)/\mu$-Lipschitz by Proposition 2.12 and the fact $\left\|\boldsymbol{X}_0 + \widetilde{\boldsymbol{\Xi}} \boldsymbol{X}_0\right\|_\infty \leq 3 \|\boldsymbol{X}_0\|_\infty /2$ shown above. Similarly, $\boldsymbol{w}^* \boldsymbol{g}(\boldsymbol{w})/\|\boldsymbol{w}\| \leq -c_\star \theta/2$ by Theorem 2.3 and $\boldsymbol{w}^* \nabla^2 g(\boldsymbol{w})\boldsymbol{w}/\|\boldsymbol{w}\|^2$ is $C_2 n^3 \log^{3/2}(np)/\mu^2$-Lipschitz. Moreover, $\widetilde{\eta}_f \leq 4\eta_f$ as shown above. Since there are only multiplicative constant changes to the various quantities, we conclude

$$\widetilde{d_{\mathrm{II}}} = c_1 d_{\mathrm{II}}, \quad \widetilde{d_{\mathrm{III}}} = c_1 d_{\mathrm{III}} \tag{3.35}$$

provided

$$\Delta < \frac{c_2 c_\star \theta \mu^2}{n^{5/2} \log^{3/2}(np)}. \tag{3.36}$$

- Lemma 3.8: η_f is changed to $\widetilde{\eta}_f$ with $\widetilde{\eta}_f \leq 4\eta_f$ as shown above.

- Lemma 3.9: By (3.13), we have

$$\left\| \nabla^2 f(\boldsymbol{q}; \boldsymbol{X}_0) - \nabla^2 f(\boldsymbol{q}; \widetilde{\boldsymbol{X}_0}) \right\| \leq \frac{1}{p} \sum_{k=1}^{p} \left\{ L_{\ddot{h}} \|\widetilde{\boldsymbol{\Xi}}\| \|\boldsymbol{x}_k\|^2 + \frac{1}{\mu} \|\boldsymbol{x}_k \boldsymbol{x}_k^* - \widetilde{\boldsymbol{x}}_k \widetilde{\boldsymbol{x}}_k^*\| \right\}$$

$$\leq \|\widetilde{\boldsymbol{\Xi}}\| \left(L_{\ddot{h}} + 2/\mu + \|\widetilde{\boldsymbol{\Xi}}\|/\mu \right) \sum_{k=1}^{p} \|\boldsymbol{x}_k\|^2 \leq \|\widetilde{\boldsymbol{\Xi}}\| \left(L_{\ddot{h}} + 3/\mu \right) n \|\boldsymbol{X}_0\|_\infty^2,$$

where $L_{\ddot{h}}$ is the Lipschitz constant for the function $\ddot{h}_\mu(\cdot)$ and we have used the fact that $\|\widetilde{\boldsymbol{\Xi}}\| \leq 1$. Similarly, by 3.12,

$$\left\| \nabla f(\boldsymbol{q}; \boldsymbol{X}_0) - \nabla f(\boldsymbol{q}; \widetilde{\boldsymbol{X}_0}) \right\| \leq \frac{1}{p} \sum_{k=1}^{p} \left\{ L_{\dot{h}_\mu} \|\widetilde{\boldsymbol{\Xi}}\| \|\boldsymbol{x}_k\| + \|\widetilde{\boldsymbol{\Xi}}\| \|\boldsymbol{x}_k\| \right\} \leq \left(L_{\dot{h}_\mu} + 1 \right) \|\widetilde{\boldsymbol{\Xi}}\| \sqrt{n} \|\boldsymbol{X}_0\|_\infty,$$

where $L_{\dot{h}}$ is the Lipschitz constant for the function $\dot{h}_\mu(\cdot)$. Since $L_{\ddot{h}} \leq 2/\mu^2$ and $L_{\dot{h}} \leq 1/\mu$, and $\|\boldsymbol{X}_0\|_\infty \leq 4\sqrt{\log(np)}$ w.h.p. (Lemma 7.11). By (3.34), w.h.p. we have

$$\left\| \nabla f(\boldsymbol{q}; \boldsymbol{X}_0) - \nabla f(\boldsymbol{q}; \widetilde{\boldsymbol{X}_0}) \right\| \leq \frac{1}{2} c_\sharp \theta, \quad \text{and} \quad \left\| \nabla^2 f(\boldsymbol{q}; \boldsymbol{X}_0) - \nabla^2 f(\boldsymbol{q}; \widetilde{\boldsymbol{X}_0}) \right\| \leq \frac{1}{2} c_\sharp \theta,$$

provided the constant C in (3.33) for p is large enough. Thus, by (3.15) and the above estimates we have

$$\left\| \operatorname{Hess} f(\boldsymbol{q}; \boldsymbol{X}_0) - \operatorname{Hess} f(\boldsymbol{q}; \widetilde{\boldsymbol{X}_0}) \right\| \leq \left\| \nabla f(\boldsymbol{q}; \boldsymbol{X}_0) - \nabla f(\boldsymbol{q}; \widetilde{\boldsymbol{X}_0}) \right\| + \left\| \nabla^2 f(\boldsymbol{q}; \boldsymbol{X}_0) - \nabla^2 f(\boldsymbol{q}; \widetilde{\boldsymbol{X}_0}) \right\|$$

$$\leq c_\sharp \theta \leq \frac{1}{2} c_\sharp \frac{\theta}{\mu},$$

provided $\mu \leq 1/2$. So we conclude

$$\operatorname{Hess} f(\boldsymbol{q}; \widetilde{\boldsymbol{X}_0}) \succeq \frac{1}{2} c_\sharp \frac{\theta}{\mu} \mathcal{P}_{T_{\boldsymbol{q}}} \mathbb{S}^{n-1} \implies \widetilde{m_H} \geq \frac{1}{2} c_\sharp \frac{\theta}{\mu}. \tag{3.37}$$

- Proposition 3.10: From the estimate of M_H above Proposition 3.10 and the last point, we have

$$\left\| \text{Hess } f(\boldsymbol{q}; \widetilde{\boldsymbol{X}}_0) \right\| \leq \frac{36}{\mu} \log(np), \quad \text{and} \quad \text{Hess } f(\boldsymbol{q}; \widetilde{\boldsymbol{X}}_0) \succeq \frac{1}{2} c_\sharp \frac{\theta}{\mu} \mathcal{P}_{T_{\boldsymbol{q}}} \mathbb{S}^{n-1}.$$

Also since $\widetilde{\eta}_f \leq 4\eta_f$ in Lemma 3.4 and Lemma 3.8, there are only multiplicative constant change to the various quantities. We conclude that

$$\widetilde{d}_{\text{I}} = c_3 d_{\text{I}} \tag{3.38}$$

provided that

$$\Delta \leq \frac{c_4 c_\sharp^2 \theta^2 \mu}{n^{5/2} \log^{5/2}(np)}. \tag{3.39}$$

- Lemma 3.11 is generic and nothing changes.

- Lemma 3.12: $\widetilde{L}_H \leq 27 L_H / 8$.

- Proposition 3.13: All the quantities involved in determining Δ, m_H, M_H, and L_H, β_{grad} are modified by at most constant multiplicative factors and changed to their respective tilde version, so we conclude that the RTM algorithm always takes unconstrained R_{I} step after taking one, provided that

$$\Delta \leq \frac{c_5 c_\sharp^3 \theta^3 \mu}{n^{7/2} \log^{7/2}(np)}. \tag{3.40}$$

- Lemma 3.14:is generic and nothing changes.

- Proposition 3.15: Again m_H, M_H, L_H are changed to $\widetilde{m_H}$, $\widetilde{M_H}$, and $\widetilde{L_H}$, respectively, differing by at most constant multiplicative factors. So we conclude for any integer $k' \geq 1$,

$$\left\| \boldsymbol{q}^{(k_0 + k')} - \boldsymbol{q}_\star \right\| \leq \frac{c_6 c_\sharp \theta \mu}{n^{3/2} \log^{3/2}(np)} 2^{-2^{k'}}, \tag{3.41}$$

provided

$$\Delta \leq \frac{c_7 c_\sharp^2 \theta^2 \mu}{n^{5/2} \log^{5/2}(np)}. \tag{3.42}$$

The final proof to Theorem 2.3 is almost identical to that of Theorem 2.1, except for

$$\Delta \leq \min \left\{ \frac{c_8 c_\star \theta \mu^2}{n^{5/2} \log^{3/2}(np)}, \frac{c_9 c_\sharp^3 \theta^3 \mu}{n^{7/2} \log^{7/2}(np)} \right\}, \tag{3.43}$$

$$\widetilde{\zeta} \doteq \min \left\{ \min_{\boldsymbol{q} \in R_{\text{II}} \cup R_{\text{III}}} f\left(\boldsymbol{q}; \widetilde{\boldsymbol{X}}_0\right), \max_{\boldsymbol{q} \in R_{\text{I}}} f\left(\boldsymbol{q}; \widetilde{\boldsymbol{X}}_0\right) \right\}, \tag{3.44}$$

39

and hence all ζ is now changed to $\widetilde{\zeta}$, and also d_{I}, d_{II}, and d_{III} are changed to $\widetilde{d_{\mathrm{I}}}$, $\widetilde{d_{\mathrm{II}}}$, and $\widetilde{d_{\mathrm{III}}}$ as defined above, respectively. The final iteration complexity to each an ε-near solution is hence

$$\#\text{Iter} \leq \left[\min \left\{ \frac{c_{10} c_\star^3 \theta^3 \mu^4}{n^6 \log^3(np)}, \frac{c_{11} c_\sharp^2 \theta^2}{n} \Delta^2 \right\} \right]^{-1} \left(f\left(\boldsymbol{q}^{(0)}\right) - f\left(\boldsymbol{q}_\star\right) \right) \; + \; \log\log\left(\frac{c_{12} c_\sharp \theta \mu}{\varepsilon n^{3/2} \log^{3/2}(np)} \right).$$

Hence overall the qualitative behavior of the algorithm is not changed, as compared to that for the orthogonal case. Above c_1 through c_{12} are all numerical constants.

4   Complete Algorithm Pipeline and Main Results

For orthogonal dictionaries, from Theorem 2.1 and its corollary, we know that all the minimizers $\widehat{\boldsymbol{q}}_\star$ are $O(\mu)$ away from their respective nearest "target" $\boldsymbol{q}_\star$, with $\boldsymbol{q}_\star^* \widehat{\boldsymbol{Y}} = \alpha \boldsymbol{e}_i^* \boldsymbol{X}_0$ for certain $\alpha \neq 0$ and $i \in [n]$; in Theorem 3.1, we have shown that w.h.p. the Riemannian TRM algorithm produces a solution $\widehat{\boldsymbol{q}} \in \mathbb{S}^{n-1}$ that is ε away to one of the minimizers, say $\widehat{\boldsymbol{q}}_\star$. Thus, the $\widehat{\boldsymbol{q}}$ returned by the TRM algorithm is $O(\varepsilon + \mu)$ away from $\boldsymbol{q}_\star$. For exact recovery, we use a simple linear programming rounding procedure, which guarantees to exactly produce the optimizer $\boldsymbol{q}_\star$. We then use deflation to sequentially recover other rows of $\boldsymbol{X}_0$. Overall, w.h.p. both the dictionary $\boldsymbol{A}_0$ and sparse coefficient $\boldsymbol{X}_0$ 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 4.1. The same procedure can be used to recover complete dictionaries, though the analysis is slightly more complicated; we present the results in Section 4.2. Our overall algorithmic pipeline for recovering orthogonal dictionaries is sketched as follows.

1. **Estimating one row of $\boldsymbol{X}_0$ by the Riemannian TRM algorithm.** By Theorem 2.1 (resp. Theorem 2.3) and Theorem 3.1 (resp. Theorem 3.2), starting from any, 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 $\widehat{\boldsymbol{q}}$, with $\boldsymbol{q}_\star$ the nearest target that exactly recovers one row of $\boldsymbol{X}_0$ and $\|\widehat{\boldsymbol{q}} - \boldsymbol{q}_\star\| \in O(\mu)$ (by setting the target accuracy of the TRM as, say, $\varepsilon = \mu$).

2. **Recovering one row of $\boldsymbol{X}_0$ by rounding.** To obtain the target solution $\boldsymbol{q}_\star$ and hence recover (up to scale) one row of $\boldsymbol{X}_0$, we solve the following linear program:

$$\text{minimize}_{\boldsymbol{q}} \left\| \boldsymbol{q}^* \widehat{\boldsymbol{Y}} \right\|_1, \quad \text{subject to} \quad \langle \boldsymbol{r}, \boldsymbol{q} \rangle = 1, \tag{4.1}$$

   with $\boldsymbol{r} = \widehat{\boldsymbol{q}}$. We show in Lemma 4.2 (resp. Lemma 4.4) that when $\langle \widehat{\boldsymbol{q}}, \boldsymbol{q}_\star \rangle$ is sufficiently large, implied by μ being sufficiently small, w.h.p. the minimizer of (4.1) is exactly $\boldsymbol{q}_\star$, and hence one row of $\boldsymbol{X}_0$ is recovered by $\boldsymbol{q}_\star^* \widehat{\boldsymbol{Y}}$.

3. **Recovering all rows of $\boldsymbol{X}_0$ by deflation.** Once ℓ rows of $\boldsymbol{X}_0$ ($1 \leq \ell \leq n-2$) have been recovered, say, by unit vectors $\boldsymbol{q}_\star^1, \ldots, \boldsymbol{q}_\star^\ell$, one takes an orthonormal basis $\boldsymbol{U}$ for $[\text{span}\left(\boldsymbol{q}_\star^1, \ldots, \boldsymbol{q}_\star^\ell\right)]^\perp$, and minimizes the new function $h(\boldsymbol{z}) \doteq f(\boldsymbol{U}\boldsymbol{z}; \widehat{\boldsymbol{Y}})$ on the sphere $\mathbb{S}^{n-\ell-1}$ with the Riemannian TRM algorithm (though conservative, one can again set parameters as $\mu_\star$, $\Delta_\star$, as in Step 1) to produce a $\widehat{\boldsymbol{z}}$. Another row of $\boldsymbol{X}_0$ is then recovered via the LP rounding (4.1) with input $\boldsymbol{r} = \boldsymbol{U}\widehat{\boldsymbol{z}}$ (to produce $\boldsymbol{q}_\star^{\ell+1}$). Finally, by repeating the procedure until depletion, one can recover all the rows of $\boldsymbol{X}_0$.

4. **Reconstructing the dictionary A_0.** By solving the linear system $Y = AX_0$, one can obtain the dictionary $A_0 = YX_0^* (X_0 X_0^*)^{-1}$.

4.1 Recovering Orthogonal Dictionaries

Theorem 4.1 (Main theorem - recovering orthogonal dictionaries) *Assume the dictionary A_0 is orthogonal and we take $\widehat{Y} = Y$. Suppose $\theta \in (0, 1/3)$, $\mu_\star < \min\left\{c_a \theta n^{-1}, c_b n^{-5/4}\right\}$, and $p \geq Cn^3 \log \frac{n}{\mu_\star \theta} / \left(\mu_\star^2 \theta^2\right)$. The above algorithmic pipeline with parameter setting*

$$\Delta_\star \leq \min\left\{\frac{c_c c_\star \theta \mu_\star^2}{n^{5/2} \log^{5/2}(np)}, \frac{c_d c_\star^3 \theta^3 \mu_\star}{n^{7/2} \log^{7/2}(np)}\right\}, \tag{4.2}$$

recovers the dictionary A_0 and X_0 in polynomial time, with failure probability bounded by $c_e p^{-6}$. Here $c_\star$ is as defined in Theorem 2.1, and c_a through c_e, and C are all positive numerical constants.

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

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

Lemma 4.2 (LP rounding - orthogonal dictionary) *There exists a positive constant C, such that for all $\theta \in (0, 1/3)$, and $p \geq Cn^2 \log(n/\theta)/\theta$, with probability at least $1 - 2p^{-10} - \theta(n-1)^{-7} p^{-7} - \exp\left(-0.3\theta(n-1)p\right)$, the rounding procedure (4.1) returns $q_\star$ for any input vector r that satisfies*

$$\langle r, q_\star \rangle \geq 249/250.$$

Proof See Page 85 under Section 9. ∎

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

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 $\{q_\star^i\}_{i=1}^n$ are orthogonal to each other. W.l.o.g., suppose we have found the first ℓ unit vectors $q_\star^1, \ldots, q_\star^\ell$ which recover the first ℓ rows of X_0. Correspondingly, we partition the target dictionary A_0 and X_0 as

$$A_0 = [V, V^\perp], \quad X_0 = \begin{bmatrix} X_0^{[\ell]} \\ X_0^{[n-\ell]} \end{bmatrix}, \tag{4.3}$$

where $V \in \mathbb{R}^{n \times \ell}$, and $X_0^{[\ell]} \in \mathbb{R}^{\ell \times n}$ denotes the submatrix with the first ℓ rows of X_0. Let us define a function: $f_{n-\ell}^\downarrow : \mathbb{R}^{n-\ell} \mapsto \mathbb{R}$ by

$$f_{n-\ell}^\downarrow(z; W) \doteq \frac{1}{p} \sum_{k=1}^p h_\mu(z^* w_k), \tag{4.4}$$

41

for any matrix $W \in \mathbb{R}^{(n-\ell) \times p}$. Then by (1.4), our objective function is equivalent to

$$h(z) = f(Uz; A_0 X_0) = f_{n-\ell}^{\downarrow}(z; U^* A_0 X_0) = f_{n-\ell}^{\downarrow}(z; U^* V X_0^{[\ell]} + U^* V^{\perp} X_0^{[n-\ell]}).$$

Since the columns of the orthogonal matrix $U \in \mathbb{R}^{n \times (n-\ell)}$ forms the orthogonal complement of span $\left(q_{\star}^1, \cdots, q_{\star}^{\ell} \right)$, it is obvious that $U^* V = 0$. Therefore, we obtain

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

Since $U^* V^{\perp}$ is orthogonal and $X_0^{[n-\ell]} \sim_{i.i.d.} \mathrm{BG}(\theta)$, 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 4.1, the conditions of Theorem 2.1 and Theorem 3.1 for all cases with reduced dimensions are still valid. So w.h.p., the TRM algorithm returns a $\widehat{z}$ such that $\|\widehat{z} - z_{\star}\| \in O(\mu_{\star})$ where $z_{\star}$ is a "target" solution that recovers a row of X_0:

$$z_{\star}^* U^* V^{\perp} X_0^{[n-\ell]} = z_{\star}^* U^* A_0 X_0 = \alpha e_i^* X_0, \quad \text{for some } i \notin [\ell].$$

So pulling everything back in the original space, the effective target is $q_{\star}^{\ell+1} \doteq U z_{\star}$, and $U \widehat{z}$ is our estimation obtained from the TRM algorithm. Moreover,

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

Thus, by Lemma 4.2, one successfully recovers $U z_{\star}$ from $U \widehat{z}$ 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.

4.2   Recovering Complete Dictionaries

By working with the preconditioned data samples $\widehat{Y} = \overline{Y} \doteq \sqrt{\theta p} \left(Y Y^* \right)^{-1/2} Y$,[28] we can use a similar procedure described above to recover complete dictionaries.

Theorem 4.3 (Main theorem - recovering complete dictionaries) *Assume the dictionary A_0 is complete with condition number $\kappa(A_0)$ and we take $\widehat{Y} = \overline{Y}$. Suppose $\theta \in (0, 1/3)$, $\mu_{\star} < \min \left\{ c_a \theta n^{-1}, c_b n^{-5/4} \right\}$, and $p \geq \frac{C}{c_{\star}^2 \theta} \max \left\{ \frac{n^4}{\mu^4}, \frac{n^5}{\mu^2} \right\} \kappa^8(A_0) \log^4 \left(\frac{\kappa(A_0) n}{\mu \theta} \right)$. The algorithmic pipeline with parameter setting*

$$\Delta_{\star} \leq \min \left\{ \frac{c_c c_{\star} \theta \mu_{\star}^2}{n^{5/2} \log^{5/2}(np)}, \frac{c_d c_{\star}^3 \theta^3 \mu_{\star}}{n^{7/2} \log^{7/2}(np)} \right\}, \tag{4.5}$$

recovers the dictionary A_0 and X_0 in polynomial time, with failure probability bounded by $c_e p^{-6}$. Here $c_{\star}$ is as defined in Theorem 2.1, and c_a through c_f, and C are all positive numerical constants.

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

---

[28] In practice, the parameter θ might not be know beforehand. However, because it only scales the problem, it does not affect the overall qualitative aspect of results.

Proof of LP rounding The result of the LP rounding is only slightly different from that of the orthogonal case in Lemma 4.2, so is the proof.

Lemma 4.4 (LP rounding - complete dictionary) *There exists a positive constant C, such that for all $\theta \in (0, 1/3)$, and $p \geq \frac{C}{c_\star^2 \theta} \max \left\{ \frac{n^4}{\mu^4}, \frac{n^5}{\mu^2} \right\} \kappa^8 (A_0) \log^4 \left(\frac{\kappa(A_0)n}{\mu\theta} \right)$, with probability at least $1 - 3p^{-8} - \theta(n - 1)^{-7}p^{-7} - \exp(-0.3\theta(n - 1)p)$, the rounding procedure (4.1) returns $q_\star$ for any input vector r that satisfies*

$$\langle r, q_\star \rangle \geq 249/250.$$

Proof See Page 86 under Section 9. ∎

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 $\{q_\star^i\}_{i=1}^n$ are not necessarily orthogonal to each other, but they are almost so. W.l.o.g., let us again assume that $q_\star^1, \ldots, q_\star^\ell$ recover the first ℓ rows of X_0, and similarly partition the matrix X_0 as in (4.3).

By Lemma 2.14 and (2.15), we can write $\overline{Y} = (Q + \Xi)X_0$ for some orthogonal matrix Q and small perturbation Ξ with $\|\Xi\| \leq \delta < 1/10$ for some large p as usual. Similar to the orthogonal case, we have

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

where $f_{n-\ell}^\downarrow$ is defined the same as in (4.4). Next, we show that the matrix $U^*(Q + \Xi)X_0$ can be decomposed as $U^*VX_0^{[n-\ell]} + \Delta$, where $V \in \mathbb{R}^{(n-\ell)\times n}$ is orthogonal and Δ is a small perturbation matrix. More specifically, we show that

Lemma 4.5 *Suppose the matrices $U \in \mathbb{R}^{n\times(n-\ell)}$, $Q \in \mathbb{R}^{n\times n}$ are orthogonal as defined above, Ξ is a perturbation matrix with $\|\Xi\| \leq 1/20$, then*

$$U^* (Q + \Xi) X_0 = U^*VX_0^{[n-\ell]} + \Delta, \tag{4.6}$$

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

$$\|\Delta\|_{\ell^1 \to \ell^2} \leq 16\sqrt{n} \|\Xi\| \|X_0\|_\infty, \quad \|\Delta\| \leq 16 \|\Xi\| \|X_0\|, \tag{4.7}$$

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

Proof See Page 87 under Section 9. ∎

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

$$h(z) = f_{n-\ell}^\downarrow(z; U^*VX_0^{[n-\ell]} + \Delta).$$

Since our perturbation analysis in proving Theorem 2.3 and Theorem 3.2 solely relies on the fact that $\|\Delta\|_{\ell^1 \to \ell^2} \leq C \|\Xi\| \sqrt{n} \|X_0\|_\infty$, it is enough to make p large enough so that the theorems are still applicable for the reduced version $f_{n-\ell}^\downarrow(z; U^*VX_0^{[n-\ell]} + \Delta)$. Thus, by invoking Theorem 2.3

43

and Theorem 3.2, the TRM algorithm provably returns one $\widehat{z}$ such that $\widehat{z}$ is near to a perturbed optimal $\widehat{z}_\star$ with

$$\widehat{z}_\star^* U^* V X_0^{[n-\ell]} = z_\star^* U^* V X_0^{[n-\ell]} + z_\star^* \Delta = \alpha e_i^* X_0, \quad \text{for some } i \notin [\ell], \tag{4.8}$$

where $z_\star$ with $\|z_\star\| = 1$ is the exact solution. More specifically, Corollary 2.4 implies

$$\|\widehat{z} - \widehat{z}_\star\| \leq \sqrt{2}\mu_\star/7.$$

Next, we show that $\widehat{z}$ is also very near to the exact solution $z_\star$. Indeed, the identity (4.8) suggests

$$(\widehat{z}_\star - z_\star)^* U^* V X_0^{[n-\ell]} = z_\star^* \Delta$$
$$\Longrightarrow \widehat{z}_\star - z_\star = \left[(X_0^{[n-\ell]})^* V^* U \right]^\dagger \Delta^* z_\star = U^* V \left[(X_0^{[n-\ell]})^* \right]^\dagger \Delta^* z_\star \tag{4.9}$$

where $W^\dagger = (W^* W)^{-1} W^*$ denotes the pseudo inverse of a matrix W with full column rank. Hence, by (4.9) we can bound the distance between $\widehat{z}_\star$ and $z_\star$ by

$$\|\widehat{z}_\star - z_\star\| \leq \left\| \left[(X_0^{[n-\ell]})^* \right]^\dagger \right\| \|\Delta\| \leq \sigma_{\min}^{-1}(X_0^{[n-\ell]}) \|\Delta\|$$

By Lemma B.3, when $p \geq \Omega(n^2 \log n)$, w.h.p.,

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

Hence, combined with Lemma 4.5, we obtain

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

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

$$\|\widehat{z} - z_\star\| \leq \|\widehat{z} - \widehat{z}_\star\| + \|\widehat{z}_\star - z_\star\| \leq \sqrt{2}\mu_\star/7 + 28 \|\Xi\|.$$

Lemma 2.14, and in particular (2.15), for our choice of p as in Theorem 2.3, $\|\Xi\| \leq c\mu_\star^2 n^{-3/2}$, where c can be made smaller by making the constant in p larger. For $\mu_\star$ sufficiently small, we conclude that

$$\|U\widehat{z} - U z_\star\| = \|\widehat{z} - z_\star\| \leq 2\mu_\star/7.$$

In words, the TRM algorithm returns a $\widehat{z}$ such that $U\widehat{z}$ is very near to one of the unit vectors $\{q_\star^i\}_{i=1}^n$, such that $(q_\star^i)^* \overline{Y} = \alpha e_i^* X_0$ for some $\alpha \neq 0$. For $\mu_\star$ smaller than a fixed constant, one will have

$$\langle U\widehat{z}, q_\star^i \rangle \geq 249/250,$$

and hence by Lemma 4.4, the LP rounding exactly returns the optimal solution $q_\star^i$ upon the input $U\widehat{z}$.

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.
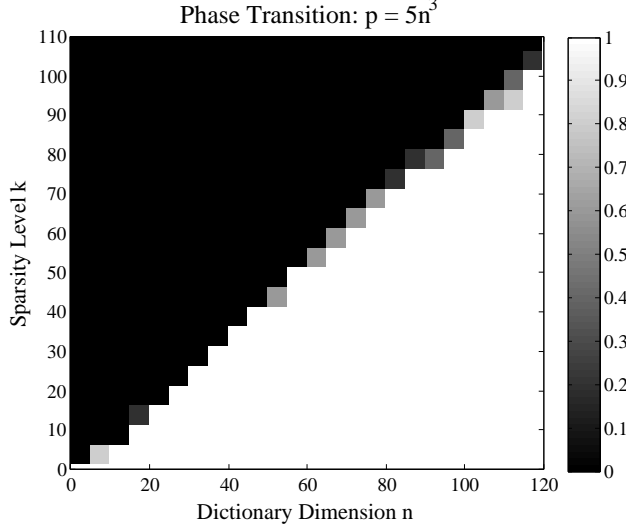
Figure 5: Phase transition for recovering a single sparse vector under the dictionary learning model with the sample complexity $p = 5n^3$

5 Simulations

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 $\boldsymbol{A}_0 = \boldsymbol{I} \in \mathbb{R}^{n \times n}$. We fix $p = 5n^3$, and each column of the coefficient matrix $\boldsymbol{X}_0 \in \mathbb{R}^{n \times p}$ 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 produce similar behavior. For the sparsity surrogate defined in (1.5), we fix the parameter $\mu = 10^{-2}$. We implement Algorithm 1 with adaptive step size instead of the fixed step size in our analysis.

To see how the allowable sparsity level varies with the dimension, which our theory primarily is about, we vary the dictionary dimension n and the sparsity k both between 1 and 120; for every pair of (k, n) we repeat the simulations independently for $T = 5$ times. Because the optimal solutions are signed coordinate vectors $\{\boldsymbol{e}_i\}_{i=1}^n$, for a solution $\widehat{\boldsymbol{q}}$ returned by the TRM algorithm, we define the reconstruction error (RE) to be

$$\mathtt{RE} = \min_{1 \le i \le n} \left(\|\widehat{\boldsymbol{q}} - \boldsymbol{e}_i\|, \|\widehat{\boldsymbol{q}} + \boldsymbol{e}_i\| \right). \tag{5.1}$$

The trial is determined to be a success once $\mathtt{RE} \le \mu$, with the idea that this indicates $\widehat{\boldsymbol{q}}$ 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). Figure 5 shows the phase transition in the (n, k) plane for the orthogonal case. It is obvious that our TRM algorithm can work well into the linear region whenever $p \in O(n^3)$. Our analysis is tight up to logarithm factors, and also the polynomial dependency on $1/\mu$, which under the theory is polynomial in n.

6 Discussion

For recovery of complete dictionaries, the LP program approach in [SWW12] that works with $\theta \leq O(1/\sqrt{n})$ only demands $p \geq \Omega(n^2 \log n^2)$, which is recently improved to $p \geq \Omega(n \log^4 n)$ [LV15], almost matching the lower bound $\Omega(n \log n)$ (i.e., when $\theta \sim 1/n$). The sample complexity stated in Theorem 4.3 is obviously much higher. It is interesting to see whether such growth in complexity is intrinsic to working in the linear regime. Though our experiments seemed to suggest the necessity of $p \sim O(n^3)$ even for the orthogonal case, there could be other efficient algorithms that demand much less. Tweaking these three points will likely improve the complexity: (1) The ℓ^1 proxy. The derivative and Hessians of the $\log \cosh$ function we adopted entail the $\tanh$ function, which is not amenable to effective approximation and affects the sample complexity; (2) Geometric characterization and algorithm analysis. It seems working directly on the sphere (i.e., in the q space) could simplify and possibly improve certain parts of the analysis; (3) treating the complete case directly, rather than using (pessimistic) bounds to treat it as a perturbation of the orthogonal case. Particularly, general linear transforms may change the space significantly, such that preconditioning and comparing to the orthogonal transforms may not be the most efficient way to proceed.

It is possible to extend the current analysis to other dictionary settings. Our geometric structures and algorithms allow plug-and-play noise analysis. Nevertheless, we believe a more stable way of dealing with noise is to directly extract the whole dictionary, i.e., to consider geometry and optimization (and perturbation) over the orthogonal group. This will require additional nontrivial technical work, but likely feasible thanks to the relatively complete knowledge of the orthogonal group [EAS98, AMS09]. A substantial leap forward would be to extend the methodology to recovery of *structured* overcomplete dictionaries, such as tight frames. Though there is no natural elimination of one variable, one can consider the marginalization of the objective function wrt the coefficients and work with hidden functions. [29] For the coefficient model, as we alluded to in Section 1.5, our analysis and results likely can be carried through to coefficients with statistical dependence and physical constraints.

The connection to ICA we discussed in Section 1.5 suggests our geometric characterization and algorithms can be modified for the ICA problem. This likely will provide new theoretical insights and computational schemes to ICA. In the surge of theoretical understanding of nonconvex heuristics [KMO10, JNS13, Har14, HW14, NNS+14, JN14, NJS13, CLS14, JO14, AGJ14, YCS13, LWB13, QSW14, LWB13, AAJ+13, AAN13, AGM13, AGMM15, ABGM14], the initialization plus local refinement strategy mostly differs from practice, whereby random initializations seem to work well, and the analytic techniques developed are mostly fragmented and highly specialized. The analytic and algorithmic we developed here hold promise to provide a coherent account of these problems. It is interesting to see to what extent we can streamline and generalize the framework.

Our motivating experiment on real images in Section 1.2 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 local maxima and saddle points (at least with high probability). The recent work [GHJY15] has showed that randomly perturbing each iterate can help gradient

---

[29]This recent work [AGMM15] on overcomplete DR has used a similar idea. The marginalization taken there is near to the global optimum of one variable, where the function is well-behaved. Studying the global properties of the marginalization may introduce additional challenges.

algorithm to escape saddle points with high probability. It would be interesting to know whether similar results can be obtained for gradient descent algorithms with random initialization. The continuous counterpart seems well understood; see, e.g., [HMG94] for discussions of Morse-Bott theorem and gradient flow convergence.

7 Proofs of Main Technical Results for High Dimensional Geometry

In this section, we provide complete proofs for technical results stated in Section 2. Before that, let us introduce some notations and common results that will be used later throughout this section. Since we deal with BG random variables and random vectors, it is often convenient to write such vector explicitly as $\boldsymbol{x} = [\Omega_1 v_1, \ldots, \Omega_n v_n] = \boldsymbol{\Omega} \odot \boldsymbol{v}$, where $\Omega_1, \ldots, \Omega_n$ are i.i.d. Bernoulli random variables and $v_1, \ldots, v_n$ are i.i.d. standard normal. For a particular realization of such random vector, we will denote the support as $\mathcal{I} \subset [n]$. Due to the particular coordinate map in use, we will often refer to subset $\mathcal{J} \doteq \mathcal{I} \setminus \{n\}$ and the random vectors $\overline{\boldsymbol{x}} \doteq [\Omega_1 v_1, \ldots, \Omega_{n-1} v_{n-1}]$ and $\overline{\boldsymbol{v}} \doteq [v_1, \ldots, v_{n-1}]$ in $\mathbb{R}^{n-1}$. By Lemma A.1, it is not hard to see that

$$\nabla_{\boldsymbol{w}} h_\mu \left(\boldsymbol{q}^* \left(\boldsymbol{w} \right) \boldsymbol{x} \right) = \tanh \left(\frac{\boldsymbol{q}^* \left(\boldsymbol{w} \right) \boldsymbol{x}}{\mu} \right) \left(\overline{\boldsymbol{x}} - \frac{x_n}{q_n \left(\boldsymbol{w} \right)} \boldsymbol{w} \right), \tag{7.1}$$

$$\nabla_{\boldsymbol{w}}^2 h_\mu \left(\boldsymbol{q}^* \left(\boldsymbol{w} \right) \boldsymbol{x} \right) = \frac{1}{\mu} \left[1 - \tanh^2 \left(\frac{\boldsymbol{q}^* \left(\boldsymbol{w} \right) \boldsymbol{x}}{\mu} \right) \right] \left(\overline{\boldsymbol{x}} - \frac{x_n}{q_n \left(\boldsymbol{w} \right)} \boldsymbol{w} \right) \left(\overline{\boldsymbol{x}} - \frac{x_n}{q_n \left(\boldsymbol{w} \right)} \boldsymbol{w} \right)^*$$

$$- x_n \tanh \left(\frac{\boldsymbol{q}^* \left(\boldsymbol{w} \right) \boldsymbol{x}}{\mu} \right) \left(\frac{1}{q_n \left(\boldsymbol{w} \right)} \boldsymbol{I} + \frac{1}{q_n^3 \left(\boldsymbol{w} \right)} \boldsymbol{w} \boldsymbol{w}^* \right). \tag{7.2}$$

7.1 Proofs for Section 2.2

7.1.1 Proof of Proposition 2.5

The proof involves some delicate analysis, particularly polynomial approximation of the function $f(t) = \frac{1}{(1+t)^2}$ over $t \in [0, 1]$. This is naturally induced by the $1 - \tanh^2(\cdot)$ function. The next lemma characterizes one polynomial approximation of $f(t)$.

Lemma 7.1 *Consider* $f(t) = \frac{1}{(1+t)^2}$ *for* $t \in [0, 1]$. *For every* $T > 1$, *there is a sequence* $b_0, b_1, \ldots$, *with* $\|\boldsymbol{b}\|_{\ell^1} = T < \infty$, *such that the polynomial* $p(t) = \sum_{k=0}^{\infty} b_k t^k$ *satisfies*

$$\|f - p\|_{L^1[0,1]} \leq \frac{1}{2\sqrt{T}}, \quad \|f - p\|_{L^\infty[0,1]} \leq \frac{1}{\sqrt{T}},$$

In particular, one can choose $b_k = (-1)^k (k+1) \beta^k$ *with* $\beta = 1 - 1/\sqrt{T} < 1$ *such that*

$$p(t) = \frac{1}{(1 + \beta t)^2} = \sum_{k=0}^{\infty} (-1)^k (k+1) \beta^k t^k.$$

Moreover, such sequence satisfies $0 < \sum_{k=0}^{\infty} \frac{b_k}{(1+k)^3} < \sum_{k=0}^{\infty} \frac{|b_k|}{(1+k)^3} < 2$.

Lemma 7.2 *Let* $X \sim \mathcal{N} \left(0, \sigma_X^2 \right)$ *and* $Y \sim \mathcal{N} \left(0, \sigma_Y^2 \right)$. *We have*

$$\mathbb{E}\left[\left(1 - \tanh^2\left(\frac{X+Y}{\mu}\right)\right)X^2 \mathbb{1}_{X+Y>0}\right] \le$$

$$\frac{1}{\sqrt{2\pi}} \frac{\mu \sigma_X^2 \sigma_Y^2}{\left(\sigma_X^2 + \sigma_Y^2\right)^{3/2}} + \frac{\mu^3 \sigma_X^2 \sigma_Y^2}{\left(\sigma_X^2 + \sigma_Y^2\right)^{3/2}} + \frac{3}{4\sqrt{2\pi}} \frac{\sigma_X^2 \mu^3}{\left(\sigma_X^2 + \sigma_Y^2\right)^{5/2}}\left(3\mu^2 + 4\sigma_X^2\right).$$

Proof For $x + y > 0$, let $z = \exp\left(-2\frac{x+y}{\mu}\right) \in [0,1]$, then $1 - \tanh^2\left(\frac{x+y}{\mu}\right) = \frac{4z}{(1+z)^2}$. Fix any $T > 1$ to be determined later, by Lemma 7.1, we choose the polynomial $p_\beta(z) = \frac{1}{(1+\beta z)^2}$ with $\beta = 1 - 1/\sqrt{T}$ to upper bound $f(z) = \frac{1}{(1+z)^2}$. So we have

$$\mathbb{E}\left[\left(1 - \tanh^2\left(\frac{X+Y}{\mu}\right)\right)X^2 \mathbb{1}_{X+Y>0}\right] = 4\mathbb{E}\left[Zf(Z)X^2\mathbb{1}_{X+Y>0}\right]$$

$$\le 4\mathbb{E}\left[Zp_\beta(Z)X^2\mathbb{1}_{X+Y>0}\right]$$

$$= 4\sum_{k=0}^{\infty}\left\{b_k \mathbb{E}\left[Z^{k+1}X^2\mathbb{1}_{X+Y>0}\right]\right\},$$

where $b_k = (-1)^k(k+1)\beta^k$, and the exchange of infinite summation and expectation above is justified in view that

$$\sum_{k=0}^{\infty}|b_k|\,\mathbb{E}\left[Z^{k+1}X^2\mathbb{1}_{X+Y>0}\right] \le \sum_{k=0}^{\infty}|b_k|\,\mathbb{E}\left[X^2\mathbb{1}_{X+Y>0}\right] \le \sigma_X^2\sum_{k=0}^{\infty}|b_k| < \infty$$

and the dominated convergence theorem (see, e.g., theorem 2.24 and 2.25 of [Fol99]). By Lemma B.1, we have

$$\sum_{k=0}^{\infty}\left\{b_k\mathbb{E}\left[Z^{k+1}X^2\mathbb{1}_{X+Y>0}\right]\right\}$$

$$= \sum_{k=0}^{\infty}(-\beta)^k(k+1)\left[\left(\sigma_X^2 + \frac{4(k+1)^2}{\mu^2}\sigma_X^4\right)\exp\left(\frac{2(k+1)^2}{\mu^2}\left(\sigma_X^2 + \sigma_Y^2\right)\right)\Phi^c\left(\frac{2(k+1)}{\mu}\sqrt{\sigma_X^2 + \sigma_Y^2}\right)\right.$$

$$\left. -\frac{2(k+1)}{\mu}\frac{\sigma_X^4}{\sqrt{2\pi}\sqrt{\sigma_X^2 + \sigma_Y^2}}\right]$$

$$\le \frac{1}{\sqrt{2\pi}}\sum_{k=0}^{\infty}(-\beta)^k(k+1)\left[\frac{\sigma_X^2\mu}{2(k+1)\sqrt{\sigma_X^2 + \sigma_Y^2}} - \frac{\sigma_X^2\mu^3}{8(k+1)^3\left(\sigma_X^2 + \sigma_Y^2\right)^{3/2}} - \frac{\mu\sigma_X^4}{2(k+1)\left(\sigma_X^2 + \sigma_Y^2\right)^{3/2}}\right]$$

$$+ \frac{3}{\sqrt{2\pi}}\sum_{k=0}^{\infty}\beta^k(k+1)\left(\sigma_X^2 + \frac{4(k+1)^2}{\mu^2}\sigma_X^4\right)\frac{\mu^5}{32(k+1)^5\left(\sigma_X^2 + \sigma_Y^2\right)^{5/2}},$$

where we have applied Type I upper and lower bounds for $\Phi^c(\cdot)$ to even k and odd k respectively and rearrange the terms to obtain the last line. Using the following estimates (see Lemma 7.1)

$$\sum_{k=0}^{\infty}(-\beta)^k = \frac{1}{1+\beta}, \quad \sum_{k=0}^{\infty}\frac{b_k}{(k+1)^3} \ge 0, \quad \sum_{k=0}^{\infty}\frac{|b_k|}{(k+1)^5} \le \sum_{k=0}^{\infty}\frac{|b_k|}{(k+1)^3} \le 2,$$

we obtain

$$\sum_{k=0}^{\infty} \left\{ b_k \mathbb{E}\left[Z^{k+1} X^2 \mathbb{1}_{X+Y>0} \right] \right\} \leq$$

$$\frac{1}{2\sqrt{2\pi}} \frac{\mu \sigma_X^2 \sigma_Y^2}{\left(\sigma_X^2 + \sigma_Y^2\right)^{3/2}} \frac{1}{1+\beta} + \frac{3}{16\sqrt{2\pi}} \frac{\sigma_X^2 \mu^3}{\left(\sigma_X^2 + \sigma_Y^2\right)^{5/2}} \left(3\mu^2 + 4\sigma_X^2\right).$$

Noticing $\frac{1}{1+\beta} < \frac{1}{2} + \frac{1}{2\sqrt{T}}$ and choosing $T = \mu^{-4}$, we obtain the desired result. ■

Lemma 7.3 *Let* $X \sim \mathcal{N}\left(0, \sigma_X^2\right)$ *and* $Y \sim \mathcal{N}\left(0, \sigma_Y^2\right)$. *We have*

$$\mathbb{E}\left[\tanh\left(\frac{X+Y}{\mu} \right) X \right] \geq$$

$$\frac{2\sigma_X^2}{\sqrt{2\pi}\sqrt{\sigma_X^2 + \sigma_Y^2}} - \frac{4\mu^2 \sigma_X^2}{\sqrt{2\pi}\sqrt{\sigma_X^2 + \sigma_Y^2}} - \frac{2\sigma_X^2 \mu^2}{\sqrt{2\pi}\left(\sigma_X^2 + \sigma_Y^2\right)^{3/2}} - \frac{3\sigma_X^2 \mu^4}{2\sqrt{2\pi}\left(\sigma_X^2 + \sigma_Y^2\right)^{5/2}}.$$

Proof By Lemma B.1, we know

$$\mathbb{E}\left[\tanh\left(\frac{X+Y}{\mu} \right) X \right] = \frac{\sigma_X^2}{\mu} \mathbb{E}\left[1 - \tanh^2\left(\frac{X+Y}{\mu} \right) \right]$$

Similar to the proof of the above lemma, for $x + y > 0$, let $z = \exp\left(-2\frac{x+y}{\mu} \right)$ and $f(z) = \frac{1}{(1+z)^2}$. Fixing any $T > 1$, we will use $4z p_\beta(z) = \frac{4z}{(1+\beta z)^2}$ to approximate the $1 - \tanh^2\left(\frac{x+y}{\mu} \right) = 4z f(z)$ function from above, where again $\beta = 1 - 1/\sqrt{T}$. So we obtain

$$\mathbb{E}\left[1 - \tanh^2\left(\frac{X+Y}{\mu} \right) \right] = 8\mathbb{E}\left[f(Z) Z \mathbb{1}_{X+Y>0} \right]$$

$$= 8\mathbb{E}\left[p_\beta(Z) Z \mathbb{1}_{X+Y>0} \right] - 8\mathbb{E}\left[\left(p_\beta(Z) - f(Z) \right) Z \mathbb{1}_{X+Y>0} \right].$$

Now for the first term, we have

$$\mathbb{E}\left[p_\beta(Z) Z \mathbb{1}_{X+Y>0} \right] = \sum_{k=0}^{\infty} b_k \mathbb{E}\left[Z^{k+1} \mathbb{1}_{X+Y>0} \right],$$

justified as $\sum_{k=0}^{\infty} |b_k| \mathbb{E}\left[Z^{k+1} \mathbb{1}_{X+Y>0} \right] \leq \sum_{k=0}^{\infty} |b_k| < \infty$ making the dominated convergence theorem (see, e.g., theorem 2.24 and 2.25 of [Fol99]) applicable. To proceed, from Lemma B.1, we obtain

$$\sum_{k=0}^{\infty} b_k \mathbb{E}\left[Z^{k+1} \mathbb{1}_{X+Y>0} \right]$$

$$= \sum_{k=0}^{\infty} (-\beta)^k (k+1) \exp\left(\frac{2}{\mu^2} (k+1)^2 \left(\sigma_X^2 + \sigma_Y^2\right) \right) \Phi^c\left(\frac{2}{\mu} (k+1) \sqrt{\sigma_X^2 + \sigma_Y^2} \right)$$

$$\geq \frac{1}{\sqrt{2\pi}} \sum_{k=0}^{\infty} (-\beta)^k (k+1) \left(\frac{\mu}{2(k+1)\sqrt{\sigma_X^2 + \sigma_Y^2}} - \frac{\mu^3}{8(k+1)^3 (\sigma_X^2 + \sigma_Y^2)^{3/2}} \right)$$

$$- \frac{3}{\sqrt{2\pi}} \sum_{k=0}^{\infty} \beta^k (k+1) \frac{\mu^5}{32(k+1)^5 (\sigma_X^2 + \sigma_Y^2)^{5/2}},$$

where we have applied Type I upper and lower bounds for $\Phi^c(\cdot)$ to odd k and even k respectively and rearrange the terms to obtain the last line. Using the following estimates (see Lemma 7.1)

$$\sum_{k=0}^{\infty} (-\beta)^k = \frac{1}{1+\beta}, \quad 0 \leq \sum_{k=0}^{\infty} \frac{b_k}{(k+1)^3} \leq \sum_{k=0}^{\infty} \frac{|b_k|}{(k+1)^5} \leq \sum_{k=0}^{\infty} \frac{|b_k|}{(k+1)^3} \leq 2,$$

we obtain

$$\sum_{k=0}^{\infty} b_k \mathbb{E}\left[Z^{k+1} \mathbb{1}_{X+Y>0} \right] \geq$$

$$\frac{\mu}{2\sqrt{2\pi}\sqrt{\sigma_X^2 + \sigma_Y^2}} \frac{1}{1+\beta} - \frac{\mu^3}{4\sqrt{2\pi}(\sigma_X^2 + \sigma_Y^2)^{3/2}} - \frac{3\mu^5}{16\sqrt{2\pi}(\sigma_X^2 + \sigma_Y^2)^{5/2}}.$$

To proceed, by Lemma B.1 and Lemma 7.1, we have

$$\mathbb{E}\left[(p_\beta(Z) - f(Z)) Z \mathbb{1}_{X+Y>0} \right] \leq \|p - f\|_{L^\infty[0,1]} \mathbb{E}\left[Z \mathbb{1}_{X+Y>0} \right] \leq \frac{\mu}{2\sqrt{2\pi T}\sqrt{\sigma_X^2 + \sigma_Y^2}},$$

where we have also used Type I upper bound for $\Phi^c(\cdot)$. Combining the above estimates, we get

$$\mathbb{E}\left[\tanh\left(\frac{X+Y}{\mu} \right) X \right] \geq$$

$$\frac{4\sigma_X^2}{\sqrt{2\pi}\sqrt{\sigma_X^2 + \sigma_Y^2}} \left(\frac{1}{1+\beta} - \frac{1}{\sqrt{T}} \right) - \frac{2\sigma_X^2 \mu^2}{\sqrt{2\pi}(\sigma_X^2 + \sigma_Y^2)^{3/2}} - \frac{3\sigma_X^2 \mu^4}{2\sqrt{2\pi}(\sigma_X^2 + \sigma_Y^2)^{5/2}}.$$

Noticing $\frac{1}{1+\beta} > \frac{1}{2}$ and taking $T = \mu^{-4}$, we obtain the claimed result. ∎

Proof [of Proposition 2.5] For any $i \in [n-1]$, we have

$$\int_0^1 \int_{\boldsymbol{x}} \left| \frac{\partial}{\partial w_i} h_\mu \left(\boldsymbol{q}^* (\boldsymbol{w}) \boldsymbol{x} \right) \right| \mu(d\boldsymbol{x}) \; dw_i \leq \int_0^1 \int_{\boldsymbol{x}} \left(|x_i| + |x_n| \frac{1}{q_n(\boldsymbol{w})} \right) \mu(d\boldsymbol{x}) \; dw_i < \infty.$$

Hence by Lemma A.4 we obtain $\frac{\partial}{\partial w_i} \mathbb{E}\left[h_\mu \left(\boldsymbol{q}^* (\boldsymbol{w}) \boldsymbol{x} \right) \right] = \mathbb{E}\left[\frac{\partial}{\partial w_i} h_\mu \left(\boldsymbol{q}^* (\boldsymbol{w}) \boldsymbol{x} \right) \right]$. Moreover for any $j \in [n-1]$,

$$\int_0^1 \int_{\boldsymbol{x}} \left| \frac{\partial^2}{\partial w_j \partial w_i} h_\mu \left(\boldsymbol{q}^* (\boldsymbol{w}) \boldsymbol{x} \right) \right| \mu(d\boldsymbol{x}) \; dw_j \leq$$

$$\int_0^1 \int_{\boldsymbol{x}} \left[\frac{1}{\mu} \left(|x_i| + \frac{|x_n|}{q_n(\boldsymbol{w})} \right) \left(|x_j| + \frac{|x_n|}{q_n(\boldsymbol{w})} \right) + |x_n| \left(\frac{1}{q_n(\boldsymbol{w})} + \frac{1}{q_n^3(\boldsymbol{w})} \right) \right] \mu(d\boldsymbol{x}) \; dw_i < \infty.$$

Invoking Lemma A.4 again we obtain

$$\frac{\partial^2}{\partial w_j \partial w_i} \mathbb{E}\left[h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right] = \frac{\partial}{\partial w_j} \mathbb{E}\left[\frac{\partial}{\partial w_i} h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right] = \mathbb{E}\left[\frac{\partial^2}{\partial w_j \partial w_i} h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right].$$

The above holds for any pair of $i, j \in [n-1]$, so it follows that

$$\nabla_{\boldsymbol{w}}^2 \mathbb{E}\left[h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right] = \mathbb{E}\left[\nabla_{\boldsymbol{w}}^2 h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right].$$

Hence it is easy to see that

$$\boldsymbol{w}^* \nabla_{\boldsymbol{w}}^2 \mathbb{E}\left[h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right] \boldsymbol{w}$$
$$= \frac{1}{\mu} \mathbb{E}\left[\left(1 - \tanh^2\left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}}{\mu}\right)\right)\left(\boldsymbol{w}^*\overline{\boldsymbol{x}} - \frac{x_n}{q_n\left(\boldsymbol{w}\right)}\|\boldsymbol{w}\|^2\right)^2\right] - \mathbb{E}\left[\tanh\left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}}{\mu}\right)\frac{x_n}{q_n^3\left(\boldsymbol{w}\right)}\|\boldsymbol{w}\|^2\right].$$

Now the first term is

$$\frac{1}{\mu} \mathbb{E}\left[\left(1 - \tanh^2\left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}}{\mu}\right)\right)\left(\boldsymbol{w}^*\overline{\boldsymbol{x}} - \frac{x_n}{q_n\left(\boldsymbol{w}\right)}\|\boldsymbol{w}\|^2\right)^2\right]$$

$$= \frac{2\left(1-\theta\right)}{\mu} \mathbb{E}\left[\left(1 - \tanh^2\left(\frac{\boldsymbol{w}^*\overline{\boldsymbol{x}}}{\mu}\right)\right)\left(\boldsymbol{w}^*\overline{\boldsymbol{x}}\right)^2 \mathbb{1}_{\boldsymbol{w}^*\overline{\boldsymbol{x}}>0}\right]$$
$$- \frac{4\theta\|\boldsymbol{w}\|^2}{\mu\, q_n^2\left(\boldsymbol{w}\right)} \mathbb{E}\left[\left(1 - \tanh^2\left(\frac{\boldsymbol{w}^*\overline{\boldsymbol{x}} + q_n\left(\boldsymbol{w}\right)x_n}{\mu}\right)\right)\left(\boldsymbol{w}^*\overline{\boldsymbol{x}}\right)\left(q_n\left(\boldsymbol{w}\right)x_n\right)\mathbb{1}_{\boldsymbol{w}^*\overline{\boldsymbol{x}}+q_n(\boldsymbol{w})x_n>0}\right]$$
$$+ \frac{2\theta}{\mu} \mathbb{E}_{\mathcal{J}}\mathbb{E}_{\boldsymbol{v}}\left[\left(1 - \tanh^2\left(\frac{\boldsymbol{w}_{\mathcal{J}}^*\overline{\boldsymbol{v}} + q_n\left(\boldsymbol{w}\right)v_n}{\mu}\right)\right)\left(\boldsymbol{w}_{\mathcal{J}}^*\overline{\boldsymbol{v}}\right)^2 \mathbb{1}_{\boldsymbol{w}_{\mathcal{J}}^*\overline{\boldsymbol{v}}+q_n(\boldsymbol{w})v_n>0}\right]$$
$$+ \frac{2\theta\|\boldsymbol{w}\|^4}{\mu\, q_n^4\left(\boldsymbol{w}\right)} \mathbb{E}_{\mathcal{J}}\mathbb{E}_{\boldsymbol{v}}\left[\left(1 - \tanh^2\left(\frac{\boldsymbol{w}_{\mathcal{J}}^*\overline{\boldsymbol{v}} + q_n\left(\boldsymbol{w}\right)v_n}{\mu}\right)\right)\left(q_n\left(\boldsymbol{w}\right)v_n\right)^2 \mathbb{1}_{\boldsymbol{w}_{\mathcal{J}}^*\overline{\boldsymbol{v}}+q_n(\boldsymbol{w})v_n>0}\right]$$

$$\leq \frac{8\left(1-\theta\right)}{\mu} \mathbb{E}\left[\exp\left(-2\frac{\boldsymbol{w}^*\overline{\boldsymbol{x}}}{\mu}\right)\left(\boldsymbol{w}^*\overline{\boldsymbol{x}}\right)^2 \mathbb{1}_{\boldsymbol{w}^*\overline{\boldsymbol{x}}>0}\right]$$
$$+ \frac{8\theta\|\boldsymbol{w}\|^2}{\mu\, q_n^2\left(\boldsymbol{w}\right)} \mathbb{E}\left[\exp\left(-\frac{2}{\mu}\left(\boldsymbol{w}^*\overline{\boldsymbol{x}} + q_n\left(\boldsymbol{w}\right)x_n\right)\right)\left(\boldsymbol{w}^*\overline{\boldsymbol{x}} + q_n\left(\boldsymbol{w}\right)x_n\right)^2 \mathbb{1}_{\boldsymbol{w}^*\overline{\boldsymbol{x}}+q_n(\boldsymbol{w})x_n>0}\right]$$
$$+ \frac{2\theta}{\mu} \mathbb{E}_{\mathcal{J}}\mathbb{E}_{X,Y}\left[\left(1 - \tanh^2\left(\frac{X+Y}{\mu}\right)\right)Y^2 \mathbb{1}_{X+Y>0}\right]$$
$$+ \frac{2\theta\|\boldsymbol{w}\|^4}{\mu\, q_n^4\left(\boldsymbol{w}\right)} \mathbb{E}_{\mathcal{J}}\mathbb{E}_{X,Y}\left[\left(1 - \tanh^2\left(\frac{X+Y}{\mu}\right)\right)X^2 \mathbb{1}_{X+Y>0}\right],$$

where conditioned on each support set $\mathcal{J}$, we let $X \doteq q_n\left(\boldsymbol{w}\right)v_n \sim \mathcal{N}\left(0, q_n^2\left(\boldsymbol{w}\right)\right)$ and $Y \doteq \boldsymbol{w}_{\mathcal{J}}^*\overline{\boldsymbol{v}} \sim \mathcal{N}\left(0, \|\boldsymbol{w}_{\mathcal{J}}\|^2\right)$. Noticing the fact $t \mapsto \exp\left(-2t/\mu\right)t^2$ for $t > 0$ is maximized at $t = \mu$ with maximum value $\exp\left(-2\right)\mu^2$, and in view of the estimate in Lemma 7.2, we obtain

$$\frac{1}{\mu} \mathbb{E}\left[\left(1 - \tanh^2\left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}}{\mu}\right)\right)\left(\boldsymbol{w}^*\overline{\boldsymbol{x}} - \frac{x_n}{q_n\left(\boldsymbol{w}\right)}\|\boldsymbol{w}\|^2\right)^2\right]$$
$$\leq 8\exp\left(-2\right)\left(1 - \theta + \frac{\|\boldsymbol{w}\|^2}{q_n^2\left(\boldsymbol{w}\right)}\theta\right)\mu$$

$$+ \frac{2\theta}{\mu}\mathbb{E}_{\mathcal{J}}\left[\frac{1}{\sqrt{2\pi}}\frac{\mu\|\boldsymbol{w}_{\mathcal{J}}\|^2 q_n^2(\boldsymbol{w})}{\|\boldsymbol{q}_{\mathcal{I}}\|^3} + \frac{\mu^3\|\boldsymbol{w}_{\mathcal{J}}\|^2 q_n^2(\boldsymbol{w})}{\|\boldsymbol{q}_{\mathcal{I}}\|^3} + \frac{3}{4\sqrt{2\pi}}\frac{\|\boldsymbol{w}_{\mathcal{J}}\|^2\mu^3}{\|\boldsymbol{q}_{\mathcal{I}}\|^5}\left(3\mu^2 + 4\|\boldsymbol{w}_{\mathcal{J}}\|^2\right)\right]$$

$$+ \frac{2\theta}{\mu}\frac{\|\boldsymbol{w}\|^4}{q_n^4(\boldsymbol{w})}\mathbb{E}_{\mathcal{J}}\left[\frac{1}{\sqrt{2\pi}}\frac{\mu\|\boldsymbol{w}_{\mathcal{J}}\|^2 q_n^2(\boldsymbol{w})}{\|\boldsymbol{q}_{\mathcal{I}}\|^3} + \frac{\mu^3\|\boldsymbol{w}_{\mathcal{J}}\|^2 q_n^2(\boldsymbol{w})}{\|\boldsymbol{q}_{\mathcal{I}}\|^3} + \frac{3}{4\sqrt{2\pi}}\frac{q_n^2(\boldsymbol{w})\mu^3}{\|\boldsymbol{q}_{\mathcal{I}}\|^5}\left(3\mu^2 + 4q_n^2(\boldsymbol{w})\right)\right]$$

$$\leq \frac{2\theta}{\sqrt{2\pi}q_n^2(\boldsymbol{w})}\mathbb{E}_{\mathcal{J}}\left[\frac{\|\boldsymbol{w}_{\mathcal{J}}\|^2}{\|\boldsymbol{q}_{\mathcal{I}}\|^3}\right] + \frac{11}{20}\mu\left(2 + \frac{1}{q_n^2(\boldsymbol{w})}\right) + 2\theta\mu^2\left(1 + \frac{3}{\sqrt{2\pi}q_n(\boldsymbol{w})} + \frac{1}{q_n^3(\boldsymbol{w})} + \frac{3}{\sqrt{2\pi}q_n^5(\boldsymbol{w})}\right),$$

where we have used $\mu < q_n(\boldsymbol{w}) \leq \|\boldsymbol{q}_{\mathcal{I}}\|$ and $\|\boldsymbol{w}_{\mathcal{J}}\| \leq \|\boldsymbol{q}_{\mathcal{I}}\|$ and $\|\boldsymbol{w}\| \leq 1$ and $\theta \in (0, 1/2)$ to simplify the intermediate quantities to obtain the last line. Similarly for the second term, we obtain

$$\mathbb{E}\left[\tanh\left(\frac{\boldsymbol{q}^*(\boldsymbol{w})\boldsymbol{x}}{\mu}\right)\frac{x_n}{q_n^3(\boldsymbol{w})}\|\boldsymbol{w}\|^2\right]$$

$$= \frac{\|\boldsymbol{w}\|^2\theta}{q_n^4(\boldsymbol{w})}\mathbb{E}_{\mathcal{J}}\mathbb{E}_{\boldsymbol{v}}\left[\tanh\left(\frac{\boldsymbol{w}_{\mathcal{J}}^*\overline{\boldsymbol{v}} + q_n(\boldsymbol{w})v_n}{\mu}\right)x_n q_n(\boldsymbol{w})\right]$$

$$\geq \frac{\|\boldsymbol{w}\|^2\theta}{q_n^4(\boldsymbol{w})}\mathbb{E}_{\mathcal{J}}\left[\frac{2q_n^2(\boldsymbol{w})}{\sqrt{2\pi}\|\boldsymbol{q}_{\mathcal{I}}\|} - \frac{4\mu^2 q_n^2(\boldsymbol{w})}{\sqrt{2\pi}\|\boldsymbol{q}_{\mathcal{I}}\|} - \frac{2q_n^2(\boldsymbol{w})\mu^2}{\sqrt{2\pi}\|\boldsymbol{q}_{\mathcal{I}}\|^3} - \frac{3q_n^2(\boldsymbol{w})\mu^4}{2\sqrt{2\pi}\|\boldsymbol{q}_{\mathcal{I}}\|^5}\right]$$

$$\geq \sqrt{\frac{2}{\pi}}\frac{\theta}{q_n^2(\boldsymbol{w})}\mathbb{E}_{\mathcal{J}}\left[\frac{\|\boldsymbol{w}\|^2}{\|\boldsymbol{q}_{\mathcal{I}}\|}\right] - \frac{4\theta\mu^2}{\sqrt{2\pi}}\left(\frac{1}{q_n^3(\boldsymbol{w})} + \frac{1}{q_n^5(\boldsymbol{w})}\right).$$

Collecting the above estimates, we obtain

$$\boldsymbol{w}^*\nabla_{\boldsymbol{w}}^2\mathbb{E}\left[h_\mu(\boldsymbol{q}^*(\boldsymbol{w})\boldsymbol{x})\right]\boldsymbol{w}$$

$$\leq \sqrt{\frac{2}{\pi}}\frac{\theta}{q_n^2(\boldsymbol{w})}\mathbb{E}_{\mathcal{J}}\left[\frac{\|\boldsymbol{w}_{\mathcal{J}}\|^2}{\|\boldsymbol{q}_{\mathcal{I}}\|^3} - \frac{\|\boldsymbol{w}\|^2\left(\|\boldsymbol{w}_{\mathcal{J}}\|^2 + q_n^2(\boldsymbol{w})\right)}{\|\boldsymbol{q}_{\mathcal{I}}\|^3}\right]$$

$$+ \frac{11}{20}\mu\left(2 + \frac{1}{q_n^2(\boldsymbol{w})}\right) + 2\theta\mu^2\left(1 + \frac{3}{\sqrt{2\pi}q_n(\boldsymbol{w})} + \frac{2}{q_n^3(\boldsymbol{w})} + \frac{5}{\sqrt{2\pi}q_n^5(\boldsymbol{w})}\right)$$

$$\leq -\sqrt{\frac{2}{\pi}}\theta\mathbb{E}\left[\frac{\|\boldsymbol{w}_{\mathcal{J}^c}\|^2}{\|\boldsymbol{q}_{\mathcal{I}}\|^3}\right] + \frac{11}{10}\mu + \frac{11}{20}\frac{\mu}{q_n(\boldsymbol{w})} + 2\theta\mu^2\left(1 + \frac{6}{q_n^5(\boldsymbol{w})}\right)$$

$$\leq -\sqrt{\frac{2}{\pi}}\theta(1 - \theta)\|\boldsymbol{w}\|^2\mathbb{E}\left[\frac{1}{\|\boldsymbol{q}_{\mathcal{I}}\|^3}\right] + \frac{11}{10}\mu + \frac{11}{20}\frac{\mu}{q_n^2(\boldsymbol{w})} + 2\theta\mu^2\left(1 + \frac{6}{q_n^5(\boldsymbol{w})}\right), \tag{7.3}$$

where to obtain the last line we have invoked the association inequality in Lemma A.3, as both $\|\boldsymbol{w}_{\mathcal{J}^c}\|^2$ and $1/\|\boldsymbol{q}_{\mathcal{I}}\|^3$ both coordinatewise nonincreasing w.r.t. the index set. Substituting the upper bound for μ into (7.3) and noting $R_h \leq \|\boldsymbol{w}\|$ and also noting the fact $q_n(\boldsymbol{w}) \geq \frac{1}{2\sqrt{n}}$ (implied by the assumption $\|\boldsymbol{w}\| \leq \sqrt{\frac{4n-1}{4n}}$), we obtain the claimed result. ∎

7.1.2 Proof of Proposition 2.6

Proof By similar consideration as proof of the above proposition, the following is justified:

$$\nabla_{\boldsymbol{w}}\mathbb{E}\left[h_\mu(\boldsymbol{q}^*(\boldsymbol{w})\boldsymbol{x})\right] = \mathbb{E}\left[\nabla_{\boldsymbol{w}}h_\mu(\boldsymbol{q}^*(\boldsymbol{w})\boldsymbol{x})\right].$$

Now consider

$$\boldsymbol{w}^* \nabla \mathbb{E}\left[h_\mu(\boldsymbol{q}^*(\boldsymbol{w})\,\boldsymbol{x})\right] = \nabla \mathbb{E}\left[\boldsymbol{w}^* h_\mu(\boldsymbol{q}^*(\boldsymbol{w})\,\boldsymbol{x})\right]$$

$$= \mathbb{E}\left[\tanh\left(\frac{\boldsymbol{q}^*(\boldsymbol{w})\,\boldsymbol{x}}{\mu}\right)(\boldsymbol{w}^*\bar{\boldsymbol{x}})\right] - \frac{\|\boldsymbol{w}\|^2}{q_n}\mathbb{E}\left[\tanh\left(\frac{\boldsymbol{q}^*(\boldsymbol{w})\,\boldsymbol{x}}{\mu}\right)x_n\right]. \qquad (7.4)$$

For (7.4), we next provide a lower bound for the first expectation and an upper bound for the second expectation. For the first, we have

$$\mathbb{E}\left[\tanh\left(\frac{\boldsymbol{q}^*(\boldsymbol{w})\,\boldsymbol{x}}{\mu}\right)(\boldsymbol{w}^*\bar{\boldsymbol{x}})\right]$$

$$= \theta\mathbb{E}_{\mathcal{J}}\left[\mathbb{E}_{\boldsymbol{v}}\left[\tanh\left(\frac{\boldsymbol{w}_{\mathcal{J}}^*\bar{\boldsymbol{v}} + q_n(\boldsymbol{w})\,v_n}{\mu}\right)(\boldsymbol{w}_{\mathcal{J}}^*\bar{\boldsymbol{v}})\right]\right] + (1-\theta)\mathbb{E}_{\mathcal{J}}\left[\mathbb{E}_{\boldsymbol{v}}\left[\tanh\left(\frac{\boldsymbol{w}_{\mathcal{J}}^*\bar{\boldsymbol{v}}}{\mu}\right)(\boldsymbol{w}_{\mathcal{J}}^*\bar{\boldsymbol{v}})\right]\right]$$

$$= \theta\mathbb{E}_{\mathcal{J}}\left[\mathbb{E}_{X,Y}\left[\tanh\left(\frac{X+Y}{\mu}\right)Y\right]\right] + (1-\theta)\mathbb{E}_{\mathcal{J}}\left[\mathbb{E}_{Y}\left[\tanh\left(\frac{Y}{\mu}\right)Y\right]\right],$$

where $X \doteq q_n(\boldsymbol{w})\,v_n \sim \mathcal{N}\left(0, q_n^2(\boldsymbol{w})\right)$ and $Y \doteq \boldsymbol{w}_{\mathcal{J}}^*\bar{\boldsymbol{v}} \sim \mathcal{N}\left(0, \|\boldsymbol{w}_{\mathcal{J}}\|^2\right)$. Now by Lemma A.3 we obtain

$$\mathbb{E}\left[\tanh\left(\frac{X+Y}{\mu}\right)Y\right] \geq \mathbb{E}\left[\tanh\left(\frac{X+Y}{\mu}\right)\right]\mathbb{E}\left[Y\right] = 0,$$

as $\tanh\left(\frac{X+Y}{\mu}\right)$ and X are both coordinatewise nondecreasing function of X and Y. Using the $\tanh(z) \geq (1 - \exp(-2z))/2$ lower bound for $z > 0$ and integral results in Lemma B.1, we obtain

$$\mathbb{E}\left[\tanh\left(\frac{Y}{\mu}\right)Y\right] = 2\mathbb{E}\left[\tanh\left(\frac{Y}{\mu}\right)Y\mathbb{1}_{Y>0}\right]$$

$$\geq \mathbb{E}\left[\left(1 - \exp\left(-\frac{2Y}{\mu}\right)\right)Y\mathbb{1}_{Y>0}\right]$$

$$= \frac{2\sigma_Y^2}{\mu}\exp\left(\frac{2\sigma_Y^2}{\mu^2}\right)\Phi^c\left(\frac{2\sigma_Y}{\mu}\right)$$

$$\geq \frac{2\sigma_Y^2}{\mu\sqrt{2\pi}}\left(\sqrt{1 + \frac{\sigma_Y^2}{\mu^2}} - \frac{\sigma_Y}{\mu}\right)$$

$$\geq \frac{2\sigma_Y^2}{\mu\sqrt{2\pi}}\left(\sqrt{1 + \frac{\|\boldsymbol{w}\|^2}{\mu^2}} - \frac{\|\boldsymbol{w}\|}{\mu}\right),$$

where at the second last inequality we have used Type III lower bound for Gaussian upper tail $\Phi^c(\cdot)$ (Lemma A.5), and at the last we have used the fact that $t \mapsto \sqrt{1+t^2} - t$ is a monotonic decreasing function over $t > 0$ and that $\sigma_Y = \|\boldsymbol{w}_{\mathcal{J}}\| \leq \|\boldsymbol{w}\|$. Collecting the above estimates, we have

$$\mathbb{E}\left[\tanh\left(\frac{\boldsymbol{q}^*(\boldsymbol{w})\,\boldsymbol{x}}{\mu}\right)(\boldsymbol{w}^*\bar{\boldsymbol{x}})\right] \geq (1-\theta)\mathbb{E}_{\mathcal{J}}\left[\frac{2\|\boldsymbol{w}_{\mathcal{J}}\|^2}{\mu\sqrt{2\pi}}\left(\sqrt{1 + \frac{\|\boldsymbol{w}\|_2^2}{\mu^2}} - \frac{\|\boldsymbol{w}\|}{\mu}\right)\right]$$

$$\geq (1-\theta)\, \mathbb{E}_{\mathcal{J}} \left[\frac{2 \left\| \boldsymbol{w}_{\mathcal{J}} \right\|^2}{\mu \sqrt{2\pi}} \frac{\mu}{10 \left\| \boldsymbol{w} \right\|} \right]$$

$$\geq \frac{\theta (1-\theta) \left\| \boldsymbol{w} \right\|}{5\sqrt{2\pi}}, \tag{7.5}$$

where at the second line we have used the assumption that $\left\| \boldsymbol{w} \right\| \geq \frac{\mu}{6\sqrt{2}}$ and also the fact that $\sqrt{1+x^2} \geq x + \frac{1}{10x}$ for $x \geq \frac{1}{6\sqrt{2}}$.

For the second expectation of (7.4), we have

$$\mathbb{E}\left[\tanh\left(\frac{\boldsymbol{q}^*(\boldsymbol{w})\,\boldsymbol{x}}{\mu} \right) x_n \right] \leq \theta \mathbb{E}\left[\left| \tanh\left(\frac{\boldsymbol{q}^*(\boldsymbol{w})\,\boldsymbol{x}}{\mu} \right) \right| |v_n| \right] \leq \theta \sqrt{\frac{2}{\pi}}, \tag{7.6}$$

as $\tanh(\cdot)$ is bounded by one in magnitude. Plugging the results of (7.5) and (7.6) into (7.4) and noticing that $q_n(\boldsymbol{w})^2 + \left\| \boldsymbol{w} \right\|^2 = 1$ we obtain

$$\boldsymbol{w}^* \nabla \mathbb{E}\left[h_\mu(\boldsymbol{q}^*(\boldsymbol{w})\,\boldsymbol{x}) \right] \geq \frac{\theta \left\| \boldsymbol{w} \right\|}{\sqrt{2\pi}} \left[\frac{1-\theta}{5} - \frac{2 \left\| \boldsymbol{w} \right\|}{\sqrt{1 - \left\| \boldsymbol{w} \right\|^2}} \right] \geq \frac{\theta (1-\theta) \left\| \boldsymbol{w} \right\|}{10\sqrt{2\pi}},$$

where we have invoked the assumption that $\left\| \boldsymbol{w} \right\| \leq \frac{1}{10\sqrt{5}}(1-\theta)$ to provide the upper bound $\frac{2\|\boldsymbol{w}\|}{\sqrt{1-\|\boldsymbol{w}\|^2}} \leq \frac{1}{10}(1-\theta)$. We then choose the particular ranges as stated for μ and θ to ensure $r_g < R_g$, completing the proof. ∎

7.1.3 Proof of Proposition 2.7

Proof By consideration similar to proof of Proposition 2.5, we can exchange the hessian and expectation, i.e.,

$$\nabla^2_{\boldsymbol{w}} \mathbb{E}\left[h_\mu(\boldsymbol{q}^*(\boldsymbol{w})\,\boldsymbol{x}) \right] = \mathbb{E}\left[\nabla^2_{\boldsymbol{w}} h_\mu(\boldsymbol{q}^*(\boldsymbol{w})\,\boldsymbol{x}) \right].$$

We are interested in the expected Hessian matrix

$$\nabla^2_{\boldsymbol{w}} \mathbb{E}\left[h_\mu(\boldsymbol{q}^*(\boldsymbol{w})\,\boldsymbol{x}) \right] = \frac{1}{\mu} \mathbb{E}\left[\left(1 - \tanh^2\left(\frac{\boldsymbol{q}^*(\boldsymbol{w})\,\boldsymbol{x}}{\mu} \right) \right) \left(\overline{\boldsymbol{x}} - \frac{x_n}{q_n(\boldsymbol{w})} \boldsymbol{w} \right) \left(\overline{\boldsymbol{x}} - \frac{x_n}{q_n(\boldsymbol{w})} \boldsymbol{w} \right)^* \right]$$
$$- \mathbb{E}\left[\tanh\left(\frac{\boldsymbol{q}^*(\boldsymbol{w})\,\boldsymbol{x}}{\mu} \right) \left(\frac{x_n}{q_n(\boldsymbol{w})} \boldsymbol{I} + \frac{x_n}{q_n^3(\boldsymbol{w})} \boldsymbol{w}\boldsymbol{w}^* \right) \right]$$

in the region that $0 \leq \left\| \boldsymbol{w} \right\| \leq \frac{\mu}{4\sqrt{2}}$.

When $\boldsymbol{w} = \boldsymbol{0}$, by Lemma B.1, we have

$$\mathbb{E}\left[\nabla^2_{\boldsymbol{w}} h_\mu(\boldsymbol{q}^*(\boldsymbol{w})\,\boldsymbol{x}) \right]\big|_{\boldsymbol{w}=\boldsymbol{0}}$$
$$= \frac{1}{\mu} \mathbb{E}\left[\left(1 - \tanh^2\left(\frac{x_n}{\mu} \right) \right) \overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}^* \right] - \mathbb{E}\left[\tanh\left(\frac{x_n}{\mu} \right) x_n \right] \boldsymbol{I}$$
$$= \frac{\theta(1-\theta)}{\mu} \boldsymbol{I} + \frac{\theta^2}{\mu} \mathbb{E}_{v_n}\left[1 - \tanh^2\left(\frac{v_n}{\mu} \right) \right] \boldsymbol{I} - \frac{\theta}{\mu} \mathbb{E}_{v_n}\left[1 - \tanh^2\left(\frac{v_n}{\mu} \right) \right] \boldsymbol{I}$$

$$= \frac{\theta(1-\theta)}{\mu} \mathbb{E}_{v_n} \left[\tanh^2 \left(\frac{q_n\left(\boldsymbol{w}\right) v_n}{\mu} \right) \right] \boldsymbol{I}.$$

Simple calculation based on Lemma B.1 shows

$$\mathbb{E}_{v_n} \left[\tanh^2 \left(\frac{v_n}{\mu} \right) \right] \geq 2 \left(1 - 4 \exp \left(\frac{2}{\mu^2} \right) \Phi^c \left(\frac{2}{\mu} \right) \right) \geq 2 \left(1 - \frac{2}{\sqrt{2\pi}} \mu \right).$$

Invoking the assumptions $\mu \leq \frac{1}{20\sqrt{n}} \leq 1/20$ and $\theta < 1/2$, we obtain

$$\mathbb{E} \left[\nabla_{\boldsymbol{w}}^2 h_\mu \left(\boldsymbol{q}^*\left(\boldsymbol{w}\right) \boldsymbol{x} \right) \right] \Big|_{\boldsymbol{w}=0} \succeq \frac{\theta\left(1-\theta\right)}{\mu} \left(2 - \frac{4}{\sqrt{2\pi}} \mu \right) \boldsymbol{I} \succeq \frac{\theta}{\mu} \left(1 - \frac{1}{10\sqrt{2\pi}} \right) \boldsymbol{I}.$$

When $0 < \|\boldsymbol{w}\| \leq \frac{\mu}{4\sqrt{2}}$, we aim to derive a semidefinite lower bound for

$$\begin{aligned}
& \mathbb{E} \left[\nabla_{\boldsymbol{w}}^2 h_\mu \left(\boldsymbol{q}^*\left(\boldsymbol{w}\right) \boldsymbol{x} \right) \right] \\
&= \frac{1}{\mu} \mathbb{E} \left[\left(1 - \tanh^2 \left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right) \boldsymbol{x}}{\mu} \right) \right) \overline{\boldsymbol{x}}\, \overline{\boldsymbol{x}}^* \right] - \frac{1}{q_n^2\left(\boldsymbol{w}\right)} \mathbb{E} \left[\tanh \left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right) \boldsymbol{x}}{\mu} \right) q_n\left(\boldsymbol{w}\right) x_n \right] \boldsymbol{I} \\
& \quad - \frac{1}{\mu q_n^2\left(\boldsymbol{w}\right)} \mathbb{E} \left[\left(1 - \tanh^2 \left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right) \boldsymbol{x}}{\mu} \right) \right) q_n\left(\boldsymbol{w}\right) x_n \left(\boldsymbol{w}\overline{\boldsymbol{x}}^* + \overline{\boldsymbol{x}}\boldsymbol{w}^* \right) \right] \\
& \quad + \frac{1}{q_n^4\left(\boldsymbol{w}\right)} \left\{ \frac{1}{\mu} \mathbb{E} \left[\left(1 - \tanh^2 \left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right) \boldsymbol{x}}{\mu} \right) \right) \left(q_n\left(\boldsymbol{w}\right) x_n \right)^2 \right] - \mathbb{E} \left[\tanh \left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right) \boldsymbol{x}}{\mu} \right) q_n\left(\boldsymbol{w}\right) x_n \right] \right\} \boldsymbol{w}\boldsymbol{w}^*.
\end{aligned}$$
$$(7.7)$$

We will first provide bounds for the last two lines and then tackle the first which is slightly more tricky. For the second line, we have

$$\begin{aligned}
& \frac{1}{\mu q_n^2\left(\boldsymbol{w}\right)} \left\| \mathbb{E} \left[\left(1 - \tanh^2 \left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right) \boldsymbol{x}}{\mu} \right) \right) q_n\left(\boldsymbol{w}\right) x_n \left(\boldsymbol{w}\overline{\boldsymbol{x}}^* + \overline{\boldsymbol{x}}\boldsymbol{w}^* \right) \right] \right\| \\
& \leq \frac{2}{\mu q_n^2\left(\boldsymbol{w}\right)} \left\| \mathbb{E} \left[\left(1 - \tanh^2 \left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right) \boldsymbol{x}}{\mu} \right) \right) q_n\left(\boldsymbol{w}\right) x_n \overline{\boldsymbol{x}} \right] \boldsymbol{w}^* \right\| \\
& \leq \frac{2}{\mu q_n^2\left(\boldsymbol{w}\right)} \left\| \mathbb{E} \left[\left(1 - \tanh^2 \left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right) \boldsymbol{x}}{\mu} \right) \right) q_n\left(\boldsymbol{w}\right) x_n \overline{\boldsymbol{x}} \right] \right\| \|\boldsymbol{w}\| \\
& \leq \frac{2}{\mu q_n\left(\boldsymbol{w}\right)} \theta^2 \mathbb{E} \left[|v_n| \right] \mathbb{E} \left[\|\overline{\boldsymbol{v}}\| \right] \|\boldsymbol{w}\| \\
& \leq \frac{4\theta^2}{\pi \mu q_n\left(\boldsymbol{w}\right)} \sqrt{n} \|\boldsymbol{w}\| \leq \frac{\theta}{\mu} \frac{4\theta\sqrt{n} \|\boldsymbol{w}\|}{\pi \sqrt{1 - \|\boldsymbol{w}\|^2}} \leq \frac{\theta}{\mu} \frac{1}{40\pi},
\end{aligned}$$

where from the third to the fourth line we have used $\left\| 1 - \tanh^2 \left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}}{\mu} \right) \right\| \leq 1$, Jensen's inequality for the $\|\cdot\|$ function, and independence of x_n and $\overline{\boldsymbol{x}}$, and to obtain the last bound we have invoked the $\|\boldsymbol{w}\| \leq \frac{\mu}{4\sqrt{2}}$, $\mu \leq \frac{1}{20\sqrt{n}}$, and $\theta < \frac{1}{2}$ assumptions. For the third line in (7.7), by Lemma A.1 and Lemma B.1,

$$\left| \frac{1}{\mu} \mathbb{E} \left[\left(1 - \tanh^2 \left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right) \boldsymbol{x}}{\mu} \right) \right) \left(q_n\left(\boldsymbol{w}\right) x_n \right)^2 \right] - \mathbb{E} \left[\tanh \left(\frac{\boldsymbol{q}^*\left(\boldsymbol{w}\right) \boldsymbol{x}}{\mu} \right) q_n x_n \right] \right|$$

$$= \left| \frac{\theta}{\mu} \mathbb{E}_{\mathcal{J}} \mathbb{E}_{\boldsymbol{v}} \left[\left(1 - \tanh^2 \left(\frac{\boldsymbol{w}_{\mathcal{J}}^* \overline{\boldsymbol{v}} + q_n(\boldsymbol{w}) v_n}{\mu} \right) (q_n(\boldsymbol{w}) v_n)^2 \right) \right] \right.$$

$$\left. - \theta \mathbb{E}_{\mathcal{J}} \mathbb{E}_{\boldsymbol{v}} \left[\tanh \left(\frac{\boldsymbol{w}_{\mathcal{J}}^* \overline{\boldsymbol{v}} + q_n(\boldsymbol{w}) v_n}{\mu} \right) q_n(\boldsymbol{w}) v_n \right] \right|$$

$$= \frac{\theta}{\mu} \mathbb{E}_{\mathcal{J}} \mathbb{E}_{\boldsymbol{v}} \left[\left(1 - \tanh^2 \left(\frac{\boldsymbol{w}_{\mathcal{J}}^* \overline{\boldsymbol{v}} + q_n(\boldsymbol{w}s) v_n}{\mu} \right) \right) ((q_n(\boldsymbol{w}) v_n)^2 + q_n^2(\boldsymbol{w})) \right]$$

$$\leq \frac{8\theta}{\mu} \mathbb{E}_{\mathcal{J}} \mathbb{E}_{\boldsymbol{v}} \left[\exp \left(-\frac{2}{\mu} \left(\boldsymbol{w}_{\mathcal{J}}^* \overline{\boldsymbol{v}} + q_n(\boldsymbol{w}) v_n \right) \right) ((q_n(\boldsymbol{w}) v_n)^2 + q_n^2(\boldsymbol{w})) \, \mathbb{1}_{\boldsymbol{w}_{\mathcal{J}}^* \overline{\boldsymbol{v}} + q_n(\boldsymbol{w}) v_n > 0} \right]$$

$$\leq \frac{8\theta}{\sqrt{2\pi}} \mathbb{E}_{\mathcal{J}} \left[\frac{q_n^2(\boldsymbol{w})}{\sqrt{q_n^2(\boldsymbol{w}) + \|\boldsymbol{w}_{\mathcal{J}}\|^2}} \right] \leq \frac{8\theta q_n(\boldsymbol{w})}{\sqrt{2\pi}}.$$

Thus, we have

$$\frac{1}{q_n^4(\boldsymbol{w})} \left\{ \frac{1}{\mu} \mathbb{E} \left[\left(1 - \tanh^2 \left(\frac{\boldsymbol{q}^* \boldsymbol{x}}{\mu} \right) \right) (q_n x_n)^2 \right] - \mathbb{E} \left[\tanh \left(\frac{\boldsymbol{q}^* \boldsymbol{x}}{\mu} \right) q_n x_n \right] \right\} \boldsymbol{w} \boldsymbol{w}^*$$

$$\succeq - \frac{8\theta}{q_n^3(\boldsymbol{w}) \sqrt{2\pi}} \|\boldsymbol{w}\|^2 \boldsymbol{I} \succeq - \frac{\theta}{\mu} \left(\frac{64 n^{3/2} \mu \|\boldsymbol{w}\|^2}{q_n^3(\boldsymbol{w}) \sqrt{2\pi}} \right) \boldsymbol{I} \succeq - \frac{\theta}{\mu} \frac{1}{4000 \sqrt{2\pi}} \boldsymbol{I},$$

where we have again used $\|\boldsymbol{w}\| \leq \frac{\mu}{4\sqrt{2}}$, $\mu \leq \frac{1}{20\sqrt{n}}$, and $q_n(\boldsymbol{w}) \geq \frac{1}{2\sqrt{n}}$ assumptions to simplify the final bound.

To derive a lower bound for the first line of (7.7), we lower bound the first term and upper bound the second. The latter is easy: using Lemma A.1 and Lemma B.1,

$$\frac{1}{q_n^2(\boldsymbol{w})} \mathbb{E} \left[\tanh \left(\frac{\boldsymbol{q}^*(\boldsymbol{w}) \boldsymbol{x}}{\mu} \right) q_n(\boldsymbol{w}) x_n \right]$$

$$= \frac{\theta}{\mu} \mathbb{E}_{\mathcal{J}} \mathbb{E}_{\boldsymbol{v}} \left[1 - \tanh^2 \left[\frac{\boldsymbol{w}_{\mathcal{J}}^* \overline{\boldsymbol{v}} + q_n(\boldsymbol{w}) v_n}{\mu} \right] \right]$$

$$\leq \frac{8\theta}{\mu} \mathbb{E}_{\mathcal{J}} \mathbb{E}_{\boldsymbol{v}} \left[\exp \left(-2 \frac{\boldsymbol{w}_{\mathcal{J}}^* \overline{\boldsymbol{v}} + q_n(\boldsymbol{w}) v_n}{\mu} \right) \mathbb{1}_{\boldsymbol{w}_{\mathcal{J}}^* \overline{\boldsymbol{v}} + q_n(\boldsymbol{w}) v_n > 0} \right]$$

$$\leq \frac{4\theta}{\sqrt{2\pi} q_n(\boldsymbol{w})} \leq \frac{\theta}{\mu} \frac{8\sqrt{n}\mu}{\sqrt{2\pi}} \leq \frac{\theta}{\mu} \frac{2}{5\sqrt{2\pi}},$$

where we have again used assumptions that $q_n(\boldsymbol{w}) \geq \frac{1}{2\sqrt{n}}$ and $\mu \leq \frac{1}{20\sqrt{n}}$ to simplify the last bound. To lower bound the first term, first note that

$$\frac{1}{\mu} \mathbb{E} \left[\left(1 - \tanh^2 \left(\frac{\boldsymbol{q}^*(\boldsymbol{w}) \boldsymbol{x}}{\mu} \right) \right) \overline{\boldsymbol{x}} \, \overline{\boldsymbol{x}}^* \right] \succeq \frac{1-\theta}{\mu} \mathbb{E}_{\overline{\boldsymbol{x}}} \left[\left(1 - \tanh^2 \left(\frac{\boldsymbol{w}^* \overline{\boldsymbol{x}}}{\mu} \right) \right) \overline{\boldsymbol{x}} \, \overline{\boldsymbol{x}}^* \right].$$

We set out to lower bound the expectation as

$$\mathbb{E}_{\overline{\boldsymbol{x}}} \left[\left(1 - \tanh^2 \left(\frac{\boldsymbol{w}^* \overline{\boldsymbol{x}}}{\mu} \right) \right) \overline{\boldsymbol{x}} \, \overline{\boldsymbol{x}}^* \right] \succeq \theta \beta \boldsymbol{I}$$

for some scalar $\beta > 0$. Suppose $\boldsymbol{w}$ has $k \in [n-1]$ nonzeros, w.l.o.g., further assume the first k elements of $\boldsymbol{w}$ are these nonzeros. It is easy to see the expectation above has a block diagonal

56

structure $\mathrm{diag}\left(\mathbf{\Sigma}; \theta \mathbf{I}_{n-1-k}\right)$. So in order to derive the $\theta\beta\mathbf{I}$ lower bound as desired, it is sufficient to show $\mathbf{\Sigma} \succeq \theta\beta\mathbf{I}$ for some $0 < \beta < 1$, i.e., letting $\widetilde{\boldsymbol{w}} \in \mathbb{R}^k$ be the subvector of nonzero elements,

$$\mathbb{E}_{\widetilde{\boldsymbol{x}} \sim i.i.d.\,\mathrm{BG}(\theta)} \left[\left(1 - \tanh^2 \left(\frac{\widetilde{\boldsymbol{w}}^* \widetilde{\boldsymbol{x}}}{\mu} \right) \right) \widetilde{\boldsymbol{x}}\, \widetilde{\boldsymbol{x}}^* \right] \succeq \theta\beta\mathbf{I},$$

which is equivalent to that for all $\boldsymbol{z} \in \mathbb{R}^k$ such that $\|\boldsymbol{z}\| = 1$,

$$\mathbb{E}_{\widetilde{\boldsymbol{x}} \sim i.i.d.\,\mathrm{BG}(\theta)} \left[\left(1 - \tanh^2 \left(\frac{\widetilde{\boldsymbol{w}}^* \widetilde{\boldsymbol{x}}}{\mu} \right) \right) (\widetilde{\boldsymbol{x}}^* \boldsymbol{z})^2 \right] \geq \theta\beta.$$

It is then sufficient to show that for any nontrivial support set $\mathcal{S} \subset [k]$ and any vector $\boldsymbol{z} \in \mathbb{R}^k$ such that $\mathrm{supp}\,(\boldsymbol{z}) = \mathcal{S}$ with $\|\boldsymbol{z}\| = 1$,

$$\mathbb{E}_{\widetilde{\boldsymbol{v}} \sim i.i.d.\,\mathcal{N}(0,1)} \left[\left(1 - \tanh^2 \left(\frac{\widetilde{\boldsymbol{w}}_{\mathcal{S}}^* \widetilde{\boldsymbol{v}}}{\mu} \right) \right) (\widetilde{\boldsymbol{v}}^* \boldsymbol{z})^2 \right] \geq \beta.$$

To see the implication, suppose the latter claimed holds, then for any $\boldsymbol{z}$ with unit norm,

$$\mathbb{E}_{\widetilde{\boldsymbol{x}} \sim i.i.d.\,\mathrm{BG}(\theta)} \left[\left(1 - \tanh^2 \left(\frac{\widetilde{\boldsymbol{w}}^* \widetilde{\boldsymbol{x}}}{\mu} \right) \right) (\widetilde{\boldsymbol{x}}^* \boldsymbol{z})^2 \right]$$

$$= \sum_{s=1}^k \theta^s (1-\theta)^{k-s} \sum_{\mathcal{S} \in \binom{[k]}{s}} \mathbb{E}_{\widetilde{\boldsymbol{v}} \sim i.i.d.\,\mathcal{N}(0,1)} \left[\left(1 - \tanh^2 \left(\frac{\widetilde{\boldsymbol{w}}_{\mathcal{S}}^* \widetilde{\boldsymbol{v}}}{\mu} \right) \right) (\widetilde{\boldsymbol{v}}^* \boldsymbol{z}_{\mathcal{S}})^2 \right]$$

$$\geq \sum_{s=1}^k \theta^s (1-\theta)^{k-s} \sum_{\mathcal{S} \in \binom{[k]}{s}} \beta \|\boldsymbol{z}_{\mathcal{S}}\|^2 = \beta\mathbb{E}_{\mathcal{S}} \left[\|\boldsymbol{z}_{\mathcal{S}}\|^2 \right] = \theta\beta.$$

Now for any fixed support set $\mathcal{S} \subset [k]$, $\boldsymbol{z} = \mathcal{P}_{\widetilde{\boldsymbol{w}}_{\mathcal{S}}} \boldsymbol{z} + \left(\mathbf{I} - \mathcal{P}_{\widetilde{\boldsymbol{w}}_{\mathcal{S}}}\right) \boldsymbol{z}$. So we have

$$\mathbb{E}_{\widetilde{\boldsymbol{v}} \sim i.i.d.\,\mathcal{N}(0,1)} \left[\left(1 - \tanh^2 \left(\frac{\widetilde{\boldsymbol{w}}_{\mathcal{S}}^* \widetilde{\boldsymbol{v}}}{\mu} \right) \right) (\widetilde{\boldsymbol{v}}^* \boldsymbol{z})^2 \right]$$

$$= \mathbb{E}_{\widetilde{\boldsymbol{v}}} \left[\left(1 - \tanh^2 \left(\frac{\widetilde{\boldsymbol{w}}_{\mathcal{S}}^* \widetilde{\boldsymbol{v}}}{\mu} \right) \right) (\widetilde{\boldsymbol{v}}^* \mathcal{P}_{\widetilde{\boldsymbol{w}}_{\mathcal{S}}} \boldsymbol{z})^2 \right] + \mathbb{E}_{\widetilde{\boldsymbol{v}}} \left[\left(1 - \tanh^2 \left(\frac{\widetilde{\boldsymbol{w}}_{\mathcal{S}}^* \widetilde{\boldsymbol{v}}}{\mu} \right) \right) \left(\widetilde{\boldsymbol{v}}^* \left(\mathbf{I} - \mathcal{P}_{\widetilde{\boldsymbol{w}}_{\mathcal{S}}}\right) \boldsymbol{z}\right)^2 \right]$$

$$= \frac{(\widetilde{\boldsymbol{w}}_{\mathcal{S}}^* \boldsymbol{z})^2}{\|\boldsymbol{w}_{\mathcal{S}}\|^4} \mathbb{E}_{\widetilde{\boldsymbol{v}}} \left[\left(1 - \tanh^2 \left(\frac{\widetilde{\boldsymbol{w}}_{\mathcal{S}}^* \widetilde{\boldsymbol{v}}}{\mu} \right) \right) (\widetilde{\boldsymbol{v}}^* \widetilde{\boldsymbol{w}}_{\mathcal{S}})^2 \right] + \mathbb{E}_{\widetilde{\boldsymbol{v}}} \left[\left(1 - \tanh^2 \left(\frac{\widetilde{\boldsymbol{w}}_{\mathcal{S}}^* \widetilde{\boldsymbol{v}}}{\mu} \right) \right) \right] \mathbb{E}_{\widetilde{\boldsymbol{v}}} \left[\left(\widetilde{\boldsymbol{v}}^* \left(\mathbf{I} - \mathcal{P}_{\widetilde{\boldsymbol{w}}_{\mathcal{S}}}\right) \boldsymbol{z}\right)^2 \right]$$

$$\geq 2\frac{(\widetilde{\boldsymbol{w}}_{\mathcal{S}}^* \boldsymbol{z})^2}{\|\boldsymbol{w}_{\mathcal{S}}\|^4} \mathbb{E}_{\widetilde{\boldsymbol{v}}} \left[\exp \left(-\frac{2\widetilde{\boldsymbol{w}}_{\mathcal{S}}^* \widetilde{\boldsymbol{v}}}{\mu} \right) (\widetilde{\boldsymbol{v}}^* \widetilde{\boldsymbol{w}}_{\mathcal{S}})^2 \mathbb{1}_{\widetilde{\boldsymbol{v}}^* \widetilde{\boldsymbol{w}}_{\mathcal{S}} > 0} \right] + 2\mathbb{E}_{\widetilde{\boldsymbol{v}}} \left[\exp \left(-\frac{2\widetilde{\boldsymbol{w}}_{\mathcal{S}}^* \widetilde{\boldsymbol{v}}}{\mu} \right) \mathbb{1}_{\widetilde{\boldsymbol{w}}_{\mathcal{S}}^* \widetilde{\boldsymbol{v}} > 0} \right] \left\|\left(\mathbf{I} - \mathcal{P}_{\widetilde{\boldsymbol{w}}_{\mathcal{S}}}\right) \boldsymbol{z}\right\|^2.$$

Using expectation result from Lemma B.1, and applying Type III lower bound for Gaussian tails, we obtain

$$\mathbb{E}_{\widetilde{\boldsymbol{v}} \sim i.i.d.\,\mathcal{N}(0,1)} \left[\left(1 - \tanh^2 \left(\frac{\widetilde{\boldsymbol{w}}_{\mathcal{S}}^* \widetilde{\boldsymbol{v}}}{\mu} \right) \right) (\widetilde{\boldsymbol{v}}^* \boldsymbol{z})^2 \right]$$

$$\geq \frac{1}{\sqrt{2\pi}} \left(\sqrt{4 + \frac{4 \|\widetilde{\boldsymbol{w}}_{\mathcal{S}}\|^2}{\mu^2}} - \frac{2 \|\widetilde{\boldsymbol{w}}_{\mathcal{S}}\|}{\mu} \right) - \frac{4 (\widetilde{\boldsymbol{w}}_{\mathcal{S}}^* \boldsymbol{z})^2}{\mu\sqrt{2\pi} \|\widetilde{\boldsymbol{w}}_{\mathcal{S}}\|}$$

$$\geq \frac{1}{\sqrt{2\pi}} \left(2 - \frac{3}{4}\sqrt{2} \right),$$

where we have used Cauchy-Schwarz to obtain $(\widetilde{\boldsymbol{v}}^*\boldsymbol{z})^2 \leq \|\widetilde{\boldsymbol{v}}^*\|^2$ and invoked the assumption $\|\boldsymbol{w}\| \leq \frac{\mu}{4\sqrt{2}}$ to simplify the last bound. So $\beta = \frac{1}{\sqrt{2\pi}}\left(2 - \frac{3}{4}\sqrt{2}\right) < 1$.

Putting together the above estimates for the case $\boldsymbol{w} \neq \boldsymbol{0}$, we obtain

$$\mathbb{E}\left[\nabla_{\boldsymbol{w}}^2 h_\mu\left(\boldsymbol{q}^*\left(\boldsymbol{w}\right)\boldsymbol{x}\right)\right] \succeq \frac{\theta}{\mu\sqrt{2\pi}}\left(1 - \frac{3}{8}\sqrt{2} - \frac{\sqrt{2\pi}}{40\pi} - \frac{1}{4000} - \frac{2}{5}\right)\boldsymbol{I} \succeq \frac{1}{25\sqrt{2\pi}}\frac{\theta}{\mu}\boldsymbol{I}.$$

Hence for all $\boldsymbol{w}$, we can take the $\frac{1}{25\sqrt{2\pi}}\frac{\theta}{\mu}$ as the lower bound, completing the proof. ∎

7.1.4 Proof of Pointwise Concentration Results

To avoid clutter of notations, in this subsection we write $\boldsymbol{X}$ to mean $\boldsymbol{X}_0$; similarly $\boldsymbol{x}_k$ for $(\boldsymbol{x}_0)_k$, the k-th column of $\boldsymbol{X}_0$. The function $g\left(\boldsymbol{w}\right)$ means $g\left(\boldsymbol{w}; \boldsymbol{X}_0\right)$. We first establish a useful comparison lemma between random i.i.d. Bernoulli random vectors random i.i.d. normal random vectors.

Lemma 7.4 *Suppose $\boldsymbol{z}, \boldsymbol{z}' \in \mathbb{R}^n$ are independent and obey $\boldsymbol{z} \sim_{i.i.d.} \mathrm{BG}\left(\theta\right)$ and $\boldsymbol{z}' \sim_{i.i.d.} \mathcal{N}\left(0,1\right)$. Then, for any fixed vector $\boldsymbol{v} \in \mathbb{R}^n$, it holds that*

$$\mathbb{E}\left[|\boldsymbol{v}^*\boldsymbol{z}|^m\right] \leq \mathbb{E}\left[\left|\boldsymbol{v}^*\boldsymbol{z}'\right|^m\right] = \mathbb{E}_{Z\sim\mathcal{N}\left(0,\|\boldsymbol{v}\|^2\right)}\left[|Z|^m\right],$$

$$\mathbb{E}\left[\|\boldsymbol{z}\|^m\right] \leq \mathbb{E}\left[\|\boldsymbol{z}'\|^m\right],$$

for all integers $m \geq 1$.

Now, we are ready to prove Proposition 2.8 to Proposition 2.10 as follows.

Proof [of Proposition 2.8] Let

$$Y_k = \frac{1}{\|\boldsymbol{w}\|^2}\boldsymbol{w}^*\nabla^2 h_\mu\left(\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}_k\right)\boldsymbol{w},$$

then $\frac{\boldsymbol{w}^*\nabla^2 g(\boldsymbol{w})\boldsymbol{w}}{\|\boldsymbol{w}\|^2} = \frac{1}{p}\sum_{k=1}^p Y_k$. For each Y_k ($k \in [p]$), from (7.2), we know that

$$Y_k = \frac{1}{\mu}\left(1 - \tanh^2\left(\frac{\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}_k}{\mu}\right)\right)\left(\frac{\boldsymbol{w}^*\overline{\boldsymbol{x}}_k}{\|\boldsymbol{w}\|} - \frac{x_k\left(n\right)\|\boldsymbol{w}\|}{q_n(\boldsymbol{w})}\right)^2 - \tanh\left(\frac{\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}_k}{\mu}\right)\frac{x_k\left(n\right)}{q_n^3(\boldsymbol{w})}.$$

Writing $Y_k = W_k + V_k$, where

$$W_k = \frac{1}{\mu}\left(1 - \tanh^2\left(\frac{\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}_k}{\mu}\right)\right)\left(\frac{\boldsymbol{w}^*\overline{\boldsymbol{x}}_k}{\|\boldsymbol{w}\|} - \frac{x_k\left(n\right)\|\boldsymbol{w}\|}{q_n(\boldsymbol{w})}\right)^2,$$

$$V_k = -\tanh\left(\frac{\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}_k}{\mu}\right)\frac{x_k\left(n\right)}{q_n^3(\boldsymbol{w})}.$$

Then by similar argument as in proof to Proposition 2.9, we have for all integers $m \geq 2$ that

$$\mathbb{E}\left[|W_k|^m\right] \leq \frac{1}{\mu^m}\mathbb{E}\left[\left|\frac{\boldsymbol{w}^*\overline{\boldsymbol{x}}_k}{\|\boldsymbol{w}\|} - \frac{x_k\left(n\right)\|\boldsymbol{w}\|}{q_n(\boldsymbol{w})}\right|^{2m}\right] \leq \frac{1}{\mu^m}\mathbb{E}_{Z\sim\mathcal{N}\left(0,1/q_n^2(\boldsymbol{w})\right)}\left[|Z|^{2m}\right]$$

$$\leq \frac{1}{\mu^m}(2m-1)!!(4n)^m \leq \frac{m!}{2}\left(\frac{4n}{\mu}\right)^m,$$

$$\mathbb{E}\left[|V_k|^m\right] \leq \frac{1}{q_n^{3m}(\boldsymbol{w})}\mathbb{E}\left[|v_k(n)|^m\right] \leq \left(2\sqrt{n}\right)^{3m}(m-1)!! \leq \frac{m!}{2}\left(8n\sqrt{n}\right)^m,$$

where we have again used the assumption that $q_n(\boldsymbol{w}) \geq \frac{1}{2\sqrt{n}}$ to simplify the result. Taking $\sigma_W^2 = 16n^2/\mu^2 \geq \mathbb{E}\left[W_k^2\right]$, $R_W = 4n/\mu$ and $\sigma_V^2 = 64n^3 \geq \mathbb{E}\left[V_k^2\right]$, $R_V = 8n\sqrt{n}$, and considering $S_W = \frac{1}{p}\sum_{k=1}^p W_k$ and $S_V = \frac{1}{p}\sum_{k=1}^p V_k$, then by Lemma A.9, we obtain

$$\mathbb{P}\left[|S_W - \mathbb{E}\left[S_W\right]| \geq \frac{t}{2}\right] \leq 2\exp\left(-\frac{p\mu^2 t^2}{128n^2 + 16n\mu t}\right),$$

$$\mathbb{P}\left[|S_V - \mathbb{E}\left[S_V\right]| \geq \frac{t}{2}\right] \leq 2\exp\left(-\frac{pt^2}{512n^3 + 32n\sqrt{n}t}\right).$$

Combining the above results, we obtain

$$\mathbb{P}\left[\left|\frac{1}{p}\sum_{k=1}^p X_k - \mathbb{E}\left[X_k\right]\right| \geq t\right] = \mathbb{P}\left[|S_W - \mathbb{E}\left[S_W\right] + S_V - \mathbb{E}\left[S_V\right]| \geq t\right]$$

$$\leq \mathbb{P}\left[|S_W - \mathbb{E}\left[S_W\right]| \geq \frac{t}{2}\right] + \mathbb{P}\left[|S_V - \mathbb{E}\left[S_V\right]| \geq \frac{t}{2}\right]$$

$$\leq 2\exp\left(-\frac{p\mu^2 t^2}{128n^2 + 16n\mu t}\right) + 2\exp\left(-\frac{pt^2}{512n^3 + 32n\sqrt{n}t}\right)$$

$$\leq 4\exp\left(-\frac{p\mu^2 t^2}{512n^2 + 32n\mu t}\right),$$

provided that $\mu \leq \frac{1}{\sqrt{n}}$, as desired. ∎

Proof [of Proposition 2.9] Let

$$X_k = \frac{\boldsymbol{w}^*}{\|\boldsymbol{w}\|_2}\nabla h_\mu\left(\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}_k\right),$$

then $\frac{\boldsymbol{w}^*\nabla g(\boldsymbol{w})}{\|\boldsymbol{w}\|_2} = \frac{1}{p}\sum_{k=1}^p X_k$. For each $X_k, k \in [p]$, from (7.1), we know that

$$|X_k| = \left|\tanh\left(\frac{\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}_k}{\mu}\right)\left(\frac{\boldsymbol{w}^*\overline{\boldsymbol{x}}_k}{\|\boldsymbol{w}\|} - \frac{\|\boldsymbol{w}\|_2 x_k(n)}{q_n(\boldsymbol{w})}\right)\right| \leq \left|\frac{\boldsymbol{w}^*\overline{\boldsymbol{x}}_k}{\|\boldsymbol{w}\|} - \frac{\|\boldsymbol{w}\|_2 x_k(n)}{q_n(\boldsymbol{w})}\right|,$$

as the magnitude of $\tanh(\cdot)$ is bounded by one. Because $\frac{\boldsymbol{w}^*\overline{\boldsymbol{x}}_k}{\|\boldsymbol{w}\|_2} - \frac{\|\boldsymbol{w}\|x_k(n)}{q_n(\boldsymbol{w})} = \left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, -\frac{\|\boldsymbol{w}\|}{q_n(\boldsymbol{w})}\right)^*\boldsymbol{x}_k$ and $\boldsymbol{x}_k \sim_{i.i.d.} \text{BG}(\theta)$, invoking Lemma 7.4, we obtain for every integer $m \geq 2$ that

$$\mathbb{E}\left[|X_k|^m\right] \leq \mathbb{E}_{Z\sim\mathcal{N}(0,1/q_n^2(\boldsymbol{w}))}\left[|Z|^m\right] \leq \frac{1}{q_n(\boldsymbol{w})^m}(m-1)!! \leq \frac{m!}{2}(4n)\left(2\sqrt{n}\right)^{m-2},$$

where the Gaussian moment can be looked up in Lemma A.6 and we used the fact that $(m-1)!! \leq m!/2$ and the assumption that $q_n(\boldsymbol{w}) \geq \frac{1}{2\sqrt{n}}$ to get the result. Thus, by taking $\sigma^2 = 4n \geq \mathbb{E}\left[X_k^2\right]$ and $R = 2\sqrt{n}$, and we obtain the claimed result by invoking Lemma A.9. ∎

Proof [of Proposition 2.10] Let $\boldsymbol{Z}_k = \nabla_{\boldsymbol{w}}^2 h_\mu\left(\boldsymbol{q}(\boldsymbol{w})^* \boldsymbol{x}_k\right)$, then $\nabla_{\boldsymbol{w}}^2 g\left(\boldsymbol{w}\right) = \frac{1}{p}\sum_{k=1}^p \boldsymbol{Z}_k$. From (7.2), we know that

$$\boldsymbol{Z}_k \;=\; \boldsymbol{W}_k + \boldsymbol{V}_k$$

where

$$\boldsymbol{W}_k \;=\; \frac{1}{\mu}\left(1 - \tanh^2\left(\frac{\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}_k}{\mu}\right)\right)\left(\overline{\boldsymbol{x}}_k - \frac{x_k\left(n\right)\boldsymbol{w}}{q_n(\boldsymbol{w})}\right)\left(\overline{\boldsymbol{x}}_k - \frac{x_k\left(n\right)\boldsymbol{w}}{q_n(\boldsymbol{w})}\right)^*$$

$$\boldsymbol{V}_k \;=\; -\tanh\left(\frac{\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}_k}{\mu}\right)\left(\frac{x_k\left(n\right)}{q_n(\boldsymbol{w})}\boldsymbol{I} + \frac{x_k\left(n\right)\boldsymbol{w}\boldsymbol{w}^*}{q_n^3(\boldsymbol{w})}\right).$$

For $\boldsymbol{W}_k$, we have

$$\boldsymbol{0} \preceq \mathbb{E}\left[\boldsymbol{W}_k^m\right] \preceq \frac{1}{\mu^m}\mathbb{E}\left[\left\|\overline{\boldsymbol{x}}_k - \frac{x_k\left(n\right)\boldsymbol{w}}{q_n(\boldsymbol{w})}\right\|^{2m-2}\left(\overline{\boldsymbol{x}}_k - \frac{x_k\left(n\right)\boldsymbol{w}}{q_n(\boldsymbol{w})}\right)\left(\overline{\boldsymbol{x}}_k - \frac{x_k\left(n\right)\boldsymbol{w}}{q_n(\boldsymbol{w})}\right)^*\right]$$

$$\preceq \frac{1}{\mu^m}\mathbb{E}\left[\left\|\overline{\boldsymbol{x}}_k - \frac{x_k\left(n\right)\boldsymbol{w}}{q_n(\boldsymbol{w})}\right\|^{2m}\right]\boldsymbol{I}$$

$$\preceq \frac{2^m}{\mu^m}\mathbb{E}\left[\left(\|\overline{\boldsymbol{x}}_k\|^2 + \frac{x_k^2\left(n\right)\|\boldsymbol{w}\|^2}{q_n^2(\boldsymbol{w})}\right)^m\right]\boldsymbol{I}$$

$$\preceq \frac{2^m}{\mu^m}\mathbb{E}\left[\|\boldsymbol{x}_k\|^{2m}\right]\boldsymbol{I} \;\preceq\; \frac{2^m}{\mu^m}\mathbb{E}_{Z\sim\chi^2(n)}\left[Z^m\right]\boldsymbol{I},$$

where we have used the fact that $\|\boldsymbol{w}\|^2/q_n^2(\boldsymbol{w}) = \|\boldsymbol{w}\|^2/(1-\|\boldsymbol{w}\|^2) \leq 1$ for $\|\boldsymbol{w}\|_2 \leq \frac{1}{4}$ and Lemma 7.4 to obtain the last line. By Lemma A.7, we obtain

$$\boldsymbol{0} \preceq \mathbb{E}\left[\boldsymbol{W}_k^m\right] \;\preceq\; \left(\frac{2}{\mu}\right)^m \frac{m!}{2}(2n)^m \boldsymbol{I} \;=\; \frac{m!}{2}\left(\frac{4n}{\mu}\right)^m \boldsymbol{I}.$$

Taking $R_W = \frac{4n}{\mu}$ and $\boldsymbol{\sigma}_W^2 = \frac{16n^2}{\mu^2} \geq \mathbb{E}\left[\boldsymbol{W}_k^2\right]$, and letting $\boldsymbol{S}_W \doteq \frac{1}{p}\sum_{k=1}^p \boldsymbol{W}_k$, by Lemma A.10, we obtain

$$\mathbb{P}\left[\|\boldsymbol{S}_W - \mathbb{E}\left[\boldsymbol{S}_W\right]\| \geq \frac{t}{2}\right] \;\leq\; 2n\exp\left(-\frac{p\mu^2 t^2}{128n^2 + 16\mu n t}\right).$$

Similarly, for $\boldsymbol{V}_k$, we have

$$\mathbb{E}\left[\boldsymbol{V}_k^m\right] \;\preceq\; \left(\frac{1}{q_n(\boldsymbol{w})} + \frac{\|\boldsymbol{w}\|^2}{q_n^3(\boldsymbol{w})}\right)^m \mathbb{E}\left[|x_k\left(n\right)|^m\right]\boldsymbol{I}$$

$$\preceq \left(8n\sqrt{n}\right)^m (m-1)!!\boldsymbol{I}$$

$$\preceq \frac{m!}{2}\left(8n\sqrt{n}\right)^m \boldsymbol{I},$$

where we have used the fact $q_n\left(\boldsymbol{w}\right) \geq \frac{1}{2\sqrt{n}}$ to simplify the result. Similar argument also shows $-\mathbb{E}\left[\boldsymbol{V}_k^m\right] \preceq m!\left(8n\sqrt{n}\right)^m \boldsymbol{I}/2$. Taking $R_V = 8n\sqrt{n}$ and $\boldsymbol{\sigma}_V^2 = 64n^3$, and letting $\boldsymbol{S}_V \doteq \frac{1}{p}\sum_{k=1}^p \boldsymbol{V}_k$, again by Lemma A.10, we obtain

$$\mathbb{P}\left[\|\boldsymbol{S}_V - \mathbb{E}\left[\boldsymbol{S}_V\right]\| \geq \frac{t}{2}\right] \;\leq\; 2n\exp\left(-\frac{pt^2}{512n^3 + 32n\sqrt{n}t}\right).$$

Combining the above results, we obtain

$$\mathbb{P}\left[\left\|\frac{1}{p}\sum_{k=1}^{p}\boldsymbol{Z}_k - \mathbb{E}\left[\boldsymbol{Z}_k\right]\right\| \geq t\right] = \mathbb{P}\left[\left\|\boldsymbol{S}_W - \mathbb{E}\left[\boldsymbol{S}_W\right] + \boldsymbol{S}_V - \mathbb{E}\left[\boldsymbol{S}_V\right]\right\| \geq t\right]$$

$$\leq \mathbb{P}\left[\left\|\boldsymbol{S}_W - \mathbb{E}\left[\boldsymbol{S}_W\right]\right\| \geq \frac{t}{2}\right] + \mathbb{P}\left[\left\|\boldsymbol{S}_V - \mathbb{E}\left[\boldsymbol{S}_V\right]\right\| \geq \frac{t}{2}\right]$$

$$\leq 2n\exp\left(-\frac{p\mu^2 t^2}{128n^2 + 16\mu n t}\right) + 2n\exp\left(-\frac{pt^2}{512n^3 + 32n\sqrt{n}t}\right)$$

$$\leq 4n\exp\left(-\frac{p\mu^2 t^2}{512n^2 + 32\mu n t}\right),$$

where we have simplified the final result based on the fact that $\mu \leq \frac{1}{\sqrt{n}}$. ∎

7.1.5 Proof of Lipschitz Results

To avoid clutter of notations, in this subsection we write $\boldsymbol{X}$ to mean $\boldsymbol{X}_0$; similarly $\boldsymbol{x}_k$ for $(\boldsymbol{x}_0)_k$, the k-th column of $\boldsymbol{X}_0$. The function $g(\boldsymbol{w})$ means $g(\boldsymbol{w}; \boldsymbol{X}_0)$. We need the following lemmas to prove the Lipschitz results.

Lemma 7.5 *Suppose that $\varphi_1 : U \to V$ is an L-Lipschitz map from a normed space U to a normed space V, and that $\varphi_2 : V \to W$ is an L'-Lipschitz map from V to a normed space W. Then the composition $\varphi_2 \circ \varphi_1 : U \to W$ is LL'-Lipschitz.*

Lemma 7.6 *Fix any $\mathcal{D} \subseteq \mathbb{R}^{n-1}$. Let $g_1, g_2 : \mathcal{D} \to \mathbb{R}$, and assume that g_1 is L_1-Lipschitz, and g_2 is L_2-Lipschitz, and that g_1 and g_2 are bounded over $\mathcal{D}$, i.e., $|g_1(\boldsymbol{x})| \leq M_1$ and $|g_2(\boldsymbol{x})| \leq M_2$ for all $x \in \mathcal{D}$ with some constants $M_1 > 0$ and $M_2 > 0$. Then the function $h(\boldsymbol{x}) = g_1(\boldsymbol{x})g_2(\boldsymbol{x})$ is L-Lipschitz, with*

$$L = M_1 L_2 + M_2 L_1.$$

Lemma 7.7 *For every $\boldsymbol{w}, \boldsymbol{w}' \in \Gamma$, and every fixed $\boldsymbol{x}$, we have*

$$\left|\dot{h}_\mu\left(\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}\right) - \dot{h}_\mu\left(\boldsymbol{q}(\boldsymbol{w}')^*\boldsymbol{x}\right)\right| \leq \frac{2\sqrt{n}}{\mu}\|\boldsymbol{x}\|\|\boldsymbol{w} - \boldsymbol{w}'\|,$$

$$\left|\ddot{h}_\mu\left(\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}\right) - \ddot{h}_\mu\left(\boldsymbol{q}(\boldsymbol{w}')^*\boldsymbol{x}\right)\right| \leq \frac{4\sqrt{n}}{\mu^2}\|\boldsymbol{x}\|\|\boldsymbol{w} - \boldsymbol{w}'\|.$$

Proof We have

$$\left|q_n(\boldsymbol{w}) - q_n(\boldsymbol{w}')\right| = \left|\sqrt{1 - \|\boldsymbol{w}\|^2} - \sqrt{1 - \|\boldsymbol{w}'\|^2}\right| = \frac{\|\boldsymbol{w} + \boldsymbol{w}'\|\|\boldsymbol{w} - \boldsymbol{w}'\|}{\sqrt{1 - \|\boldsymbol{w}\|^2} + \sqrt{1 - \|\boldsymbol{w}'\|^2}}$$

$$\leq \frac{\max\left(\|\boldsymbol{w}\|, \|\boldsymbol{w}'\|\right)}{\min\left(q_n(\boldsymbol{w}), q_n(\boldsymbol{w}')\right)}\|\boldsymbol{w} - \boldsymbol{w}'\|.$$

Hence it holds that

$$\left\|\boldsymbol{q}(\boldsymbol{w}) - \boldsymbol{q}(\boldsymbol{w}')\right\|^2 = \left\|\boldsymbol{w} - \boldsymbol{w}'\right\|^2 + \left|q_n(\boldsymbol{w}) - q_n(\boldsymbol{w}')\right|^2 \leq \left(1 + \frac{\max\left(\|\boldsymbol{w}\|^2, \|\boldsymbol{w}'\|^2\right)}{\min\left(q_n^2(\boldsymbol{w}), q_n^2(\boldsymbol{w}')\right)}\right)\left\|\boldsymbol{w} - \boldsymbol{w}'\right\|^2$$

$$= \frac{1}{\min\left(q_n^2\left(\boldsymbol{w}\right), q_n^2\left(\boldsymbol{w}'\right)\right)} \left\|\boldsymbol{w} - \boldsymbol{w}'\right\|^2 \leq 4n \left\|\boldsymbol{w} - \boldsymbol{w}'\right\|^2,$$

where we have used the fact $q_n\left(\boldsymbol{w}\right) \geq \frac{1}{2\sqrt{n}}$ to get the final result. Hence the mapping $\boldsymbol{w} \mapsto \boldsymbol{q}(\boldsymbol{w})$ is $2\sqrt{n}$-Lipschitz over Γ. Moreover it is easy to see $\boldsymbol{q} \mapsto \boldsymbol{q}^*\boldsymbol{x}$ is $\|\boldsymbol{x}\|_2$-Lipschitz. By Lemma A.1 and the composition rule in Lemma 7.5, we obtain the desired claims. ∎

Lemma 7.8 *For any fixed $\boldsymbol{x}$, consider the function*

$$t_{\boldsymbol{x}}(\boldsymbol{w}) \doteq \frac{\boldsymbol{w}^*\overline{\boldsymbol{x}}}{\|\boldsymbol{w}\|} - \frac{x_n}{q_n(\boldsymbol{w})} \|\boldsymbol{w}\|$$

defined over $\boldsymbol{w} \in \Gamma$. Then, for all $\boldsymbol{w}, \boldsymbol{w}'$ in Γ such that $\|\boldsymbol{w}\| \geq r$ and $\|\boldsymbol{w}'\| \geq r$ for some constant $r \in (0, 1)$, it holds that

$$\left|t_{\boldsymbol{x}}(\boldsymbol{w}) - t_{\boldsymbol{x}}(\boldsymbol{w}')\right| \leq 2\left(\frac{\|\boldsymbol{x}\|}{r} + 4n^{3/2} \|\boldsymbol{x}\|_\infty\right) \left\|\boldsymbol{w} - \boldsymbol{w}'\right\|,$$

$$\left|t_{\boldsymbol{x}}(\boldsymbol{w})\right| \leq 2\sqrt{n} \|\boldsymbol{x}\|,$$

$$\left|t_{\boldsymbol{x}}^2(\boldsymbol{w}) - t_{\boldsymbol{x}}^2(\boldsymbol{w}')\right| \leq 8\sqrt{n} \|\boldsymbol{x}\| \left(\frac{\|\boldsymbol{x}\|}{r} + 4n^{3/2} \|\boldsymbol{x}\|_\infty\right) \left\|\boldsymbol{w} - \boldsymbol{w}'\right\|,$$

$$\left|t_{\boldsymbol{x}}^2(\boldsymbol{w})\right| \leq 4n \|\boldsymbol{x}\|^2.$$

Proof First of all, we have

$$\left|t_{\boldsymbol{x}}(\boldsymbol{w})\right| = \left[\frac{\boldsymbol{w}^*}{\|\boldsymbol{w}\|}, -\frac{\|\boldsymbol{w}\|}{q_n(\boldsymbol{w})}\right] \boldsymbol{x} \leq \|\boldsymbol{x}\| \left(1 + \frac{\|\boldsymbol{w}\|^2}{q_n^2(\boldsymbol{w})}\right)^{1/2} = \frac{\|\boldsymbol{x}\|}{|q_n(\boldsymbol{w})|} \leq 2\sqrt{n} \|\boldsymbol{x}\|,$$

where we have used the assumption that $q_n\left(\boldsymbol{w}\right) \geq \frac{1}{2\sqrt{n}}$ to simplify the final result. The claim about $\left|t_{\boldsymbol{x}}^2\left(\boldsymbol{w}\right)\right|$ follows immediately. Now

$$\left|t_{\boldsymbol{x}}(\boldsymbol{w}) - t_{\boldsymbol{x}}(\boldsymbol{w}')\right| \leq \left|\left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} - \frac{\boldsymbol{w}'}{\|\boldsymbol{w}'\|}\right)^* \overline{\boldsymbol{x}}\right| + |x_n| \left|\frac{\|\boldsymbol{w}\|}{q_n(\boldsymbol{w})} - \frac{\|\boldsymbol{w}'\|}{q_n(\boldsymbol{w}')}\right|.$$

Moreover we have

$$\left|\left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} - \frac{\boldsymbol{w}'}{\|\boldsymbol{w}'\|}\right)^* \overline{\boldsymbol{x}}\right| \leq \|\overline{\boldsymbol{x}}\| \left\|\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} - \frac{\boldsymbol{w}'}{\|\boldsymbol{w}'\|}\right\| \leq \|\boldsymbol{x}\| \frac{\|\boldsymbol{w} - \boldsymbol{w}'\| \|\boldsymbol{w}'\| + \|\boldsymbol{w}'\| \|\|\boldsymbol{w}\| - \|\boldsymbol{w}'\|\|}{\|\boldsymbol{w}\| \|\boldsymbol{w}'\|}$$

$$\leq \frac{2 \|\boldsymbol{x}\|}{r} \left\|\boldsymbol{w} - \boldsymbol{w}'\right\|,$$

where we have used the assumption that $\|\boldsymbol{w}\| \geq r$ to simplify the result. Noticing that $t \mapsto t/\sqrt{1 - t^2}$ is continuous over $[a, b]$ and differentiable over (a, b) for any $0 < a < b < 1$, by mean value theorem,

$$\left|\frac{\|\boldsymbol{w}\|}{q_n(\boldsymbol{w})} - \frac{\|\boldsymbol{w}'\|}{q_n(\boldsymbol{w}')}\right| \leq \sup_{\boldsymbol{w} \in \Gamma} \frac{1}{\left(1 - \|\boldsymbol{w}\|^2\right)^{3/2}} \left\|\boldsymbol{w} - \boldsymbol{w}'\right\| \leq 8n^{3/2} \left\|\boldsymbol{w} - \boldsymbol{w}'\right\|,$$

where we have again used the assumption that $q_n(w) \geq \frac{1}{2\sqrt{n}}$ to simplify the last result. Collecting the above estimates, we obtain

$$\left|t_x(w) - t_x(w')\right| \leq \left(2\frac{\|x\|}{r} + 8n^{3/2}\|x\|_\infty\right)\|w - w'\|,$$

as desired. For the last one, we have

$$\begin{aligned}\left|t_x^2(w) - t_x^2(w')\right| &= \left|t_x(w) - t_x(w')\right|\left|t_x(w) + t_x(w')\right| \\ &\leq 2\sup_{s\in\Gamma}\left|t_x(s)\right|\left|t_x(w) - t_x(w')\right|,\end{aligned}$$

leading to the claimed result once we substitute estimates of the involved quantities. ∎

Lemma 7.9 *For any fixed x, consider the function*

$$\Phi_x(w) = \frac{x_n}{q_n(w)}I + \frac{x_n}{q_n^3(w)}ww^*$$

defined over $w \in \Gamma$. Then, for all $w, w' \in \Gamma$ such that $\|w\| < r$ and $\|w'\| < r$ with some constant $r \in \left(0, \frac{1}{2}\right)$, it holds that

$$\begin{aligned}\|\Phi_x(w)\| &\leq 2\|x\|_\infty, \\ \|\Phi_x(w) - \Phi_x(w')\| &\leq 4\|x\|_\infty\|w - w'\|.\end{aligned}$$

Proof Simple calculation shows

$$\|\Phi_x(w)\| \leq \|x\|_\infty\left(\frac{1}{q_n(w)} + \frac{\|w\|^2}{q_n^3(w)}\right) = \frac{\|x\|_\infty}{q_n^3(w)} \leq \frac{\|x\|_\infty}{(1 - r^2)^{3/2}} \leq 2\|x\|_\infty.$$

For the second one, we have

$$\begin{aligned}\|\Phi_x(w) - \Phi_x(w')\| &\leq \|x\|_\infty\left\|\frac{1}{q_n(w)}I + \frac{1}{q_n^3(w)}ww^* - \frac{1}{q_n(w')}I - \frac{1}{q_n^3(w')}w'(w')^*\right\| \\ &\leq \|x\|_\infty\left(\left|\frac{1}{q_n(w)} - \frac{1}{q_n(w')}\right| + \left|\frac{\|w\|^2}{q_n^3(w)} - \frac{\|w'\|^2}{q_n^3(w')}\right|\right).\end{aligned}$$

Now

$$\left|\frac{1}{q_n(w)} - \frac{1}{q_n(w')}\right| = \frac{\left|q_n(w) - q_n(w')\right|}{q_n(w)q_n(w')} \leq \frac{\max(\|w\|, \|w'\|)}{\min(q_n^3(w), q_n^3(w'))}\|w - w'\| \leq \frac{4}{3\sqrt{3}}\|w - w'\|,$$

where we have applied the estimate for $|q_n(w) - q_n(w')|$ as established in Lemma 7.7 and also used $\|w\| \leq 1/2$ and $\|w'\| \leq 1/2$ to simplify the above result. Further noticing $t \mapsto t^2/\left(1 - t^2\right)^{3/2}$ is differentiable over $t \in (0, 1)$, we apply the mean value theorem and obtain

$$\left|\frac{\|w\|^2}{q_n^3(w)} - \frac{\|w'\|^2}{q_n^3(w')}\right| \leq \sup_{s\in\Gamma, \|s\|\leq r<\frac{1}{2}}\frac{\|s\|^3 + 2\|s\|}{\left(1 - \|s\|^2\right)^{5/2}}\|w - w'\| \leq \frac{4}{\sqrt{3}}\|w - w'\|.$$

Combining the above estimates gives the claimed result. ∎

Lemma 7.10 *For any fixed $\boldsymbol{x}$, consider the function*

$$\boldsymbol{\zeta}_{\boldsymbol{x}}(\boldsymbol{w}) = \overline{\boldsymbol{x}} - \frac{x_n}{q_n(\boldsymbol{w})}\boldsymbol{w}$$

defined over $\boldsymbol{w} \in \Gamma$. Then, for all $\boldsymbol{w}, \boldsymbol{w}' \in \Gamma$ such that $\|\boldsymbol{w}\| \leq r$ and $\|\boldsymbol{w}'\| \leq r$ for some constant $r \in \left(0, \frac{1}{2}\right)$, it holds that

$$\|\boldsymbol{\zeta}_{\boldsymbol{x}}(\boldsymbol{w})\boldsymbol{\zeta}_{\boldsymbol{x}}(\boldsymbol{w})^*\| \leq 2n\|\boldsymbol{x}\|_\infty^2,$$

$$\left\|\boldsymbol{\zeta}_{\boldsymbol{x}}(\boldsymbol{w})\boldsymbol{\zeta}_{\boldsymbol{x}}(\boldsymbol{w})^* - \boldsymbol{\zeta}_{\boldsymbol{x}}(\boldsymbol{w}')\boldsymbol{\zeta}_{\boldsymbol{x}}(\boldsymbol{w}')^*\right\| \leq 8\sqrt{2}\sqrt{n}\|\boldsymbol{x}\|_\infty^2 \|\boldsymbol{w} - \boldsymbol{w}'\|.$$

Proof We have $\|\boldsymbol{w}\|^2/q_n^2(\boldsymbol{w}) \leq 1/3$ when $\|\boldsymbol{w}\| \leq r < 1/2$, hence it holds that

$$\|\boldsymbol{\zeta}_{\boldsymbol{x}}(\boldsymbol{w})\boldsymbol{\zeta}_{\boldsymbol{x}}(\boldsymbol{w})^*\| \leq \|\boldsymbol{\zeta}_{\boldsymbol{x}}(\boldsymbol{w})\|^2 \leq 2\|\overline{\boldsymbol{x}}\|^2 + 2x_n^2\frac{\|\boldsymbol{w}\|}{q_n^2(\boldsymbol{w})} \leq 2n\|\boldsymbol{x}\|_\infty^2.$$

For the second, we first estimate

$$
\begin{aligned}
\|\boldsymbol{\zeta}(\boldsymbol{w}) - \boldsymbol{\zeta}(\boldsymbol{w}')\| &= \left\|x_n\left(\frac{\boldsymbol{w}}{q_n(\boldsymbol{w})} - \frac{\boldsymbol{w}'}{q_n(\boldsymbol{w}')}\right)\right\| \leq \|\boldsymbol{x}\|_\infty \left\|\frac{\boldsymbol{w}}{q_n(\boldsymbol{w})} - \frac{\boldsymbol{w}'}{q_n(\boldsymbol{w}')}\right\| \\
&\leq \|\boldsymbol{x}\|_\infty \left(\frac{1}{q_n(\boldsymbol{w})}\|\boldsymbol{w} - \boldsymbol{w}'\| + \|\boldsymbol{w}'\|\left|\frac{1}{q_n(\boldsymbol{w})} - \frac{1}{q_n(\boldsymbol{w}')}\right|\right) \\
&\leq \|\boldsymbol{x}\|_\infty \left(\frac{1}{q_n(\boldsymbol{w})} + \frac{\|\boldsymbol{w}'\|}{\min\{q_n^3(\boldsymbol{w}), q_n^3(\boldsymbol{w}')\}}\right)\|\boldsymbol{w} - \boldsymbol{w}'\| \\
&\leq \|\boldsymbol{x}\|_\infty \left(\frac{2}{\sqrt{3}} + \frac{4}{3\sqrt{3}}\right)\|\boldsymbol{w} - \boldsymbol{w}'\| \leq 4\|\boldsymbol{x}\|_\infty\|\boldsymbol{w} - \boldsymbol{w}'\|.
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
\left\|\boldsymbol{\zeta}_{\boldsymbol{x}}(\boldsymbol{w})\boldsymbol{\zeta}_{\boldsymbol{x}}(\boldsymbol{w})^* - \boldsymbol{\zeta}_{\boldsymbol{x}}(\boldsymbol{w}')\boldsymbol{\zeta}_{\boldsymbol{x}}(\boldsymbol{w}')^*\right\| &\leq \|\boldsymbol{\zeta}(\boldsymbol{w})\|\|\boldsymbol{\zeta}(\boldsymbol{w}) - \boldsymbol{\zeta}(\boldsymbol{w}')\| + \|\boldsymbol{\zeta}(\boldsymbol{w}) - \boldsymbol{\zeta}(\boldsymbol{w}')\|\|\boldsymbol{\zeta}(\boldsymbol{w}')\| \\
&\leq 8\sqrt{2}\sqrt{n}\|\boldsymbol{x}\|_\infty^2\|\boldsymbol{w} - \boldsymbol{w}'\|,
\end{aligned}
$$

as desired. ∎

Now, we are ready to prove all the Lipschitz propositions.

Proof [of Proposition 2.11] Let

$$F_k(\boldsymbol{w}) = \ddot{h}_\mu\left(\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}_k\right)t_{\boldsymbol{x}_k}^2(\boldsymbol{w}) + \dot{h}_\mu\left(\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}_k\right)\frac{x_k(n)}{q_n^3(\boldsymbol{w})}.$$

Then, $\frac{1}{\|\boldsymbol{w}\|^2}\boldsymbol{w}^*\nabla^2 g(\boldsymbol{w})\boldsymbol{w} = \frac{1}{p}\sum_{k=1}^p F_k(\boldsymbol{w})$. Noticing that $\ddot{h}_\mu\left(\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}_k\right)$ is bounded by $1/\mu$ and $\dot{h}_\mu\left(\boldsymbol{q}(\boldsymbol{w})^*\boldsymbol{x}_k\right)$ is bounded by 1, both in magnitude. Applying Lemma 7.6, Lemma 7.7 and Lemma 7.8, we can see $F_k(\boldsymbol{w})$ is $L_\frown^k$-Lipschitz with

$$
\begin{aligned}
L_\frown^k = {} & 4n\|\boldsymbol{x}_k\|^2\frac{4\sqrt{n}}{\mu^2}\|\boldsymbol{x}_k\| + \frac{1}{\mu}8\sqrt{n}\|\boldsymbol{x}_k\|\left(\frac{\|\boldsymbol{x}_k\|}{r_\frown} + 4n^{3/2}\|\boldsymbol{x}_k\|_\infty\right) \\
& + (2\sqrt{n})^3\|\boldsymbol{x}_k\|_\infty\frac{2\sqrt{n}}{\mu}\|\boldsymbol{x}_k\| + \sup_{r_\frown < a < \sqrt{\frac{2n-1}{2n}}}\frac{3}{(1-a^2)^{5/2}}\|\boldsymbol{x}_k\|_\infty
\end{aligned}
$$

$$= \frac{16n^{3/2}}{\mu^2} \|\boldsymbol{x}_k\|^3 + \frac{8\sqrt{n}}{\mu r_\cap} \|\boldsymbol{x}_k\|^2 + \frac{48n^2}{\mu} \|\boldsymbol{x}_k\| \|\boldsymbol{x}_k\|_\infty + 96n^{5/2} \|\boldsymbol{x}_k\|_\infty .$$

Thus, $\frac{1}{\|\boldsymbol{w}\|_2} \boldsymbol{w}^* \nabla^2 g(\boldsymbol{w}) \boldsymbol{w}$ is $L_\cap$-Lipschitz with

$$L_\cap \leq \frac{1}{p} \sum_{k=1}^p L_\cap^k \leq \frac{16n^3}{\mu^2} \|\boldsymbol{X}\|_\infty^3 + \frac{8n^{3/2}}{\mu r_\cap} \|\boldsymbol{X}\|_\infty^2 + \frac{48n^{5/2}}{\mu} \|\boldsymbol{X}\|_\infty^2 + 96n^{5/2} \|\boldsymbol{X}\|_\infty ,$$

as desired. ∎

Proof [of Proposition 2.12] We have

$$\left\| \frac{\boldsymbol{w}^*}{\|\boldsymbol{w}\|} \nabla g(\boldsymbol{w}) - \frac{\boldsymbol{w}'^*}{\|\boldsymbol{w}'\|} \nabla g(\boldsymbol{w}') \right\| \leq \frac{1}{p} \sum_{k=1}^p \left\| \dot{h}_\mu \left(\boldsymbol{q}(\boldsymbol{w})^* \boldsymbol{x}_k\right) t_{\boldsymbol{x}_k}(\boldsymbol{w}) - \dot{h}_\mu \left(\boldsymbol{q}(\boldsymbol{w}')^* \boldsymbol{x}_k\right) t_{\boldsymbol{x}_k}(\boldsymbol{w}') \right\|$$

where $\dot{h}_\mu(t) = \tanh(t/\mu)$ is bounded by one in magnitude, and $t_{\boldsymbol{x}_k}(\boldsymbol{w})$ and $t_{\boldsymbol{x}'_k}(\boldsymbol{w})$ is defined as in Lemma 7.8. By Lemma 7.6, Lemma 7.7 and Lemma 7.8, we know that $\dot{h}_\mu \left(\boldsymbol{q}(\boldsymbol{w})^* \boldsymbol{x}_k\right) t_{\boldsymbol{x}_k}(\boldsymbol{w})$ is L_k-Lipschitz with constant

$$L_k = \frac{2 \|\boldsymbol{x}_k\|}{r_g} + 8n^{3/2} \|\boldsymbol{x}_k\|_\infty + \frac{4n}{\mu} \|\boldsymbol{x}_k\|^2 .$$

Therefore, we have

$$\left\| \frac{\boldsymbol{w}^*}{\|\boldsymbol{w}\|} \nabla g(\boldsymbol{w}) - \frac{\boldsymbol{w}^*}{\|\boldsymbol{w}\|} \nabla g(\boldsymbol{w}') \right\| \leq \frac{1}{p} \sum_{k=1}^p \left(\frac{2 \|\boldsymbol{x}_k\|}{r_g} + 8n^{3/2} \|\boldsymbol{x}_k\|_\infty + \frac{4n}{\mu} \|\boldsymbol{x}_k\|^2 \right) \|\boldsymbol{w} - \boldsymbol{w}'\|$$

$$\leq \left(\frac{2\sqrt{n}}{r_g} \|\boldsymbol{X}\|_\infty + 8n^{3/2} \|\boldsymbol{X}\|_\infty + \frac{4n^2}{\mu} \|\boldsymbol{X}\|_\infty^2 \right) \|\boldsymbol{w} - \boldsymbol{w}'\| ,$$

as desired. ∎

Proof [of Proposition 2.13] Let

$$\boldsymbol{F}_k(\boldsymbol{w}) = \ddot{h}_\mu(\boldsymbol{q}(\boldsymbol{w})^* \boldsymbol{x}_k) \boldsymbol{\zeta}_k(\boldsymbol{w}) \boldsymbol{\zeta}_k(\boldsymbol{w})^* - \dot{h}_\mu \left(\boldsymbol{q}(\boldsymbol{w})^* \boldsymbol{x}_k\right) \boldsymbol{\Phi}_k(\boldsymbol{w})$$

with $\boldsymbol{\zeta}_k(\boldsymbol{w}) = \overline{\boldsymbol{x}}_k - \frac{x_k(n)}{q_n(\boldsymbol{w})} \boldsymbol{w}$ and $\boldsymbol{\Phi}_k(\boldsymbol{w}) = \frac{x_k(n)}{q_n(\boldsymbol{w})} \boldsymbol{I} + \frac{x_{n,k}}{q_n(\boldsymbol{w})} \boldsymbol{w} \boldsymbol{w}^*$. Then, $\nabla^2 g(\boldsymbol{w}) = \frac{1}{p} \sum_{k=1}^p \boldsymbol{F}_k(\boldsymbol{w})$. Using Lemma 7.6, Lemma 7.7, Lemma 7.9 and Lemma 7.10, and the facts that $\ddot{h}_\mu(t)$ is bounded by $1/\mu$ and that $\ddot{h}_\mu(t)$ is bounded by 1 in magnitude, we can see $\boldsymbol{F}_k(\boldsymbol{w})$ is $L_\smile^k$-Lipschitz continuous with

$$L_\smile^k = \frac{1}{\mu} \times 8\sqrt{2}\sqrt{n} \|\boldsymbol{x}_k\|_\infty^2 + \frac{2\sqrt{n}}{\mu^2} \|\boldsymbol{x}_k\| \times 2n \|\boldsymbol{x}_k\|_\infty^2 + 4 \|\boldsymbol{x}_k\|_\infty + \frac{2\sqrt{n}}{\mu} \|\boldsymbol{x}_k\| \times 2 \|\boldsymbol{x}_k\|_\infty$$

$$\leq \frac{4n^{3/2}}{\mu^2} \|\boldsymbol{x}_k\| \|\boldsymbol{x}_k\|_\infty^2 + \frac{4\sqrt{n}}{\mu} \|\boldsymbol{x}_k\| \|\boldsymbol{x}_k\|_\infty + \frac{8\sqrt{2}\sqrt{n}}{\mu} \|\boldsymbol{x}_k\|_\infty^2 + 4 \|\boldsymbol{x}_k\|_\infty .$$

Thus, we have

$$L_\smile \leq \frac{1}{p} \sum_{k=1}^p L_\smile^k \leq \frac{4n^2}{\mu^2} \|\boldsymbol{X}\|_\infty^3 + \frac{4n}{\mu} \|\boldsymbol{X}\|_\infty^2 + \frac{8\sqrt{2}\sqrt{n}}{\mu} \|\boldsymbol{X}\|_\infty^2 + 8 \|\boldsymbol{X}\|_\infty ,$$

as desired. ∎

7.2 Proofs of Theorem 2.1

To avoid clutter of notations, in this subsection we write $\boldsymbol{X}$ to mean $\boldsymbol{X}_0$; similarly $\boldsymbol{x}_k$ for $(\boldsymbol{x}_0)_k$, the k-th column of $\boldsymbol{X}_0$. The function $g(\boldsymbol{w})$ means $g(\boldsymbol{w}; \boldsymbol{X}_0)$. Before proving Theorem 2.1, we record one useful lemma.

Lemma 7.11 *For any $\theta \in (0, 1)$, consider the random matrix $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$ with $\boldsymbol{X} \sim_{i.i.d.} \mathrm{BG}(\theta)$. Define the event $\mathcal{E}_\infty \doteq \left\{ 1 \leq \|\boldsymbol{X}\|_\infty \leq 4\sqrt{\log(np)} \right\}$. It holds that*

$$\mathbb{P}\left[\mathcal{E}_\infty^c\right] \leq \theta(np)^{-7} + \exp\left(-0.3\theta np\right).$$

For convenience, we define three regions for the range of $\boldsymbol{w}$:

$$R_1 \doteq \left\{ \boldsymbol{w} \;\middle|\; \|\boldsymbol{w}\| \leq \frac{\mu}{4\sqrt{2}} \right\}, \qquad R_2 \doteq \left\{ \boldsymbol{w} \;\middle|\; \frac{\mu}{4\sqrt{2}} \leq \|\boldsymbol{w}\| \leq \frac{1}{20\sqrt{5}} \right\},$$

$$R_3 \doteq \left\{ \boldsymbol{w} \;\middle|\; \frac{1}{20\sqrt{5}} \leq \|\boldsymbol{w}\| \leq \sqrt{\frac{4n-1}{4n}} \right\}.$$

Proof [of Theorem 2.1] We will focus on deriving the qualitative result and hence be sloppy about constants. All indexed capital C or small c are numerical constants.

Strong convexity in region R_1. Proposition 2.7 shows that for any $\boldsymbol{w} \in R_1$, $\mathbb{E}\left[\nabla^2 g(\boldsymbol{w})\right] \succeq \frac{c_1 \theta}{\mu} \boldsymbol{I}$. For any $\varepsilon \in (0, \mu/(4\sqrt{2}))$, R_1 has an ε-net N_1 of size at most $(3\mu/(4\sqrt{2}\varepsilon))^n$. On $\mathcal{E}_\infty$, $\nabla^2 g$ is

$$L_1 \doteq \frac{C_2 n^2}{\mu^2} \log^{3/2}(np)$$

Lipschitz by Proposition 2.13. Set $\varepsilon = \frac{c_1 \theta}{3\mu L_1}$, so

$$\#N_1 \leq \exp\left(2n \log\left(\frac{C_3 n \log(np)}{\theta}\right)\right).$$

Let $\mathcal{E}_1$ denote the event

$$\mathcal{E}_1 = \left\{ \max_{\boldsymbol{w} \in N_1} \left\|\nabla^2 g(\boldsymbol{w}) - \mathbb{E}\left[\nabla^2 g(\boldsymbol{w})\right]\right\| \leq \frac{c_1 \theta}{3\mu} \right\}.$$

On $\mathcal{E}_1 \cap \mathcal{E}_\infty$,

$$\sup_{\|\boldsymbol{w}\| \leq \mu/(4\sqrt{2})} \left\|\nabla^2 g(\boldsymbol{w}) - \mathbb{E}\left[\nabla^2 g(\boldsymbol{w})\right]\right\| \leq \frac{2c_1 \theta}{3\mu},$$

and so on $\mathcal{E}_1 \cap \mathcal{E}_\infty$, (2.4) holds for any constant $c_\star \leq c_1/3$. Setting $t = c_1\theta/3\mu$ in Proposition 2.10, we obtain that for any fixed $\boldsymbol{w}$,

$$\mathbb{P}\left[\left\|\nabla^2 g(\boldsymbol{w}) - \mathbb{E}\left[\nabla^2 g(\boldsymbol{w})\right]\right\| \geq \frac{c_1 \theta}{3\mu}\right] \leq 4n \exp\left(-\frac{c_4 p \theta^2}{n^2}\right).$$

Taking a union bound, we obtain that

$$\mathbb{P}\left[\mathcal{E}_1^c\right] \leq 4n \exp\left(-\frac{c_4 p \theta^2}{n^2} + C_5 n \log(n) + C_5 n \log\log(p)\right).$$

Large gradient in region R_2. Similarly, for the gradient quantity, for $\boldsymbol{w} \in R_2$, Proposition 2.6 shows that

$$\mathbb{E}\left[\frac{\boldsymbol{w}^* \nabla g(\boldsymbol{w})}{\|\boldsymbol{w}\|}\right] \geq c_6 \theta.$$

Moreover, on $\mathcal{E}_\infty$, $\frac{\boldsymbol{w}^* \nabla g(\boldsymbol{w})}{\|\boldsymbol{w}\|}$ is

$$L_2 \doteq \frac{C_7 n^2}{\mu} \log(np)$$

Lipschitz by Proposition 2.12. For any $\varepsilon < \frac{1}{20\sqrt{5}}$, the set R_2 has an ε-net N_2 of size at most $\left(\frac{3}{20\varepsilon\sqrt{5}}\right)^n$.
Set $\varepsilon = \frac{c_6 \theta}{3 L_2}$, so

$$\# N_2 \leq \exp\left(n \log\left(\frac{C_8 n^2 \log(np)}{\theta \mu}\right)\right).$$

Let $\mathcal{E}_2$ denote the event

$$\mathcal{E}_2 = \left\{\max_{\boldsymbol{w} \in N_2} \left|\frac{\boldsymbol{w}^* \nabla g(\boldsymbol{w})}{\|\boldsymbol{w}\|} - \mathbb{E}\left[\frac{\boldsymbol{w}^* \nabla g(\boldsymbol{w})}{\|\boldsymbol{w}\|}\right]\right| \leq \frac{c_6 \theta}{3}\right\}.$$

On $\mathcal{E}_2 \cap \mathcal{E}_\infty$,

$$\sup_{\boldsymbol{w} \in R_2} \left|\frac{\boldsymbol{w}^* \nabla g(\boldsymbol{w})}{\|\boldsymbol{w}\|} - \mathbb{E}\left[\frac{\boldsymbol{w}^* \nabla g(\boldsymbol{w})}{\|\boldsymbol{w}\|}\right]\right| \leq \frac{2c_6 \theta}{3}, \tag{7.8}$$

and so on $\mathcal{E}_2 \cap \mathcal{E}_\infty$, (2.5) holds for any constant $c_\star \leq c_6/3$. Setting $t = c_6 \theta/3$ in Proposition 2.9, we obtain that for any fixed $\boldsymbol{w} \in R_2$,

$$\mathbb{P}\left[\left|\frac{\boldsymbol{w}^* \nabla g(\boldsymbol{w})}{\|\boldsymbol{w}\|} - \mathbb{E}\left[\frac{\boldsymbol{w}^* \nabla g(\boldsymbol{w})}{\|\boldsymbol{w}\|}\right|\right]\right] \leq 2\exp\left(-\frac{c_9 p \theta^2}{n}\right),$$

and so

$$\mathbb{P}\left[\mathcal{E}_2^c\right] \leq 2\exp\left(-\frac{c_9 p \theta^2}{n} + n \log\left(\frac{C_8 n^2 \log(np)}{\theta \mu}\right)\right). \tag{7.9}$$

Existence of negative curvature direction in R_3. Finally, for any $\boldsymbol{w} \in R_3$, Proposition 2.5 shows that

$$\mathbb{E}\left[\frac{\boldsymbol{w}^* \nabla^2 g(\boldsymbol{w}) \boldsymbol{w}}{\|\boldsymbol{w}\|^2}\right] \leq -c_9 \theta.$$

On $\mathcal{E}_\infty$, $\frac{\boldsymbol{w}^* \nabla^2 g(\boldsymbol{w}) \boldsymbol{w}}{\|\boldsymbol{w}\|^2}$ is

$$L_3 = \frac{C_{10} n^3}{\mu^2} \log^{3/2}(np)$$

Lipschitz by Proposition 2.11. As above, for any $\varepsilon \leq \sqrt{\frac{4n-1}{4n}}$, R_3 has an ε-net N_3 of size at most $(3/\varepsilon)^n$. Set $\varepsilon = c_9 \theta/3 L_3$. Then

$$\# N_3 \leq \exp\left(n \log\left(\frac{C_{11} n^3 \log^{3/2}(np)}{\theta \mu^2}\right)\right).$$

Let $\mathcal{E}_3$ denote the event

$$\mathcal{E}_3 = \left\{ \max_{\boldsymbol{w} \in N_3} \left| \frac{\boldsymbol{w}^* \nabla^2 g(\boldsymbol{w}) \boldsymbol{w}}{\|\boldsymbol{w}\|^2} - \mathbb{E}\left[\frac{\boldsymbol{w}^* \nabla^2 g(\boldsymbol{w}) \boldsymbol{w}}{\|\boldsymbol{w}\|^2} \right] \right| \leq \frac{c_9 \theta}{3} \right\}$$

On $\mathcal{E}_3 \cap \mathcal{E}_\infty$,

$$\sup_{\boldsymbol{w} \in R_3} \left| \frac{\boldsymbol{w}^* \nabla^2 g(\boldsymbol{w}) \boldsymbol{w}}{\|\boldsymbol{w}\|^2} - \mathbb{E}\left[\frac{\boldsymbol{w}^* \nabla^2 g(\boldsymbol{w}) \boldsymbol{w}}{\|\boldsymbol{w}\|^2} \right] \right| \leq \frac{2 c_9 \theta}{3},$$

and (2.6) holds with any constant $c_\star < c_9/3$. Setting $t = c_9\theta/3$ in Proposition 2.8 and taking a union bound, we obtain

$$\mathbb{P}\left[\mathcal{E}_3^c \right] \leq 4 \exp\left(-\frac{c_{12} p \mu^2 \theta^2}{n^2} + n \log\left(\frac{C_{11} n^3 \log^{3/2}(np)}{\theta \mu^2} \right) \right).$$

The unique local minimizer located near 0. Let $\mathcal{E}_g$ be the event that the bounds (2.4)-(2.6) hold. On $\mathcal{E}_g$, the function g is $\frac{c_\star \theta}{\mu}$-strongly convex over $R_1 = \left\{ \boldsymbol{w} \mid \|\boldsymbol{w}\| \leq \mu/\left(4\sqrt{2}\right) \right\}$. This implies that f has at most one local minimum on R_1. It also implies that for any $\boldsymbol{w} \in R_1$,

$$g(\boldsymbol{w}) \geq g(\boldsymbol{0}) + \langle \nabla g(\boldsymbol{0}), \boldsymbol{w} \rangle + \frac{c\theta}{2\mu} \|\boldsymbol{w}\|^2 \geq g(\boldsymbol{0}) - \|\boldsymbol{w}\| \|\nabla g(\boldsymbol{0})\| + \frac{c_\star \theta}{2\mu} \|\boldsymbol{w}\|^2.$$

So, if $g(\boldsymbol{w}) \leq g(\boldsymbol{0})$, we necessarily have

$$\|\boldsymbol{w}\| \leq \frac{2\mu}{c_\star \theta} \|\nabla g(\boldsymbol{0})\|.$$

Suppose that

$$\|\nabla g(\boldsymbol{0})\| \leq \frac{c_\star \theta}{32}. \tag{7.10}$$

Then $g(\boldsymbol{w}) \leq g(\boldsymbol{0})$ implies that $\|\boldsymbol{w}\| \leq \mu/16$. By Wierstrass's theorem, $g(\boldsymbol{w})$ has at least one minimizer $\boldsymbol{w}_\star$ over the compact set $S = \{ \boldsymbol{w} \mid \|\boldsymbol{w}\| \leq \mu/10 \}$. By the above reasoning, $\|\boldsymbol{w}_\star\| \leq \mu/16$, and hence $\boldsymbol{w}_\star$ does not lie on the boundary of S. This implies that $\boldsymbol{w}_\star$ is a local minimizer of g. Moreover, as above,

$$\|\boldsymbol{w}_\star\| \leq \frac{2\mu}{c_\star \theta} \|\nabla g(\boldsymbol{0})\|.$$

We now use the vector Bernstein inequality to show that with our choice of p, (7.10) is satisifed with high probability. Notice that

$$\nabla g(\boldsymbol{0}) = \frac{1}{p} \sum_{i=1}^{p} \dot{h}_\mu(x_i(n)) \overline{\boldsymbol{x}}_i,$$

and $\dot{h}_\mu$ is bounded by one in magnitude, so for any integer $m \geq 2$,

$$\mathbb{E}\left[\left\| \dot{h}_\mu(x_i(n)) \overline{\boldsymbol{x}}_i \right\|^m \right] \leq \mathbb{E}\left[\|\boldsymbol{x}_i\|^m \right] \leq \mathbb{E}_{Z \sim \chi(n)}\left[Z^m \right] \leq m! n^{m/2},$$

where we have applied the moment estimate for the $\chi(n)$ distribution shown in Lemma A.8. Applying the vector Bernstein inequality in Corollary A.11 with $R = \sqrt{n}$ and $\sigma^2 = 2n$, we obtain

$$\mathbb{P}\left[\|\nabla g(\mathbf{0})\| \geq t\right] \leq 2(n+1)\exp\left(-\frac{pt^2}{4n + 2\sqrt{n}t}\right)$$

for all $t > 0$. Using this inequality, it is not difficult to show that there exist constants $C_{13}, C_{14} > 0$ such that when $p \geq C_{13}n\log n$, with probability at least $1 - 4np^{-10}$,

$$\|\nabla g(\mathbf{0})\| \leq C_3\sqrt{\frac{n\log p}{p}}. \tag{7.11}$$

When $\frac{p}{\log p} \geq \frac{C_{14}n}{\theta^2}$, for appropriately large C_{14}, (7.11) implies (7.10). Summing up failure probabilities completes the proof. ∎

7.3 Proofs for Section 2.3 and Theorem 2.3

Proof [of Lemma 2.14] By the generative model,

$$\overline{\mathbf{Y}} = \left(\frac{1}{p\theta}\mathbf{Y}\mathbf{Y}^*\right)^{-1/2}\mathbf{Y} = \left(\frac{1}{p\theta}\mathbf{A}_0\mathbf{X}_0\mathbf{X}_0^*\mathbf{A}_0^*\right)^{-1/2}\mathbf{A}_0\mathbf{X}_0.$$

Since $\mathbb{E}\left[\mathbf{X}_0\mathbf{X}_0^*/(p\theta)\right] = \mathbf{I}$, we will compare $\left(\frac{1}{p\theta}\mathbf{A}_0\mathbf{X}_0\mathbf{X}_0^*\mathbf{A}_0^*\right)^{-1/2}\mathbf{A}_0$ with $\left(\mathbf{A}_0\mathbf{A}_0^*\right)^{-1/2}\mathbf{A}_0 = \mathbf{U}\mathbf{V}^*$. By Lemma B.2, we have

$$\left\|\left(\frac{1}{p\theta}\mathbf{A}_0\mathbf{X}_0\mathbf{X}_0^*\mathbf{A}_0^*\right)^{-1/2}\mathbf{A}_0 - \left(\mathbf{A}_0\mathbf{A}_0^*\right)^{-1/2}\mathbf{A}_0\right\|$$

$$\leq \|\mathbf{A}_0\|\left\|\left(\frac{1}{p\theta}\mathbf{A}_0\mathbf{X}_0\mathbf{X}_0^*\mathbf{A}_0^*\right)^{-1/2} - \left(\mathbf{A}_0\mathbf{A}_0^*\right)^{-1/2}\right\|$$

$$\leq \|\mathbf{A}_0\|\frac{2\|\mathbf{A}_0\|^3}{\sigma_{\min}^4(\mathbf{A}_0)}\left\|\frac{1}{p\theta}\mathbf{X}_0\mathbf{X}_0^* - \mathbf{I}\right\| = 2\kappa^4(\mathbf{A}_0)\left\|\frac{1}{p\theta}\mathbf{X}_0\mathbf{X}_0^* - \mathbf{I}\right\|$$

provided

$$\|\mathbf{A}_0\|^2\left\|\frac{1}{p\theta}\mathbf{X}_0\mathbf{X}_0^* - \mathbf{I}\right\| \leq \frac{\sigma_{\min}^2(\mathbf{A}_0)}{2} \iff \left\|\frac{1}{p\theta}\mathbf{X}_0\mathbf{X}_0^* - \mathbf{I}\right\| \leq \frac{1}{2\kappa^2(\mathbf{A}_0)}.$$

On the other hand, by Lemma B.3, when $p \geq C_1 n^2 \log n$ for some large constant C_1, $\left\|\frac{1}{p\theta}\mathbf{X}_0\mathbf{X}_0^* - \mathbf{I}\right\| \leq 10\sqrt{\frac{\theta n\log p}{p}}$ with probability at least $1 - p^{-8}$. Thus, when $p \geq C_2\kappa^4(\mathbf{A}_0)\theta n^2\log(n\theta\kappa(\mathbf{A}_0))$,

$$\left\|\left(\frac{1}{p\theta}\mathbf{A}_0\mathbf{X}_0\mathbf{X}_0^*\mathbf{A}_0^*\right)^{-1/2}\mathbf{A}_0 - \left(\mathbf{A}_0\mathbf{A}_0^*\right)^{-1/2}\mathbf{A}_0\right\| \leq 20\kappa^4(\mathbf{A}_0)\sqrt{\frac{\theta n\log p}{p}},$$

as desired. ∎

Proof [of Lemma 2.15] To avoid clutter in notation, we write $\boldsymbol{X}$ to mean $\boldsymbol{X}_0$, and $\boldsymbol{x}_k$ to mean $(\boldsymbol{x}_0)_k$ in this proof. We also let $\widetilde{\boldsymbol{Y}} \doteq \boldsymbol{X}_0 + \widetilde{\boldsymbol{\Xi}}\boldsymbol{X}_0$. Note the Jacobian matrix for the mapping $\boldsymbol{q}(\boldsymbol{w})$ is $\nabla_{\boldsymbol{w}}\boldsymbol{q}(\boldsymbol{w}) = \left[\boldsymbol{I}, -\boldsymbol{w}/\sqrt{1 - \|\boldsymbol{w}\|^2}\right]$. Hence for any vector $\boldsymbol{z} \in \mathbb{R}^n$ and all $\boldsymbol{w} \in \Gamma$,

$$\|\nabla_{\boldsymbol{w}}\boldsymbol{q}(\boldsymbol{w})\boldsymbol{z}\| \leq \sqrt{n-1}\|\boldsymbol{z}\|_\infty + \frac{\|\boldsymbol{w}\|}{\sqrt{1 - \|\boldsymbol{w}\|^2}}\|\boldsymbol{z}\|_\infty \leq 3\sqrt{n}\|\boldsymbol{z}\|_\infty.$$

Now we have

$$\left\|\nabla_{\boldsymbol{w}}g\left(\boldsymbol{w}; \widetilde{\boldsymbol{Y}}\right) - \nabla_{\boldsymbol{w}}g\left(\boldsymbol{w}; \boldsymbol{X}\right)\right\|$$

$$= \left\|\frac{1}{p}\sum_{k=1}^p \dot{h}_\mu\left(\boldsymbol{q}^*(\boldsymbol{w})\boldsymbol{x}_k + \boldsymbol{q}^*(\boldsymbol{w})\widetilde{\boldsymbol{\Xi}}\boldsymbol{x}_k\right)\nabla_{\boldsymbol{w}}\boldsymbol{q}(\boldsymbol{w})\left(\boldsymbol{x}_k + \widetilde{\boldsymbol{\Xi}}\boldsymbol{x}_k\right) - \frac{1}{p}\sum_{k=1}^p \dot{h}_\mu\left(\boldsymbol{q}^*(\boldsymbol{w})\boldsymbol{x}_k\right)\nabla_{\boldsymbol{w}}\boldsymbol{q}(\boldsymbol{w})\boldsymbol{x}_k\right\|$$

$$\leq \left\|\frac{1}{p}\sum_{k=1}^p \dot{h}_\mu\left(\boldsymbol{q}^*(\boldsymbol{w})\boldsymbol{x}_k + \boldsymbol{q}^*(\boldsymbol{w})\widetilde{\boldsymbol{\Xi}}\boldsymbol{x}_k\right)\nabla_{\boldsymbol{w}}\boldsymbol{q}(\boldsymbol{w})\left(\boldsymbol{x}_k + \widetilde{\boldsymbol{\Xi}}\boldsymbol{x}_k - \boldsymbol{x}_k\right)\right\|$$

$$+ \left\|\frac{1}{p}\sum_{k=1}^p \left[\dot{h}_\mu\left(\boldsymbol{q}^*(\boldsymbol{w})\boldsymbol{x}_k + \boldsymbol{q}^*(\boldsymbol{w})\widetilde{\boldsymbol{\Xi}}\boldsymbol{x}_k\right) - \dot{h}_\mu\left(\boldsymbol{q}^*(\boldsymbol{w})\boldsymbol{x}_k\right)\right]\nabla_{\boldsymbol{w}}\boldsymbol{q}(\boldsymbol{w})\boldsymbol{x}_k\right\|$$

$$\leq \left\|\widetilde{\boldsymbol{\Xi}}\right\|\left(\max_t \dot{h}_\mu(t)\, 3n\|\boldsymbol{X}\|_\infty + L_{\dot{h}_\mu}3n\|\boldsymbol{X}\|_\infty^2\right),$$

where $L_{\dot{h}_\mu}$ denotes the Lipschitz constant for $\dot{h}_\mu(\cdot)$. Similarly, suppose $\left\|\widetilde{\boldsymbol{\Xi}}\right\| \leq \frac{1}{2n}$, and also notice that

$$\left\|\frac{\boldsymbol{I}}{q_n(\boldsymbol{w})} + \frac{\boldsymbol{w}\boldsymbol{w}^*}{q_n^3(\boldsymbol{w})}\right\| \leq \frac{1}{q_n(\boldsymbol{w})} + \frac{\|\boldsymbol{w}\|^2}{q_n^3(\boldsymbol{w})} = \frac{1}{q_n^3(\boldsymbol{w})} \leq 2\sqrt{2}n^{3/2},$$

we obtain that

$$\left\|\nabla_{\boldsymbol{w}}^2 g\left(\boldsymbol{w}; \widetilde{\boldsymbol{Y}}\right) - \nabla_{\boldsymbol{w}}^2 g\left(\boldsymbol{w}; \boldsymbol{X}\right)\right\|$$

$$\leq \left\|\frac{1}{p}\sum_{k=1}^p \left[\ddot{h}\left(\boldsymbol{q}^*(\boldsymbol{w})\widetilde{\boldsymbol{y}}_k\right)\nabla_{\boldsymbol{w}}\boldsymbol{q}(\boldsymbol{w})\widetilde{\boldsymbol{y}}_k\widetilde{\boldsymbol{y}}_k^*(\nabla_{\boldsymbol{w}}\boldsymbol{q}(\boldsymbol{w}))^* - \ddot{h}\left(\boldsymbol{q}^*(\boldsymbol{w})\boldsymbol{x}_k\right)\nabla_{\boldsymbol{w}}\boldsymbol{q}(\boldsymbol{w})\boldsymbol{x}_k\boldsymbol{x}_k^*(\nabla_{\boldsymbol{w}}\boldsymbol{q}(\boldsymbol{w}))^*\right]\right\|$$

$$+ \left\|\frac{1}{p}\sum_{k=1}^p \left[\dot{h}\left(\boldsymbol{q}^*(\boldsymbol{w})\widetilde{\boldsymbol{y}}_k\right)\left(\frac{\boldsymbol{I}}{q_n(\boldsymbol{w})} + \frac{\boldsymbol{w}\boldsymbol{w}^*}{q_n^3}\right)\widetilde{\boldsymbol{y}}_k(n) - \dot{h}\left(\boldsymbol{q}^*(\boldsymbol{w})\boldsymbol{x}_k\right)\left(\frac{\boldsymbol{I}}{q_n(\boldsymbol{w})} + \frac{\boldsymbol{w}\boldsymbol{w}^*}{q_n^3}\right)\boldsymbol{x}_k(n)\right]\right\|$$

$$\leq \frac{45}{2}L_{\ddot{h}_\mu}n^{3/2}\|\boldsymbol{X}\|_\infty^3\left\|\widetilde{\boldsymbol{\Xi}}\right\| + \max_t \ddot{h}_\mu(t)\left(18n^{3/2}\|\boldsymbol{X}\|_\infty^2\left\|\widetilde{\boldsymbol{\Xi}}\right\| + 10n^2\|\boldsymbol{X}\|_\infty^2\left\|\widetilde{\boldsymbol{\Xi}}\right\|^2\right)$$

$$+ 3\sqrt{2}L_{\dot{h}_\mu}n^2\left\|\widetilde{\boldsymbol{\Xi}}\right\|\|\boldsymbol{X}\|_\infty^2 + \max_t \dot{h}(t)\, 2\sqrt{2}n^2\left\|\widetilde{\boldsymbol{\Xi}}\right\|\|\boldsymbol{X}\|_\infty,$$

where $L_{\ddot{h}_\mu}$ denotes the Lipschitz constant for $\ddot{h}_\mu(\cdot)$. Since

$$\max_t \dot{h}_\mu(t) \leq 1, \quad \max_t \ddot{h}_\mu(t) \leq \frac{1}{\mu}, \quad L_{h_\mu} \leq 1, \quad L_{\dot{h}_\mu} \leq \frac{1}{\mu}, \quad L_{\ddot{h}_\mu} \leq \frac{2}{\mu^2},$$

70

and by Lemma 7.11, $\|\boldsymbol{X}\|_\infty \leq 4\sqrt{\log(np)}$ with probability at least $1 - \theta(np)^{-7} - \exp(-0.3\theta np)$, we obtain

$$\left\|\nabla_{\boldsymbol{w}} g\left(\boldsymbol{w}; \widetilde{\boldsymbol{Y}}\right) - \nabla_{\boldsymbol{w}} g\left(\boldsymbol{w}; \boldsymbol{X}\right)\right\| \leq C_1 \frac{n}{\mu} \log(np) \left\|\widetilde{\boldsymbol{\Xi}}\right\|,$$

$$\left\|\nabla_{\boldsymbol{w}}^2 g\left(\boldsymbol{w}; \widetilde{\boldsymbol{Y}}\right) - \nabla_{\boldsymbol{w}}^2 g\left(\boldsymbol{w}; \boldsymbol{X}\right)\right\| \leq C_2 \max\left\{\frac{n^{3/2}}{\mu^2}, \frac{n^2}{\mu}\right\} \log^{3/2}(np) \left\|\widetilde{\boldsymbol{\Xi}}\right\|$$

for numerical constants $C_1, C_2 > 0$. ∎

Proof [of Theorem 2.3] Assume the constant $c_\star$ as defined in Theorem 2.1. By Lemma 2.14, when

$$p \geq \frac{C_1}{c_\star^2 \theta} \max\left\{\frac{n^4}{\mu^4}, \frac{n^5}{\mu^2}\right\} \kappa^8(\boldsymbol{A}_0) \log^4\left(\frac{\kappa(\boldsymbol{A}_0) n}{\mu\theta}\right),$$

the magnitude of the perturbation is bounded as

$$\left\|\widetilde{\boldsymbol{\Xi}}\right\| \leq C_2 c_\star \theta \left(\max\left\{\frac{n^{3/2}}{\mu^2}, \frac{n^2}{\mu}\right\} \log^{3/2}(np)\right)^{-1},$$

where C_2 can be made arbitrarily small by making C_1 large. Combining this result with Lemma 2.15, we obtain that for all $\boldsymbol{w} \in \Gamma$,

$$\left\|\nabla_{\boldsymbol{w}} g\left(\boldsymbol{w}; \boldsymbol{X}_0 + \widetilde{\boldsymbol{\Xi}}\boldsymbol{X}_0\right) - \nabla_{\boldsymbol{w}} g\left(\boldsymbol{w}; \boldsymbol{X}\right)\right\| \leq \frac{c_\star \theta}{2}$$

$$\left\|\nabla_{\boldsymbol{w}}^2 g\left(\boldsymbol{w}; \boldsymbol{X}_0 + \widetilde{\boldsymbol{\Xi}}\boldsymbol{X}_0\right) - \nabla_{\boldsymbol{w}}^2 g\left(\boldsymbol{w}; \boldsymbol{X}\right)\right\| \leq \frac{c_\star \theta}{2},$$

with probability at least $1 - p^{-8} - \theta(np)^{-7} - \exp(-0.3\theta np)$. In view of (2.11) in Theorem 2.1, we have

$$\frac{\boldsymbol{w}^* g\left(\boldsymbol{w}; \boldsymbol{X}_0 + \widetilde{\boldsymbol{\Xi}}\boldsymbol{X}_0\right) \boldsymbol{w}}{\|\boldsymbol{w}\|^2} = \frac{\boldsymbol{w}^* g\left(\boldsymbol{w}; \boldsymbol{X}_0\right) \boldsymbol{w}}{\|\boldsymbol{w}\|^2} + \frac{\boldsymbol{w}^* g\left(\boldsymbol{w}; \boldsymbol{X}_0 + \widetilde{\boldsymbol{\Xi}}\boldsymbol{X}_0\right) \boldsymbol{w}}{\|\boldsymbol{w}\|^2} - \frac{\boldsymbol{w}^* g\left(\boldsymbol{w}; \boldsymbol{X}_0\right) \boldsymbol{w}}{\|\boldsymbol{w}\|^2}$$

$$\leq -c_\star \theta + \left\|\nabla_{\boldsymbol{w}}^2 g\left(\boldsymbol{w}; \boldsymbol{X}_0 + \widetilde{\boldsymbol{\Xi}}\boldsymbol{X}_0\right) - \nabla_{\boldsymbol{w}}^2 g\left(\boldsymbol{w}; \boldsymbol{X}\right)\right\| \leq -\frac{1}{2}c_\star \theta.$$

By similar arguments, we obtain (2.9) through (2.11) in Theorem 2.3.

To show the unique local minimizer over Γ is near $\boldsymbol{0}$, we note that (recall the last part of proof of Theorem 2.1 in Section 7.2) $g\left(\boldsymbol{w}; \boldsymbol{X}_0 + \widetilde{\boldsymbol{\Xi}}\boldsymbol{X}_0\right)$ being $\frac{c_\star \theta}{2\mu}$ strongly convex near $\boldsymbol{0}$ implies that

$$\|\boldsymbol{w}_\star\| \leq \frac{4\mu}{c_\star \theta} \left\|\nabla g\left(\boldsymbol{0}; \boldsymbol{X}_0 + \widetilde{\boldsymbol{\Xi}}\boldsymbol{X}_0\right)\right\|.$$

The above perturbation analysis implies there exists $C_3 > 0$ such that when

$$p \geq \frac{C_3}{c_\star^2 \theta} \max\left\{\frac{n^4}{\mu^4}, \frac{n^5}{\mu^2}\right\} \kappa^8(\boldsymbol{A}_0) \log^4\left(\frac{\kappa(\boldsymbol{A}_0) n}{\mu\theta}\right),$$

it holds that

$$\left\|\nabla_{\boldsymbol{w}}g\left(\boldsymbol{0};\boldsymbol{X}_0+\widetilde{\boldsymbol{\Xi}}\boldsymbol{X}_0\right)-\nabla_{\boldsymbol{w}}g\left(\boldsymbol{0};\boldsymbol{X}\right)\right\|\le\frac{c_\star\theta}{400},$$

which in turn implies

$$\|\boldsymbol{w}_\star\|\le\frac{4\mu}{c_\star\theta}\|\nabla g\left(\boldsymbol{0};\boldsymbol{X}_0\right)\|+\frac{4\mu}{c_\star\theta}\frac{c_\star\theta}{400}\le\frac{\mu}{8}+\frac{\mu}{100}<\frac{\mu}{7},$$

where we have recall the result that $\frac{2\mu}{c_\star\theta}\|\nabla g\left(\boldsymbol{0};\boldsymbol{X}_0\right)\|\le\mu/16$ from proof of Theorem 2.1. A simple union bound with careful bookkeeping gives the success probability. ∎

8 Proof of Convergence for the Trust-Region Algorithm

Proof [of Lemma 3.3] Using the fact $\tanh(\cdot)$ and $1-\tanh^2(\cdot)$ are bounded by one in magnitude, by (3.12) and (3.13) we have

$$\|\nabla f\left(\boldsymbol{q}\right)\|\le\frac{1}{p}\sum_{k=1}^p\|\boldsymbol{x}_k\|\le\sqrt{n}\,\|\boldsymbol{X}\|_\infty,$$

$$\|\nabla^2 f\left(\boldsymbol{q}\right)\|\le\frac{1}{p}\sum_{k=1}^p\frac{1}{\mu}\|\boldsymbol{x}_k\|^2\le\frac{n}{\mu}\|\boldsymbol{X}\|_\infty^2,$$

for any $\boldsymbol{q}\in\mathbb{S}^{n-1}$. Moreover,

$$\sup_{\boldsymbol{q},\boldsymbol{q}'\in\mathbb{S}^{n-1},\boldsymbol{q}\neq\boldsymbol{q}'}\frac{\|\nabla f\left(\boldsymbol{q}\right)-\nabla f\left(\boldsymbol{q}'\right)\|}{\|\boldsymbol{q}-\boldsymbol{q}'\|}\le\frac{1}{p}\sum_{k=1}^p\|\boldsymbol{x}_k\|\sup_{\boldsymbol{q},\boldsymbol{q}'\in\mathbb{S}^{n-1},\boldsymbol{q}\neq\boldsymbol{q}'}\frac{\left|\tanh\left(\frac{\boldsymbol{q}^*\boldsymbol{x}_k}{\mu}\right)-\tanh\left(\frac{\boldsymbol{q}'^*\boldsymbol{x}_k}{\mu}\right)\right|}{\|\boldsymbol{q}-\boldsymbol{q}'\|}$$

$$\le\frac{1}{p}\sum_{k=1}^p\|\boldsymbol{x}_k\|\frac{\|\boldsymbol{x}_k\|}{\mu}\le\frac{n}{\mu}\|\boldsymbol{X}\|_\infty^2,$$

where at the last line we have used the fact the mapping $\boldsymbol{q}\mapsto\boldsymbol{q}^*\boldsymbol{x}_k/\mu$ is $\|\boldsymbol{x}_k\|/\mu$ Lipschitz, and $x\mapsto\tanh(x)$ is 1-Lipschitz, and the composition rule in Lemma 7.5. Similar argument yields the final bound. ∎

Proof [of Lemma 3.4] Suppose we can establish

$$\left|f\left(\exp_{\boldsymbol{q}}(\boldsymbol{\delta})\right)-\widehat{f}\left(\boldsymbol{q},\boldsymbol{\delta}\right)\right|\le\frac{1}{6}\eta_f\|\boldsymbol{\delta}\|^3.$$

Applying this twice we obtain

$$f(\exp_{\boldsymbol{q}}(\boldsymbol{\delta}_\star))\le\widehat{f}(\boldsymbol{q},\boldsymbol{\delta}_\star)+\frac{1}{6}\eta_f\Delta^3\le\widehat{f}(\boldsymbol{q},\boldsymbol{\delta})+\frac{1}{6}\eta_f\Delta^3\le f(\exp_{\boldsymbol{q}}(\boldsymbol{\delta}))+\frac{1}{3}\eta_f\Delta^3\le f(\boldsymbol{q})-s+\frac{1}{3}\eta_f\Delta^3,$$

as claimed. Next we establish the first result. Let $\boldsymbol{\delta}_0=\frac{\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|}$, and $t=\|\boldsymbol{\delta}\|$. Consider the composite function

$$\zeta(t)\doteq f(\exp_{\boldsymbol{q}}(t\boldsymbol{\delta}_0))=f(\boldsymbol{q}\cos(t)+\boldsymbol{\delta}_0\sin(t)),$$

and also

$$\dot{\zeta}(t) = \langle \nabla f\left(\boldsymbol{q}\cos(t) + \boldsymbol{\delta}_0 \sin(t)\right), -\boldsymbol{q}\sin(t) + \boldsymbol{\delta}_0 \cos(t)\rangle$$

$$\ddot{\zeta}(t) = \langle \nabla^2 f\left(\boldsymbol{q}\cos(t) + \boldsymbol{\delta}_0 \sin(t)\right)\left(-\boldsymbol{q}\sin(t) + \boldsymbol{\delta}_0 \cos(t)\right), -\boldsymbol{q}\sin(t) + \boldsymbol{\delta}_0 \cos(t)\rangle$$
$$+ \quad \langle \nabla f\left(\boldsymbol{q}\cos(t) + \boldsymbol{\delta}_0 \sin(t)\right), -\boldsymbol{q}\cos(t) - \boldsymbol{\delta}_0 \sin(t)\rangle .$$

In particular, this gives that

$$\zeta(0) = f(\boldsymbol{q})$$
$$\dot{\zeta}(0) = \langle \boldsymbol{\delta}_0, \nabla f(\boldsymbol{q})\rangle$$
$$\ddot{\zeta}(0) = \boldsymbol{\delta}_0^* \left(\nabla^2 f(\boldsymbol{q}) - \langle \nabla f(\boldsymbol{q}), \boldsymbol{q}\rangle \boldsymbol{I}\right) \boldsymbol{\delta}_0.$$

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

$$\left|\ddot{\zeta}(t) - \ddot{\zeta}(0)\right|$$
$$\leq \left|\langle \nabla^2 f\left(\boldsymbol{q}\cos(t) + \boldsymbol{\delta}_0 \sin(t)\right)\left(-\boldsymbol{q}\sin(t) + \boldsymbol{\delta}_0 \cos(t)\right), -\boldsymbol{q}\sin(t) + \boldsymbol{\delta}_0 \cos(t)\rangle - \boldsymbol{\delta}_0^* \nabla^2 f(\boldsymbol{q})\boldsymbol{\delta}_0\right|$$
$$+ \quad \left|\langle \nabla f\left(\boldsymbol{q}\cos(t) + \boldsymbol{\delta}_0 \sin(t)\right), -\boldsymbol{q}\cos(t) - \boldsymbol{\delta}_0 \sin(t)\rangle + \langle \nabla f(\boldsymbol{q}), \boldsymbol{q}\rangle\right|$$
$$\leq \left|\langle \left[\nabla^2 f(\boldsymbol{q}\cos(t) + \boldsymbol{\delta}_0 \sin(t)) - \nabla^2 f(\boldsymbol{q})\right]\left(-\boldsymbol{q}\sin(t) + \boldsymbol{\delta}_0 \cos(t)\right), -\boldsymbol{q}\sin(t) + \boldsymbol{\delta}_0 \cos(t)\rangle\right|$$
$$+ \quad \left|\langle \nabla^2 f(\boldsymbol{q})\left(-\boldsymbol{q}\sin(t) + \boldsymbol{\delta}_0 \cos(t) - \boldsymbol{\delta}_0\right), -\boldsymbol{q}\sin(t) + \boldsymbol{\delta}_0 \cos(t)\rangle\right|$$
$$+ \quad \left|\langle \nabla^2 f(\boldsymbol{q})\boldsymbol{\delta}_0, -\boldsymbol{q}\sin(t) + \boldsymbol{\delta}_0 \cos(t) - \boldsymbol{\delta}_0\rangle\right|$$
$$+ \quad \left|\langle \nabla f(\boldsymbol{q}\cos(t) + \boldsymbol{\delta}_0 \sin(t)), -\boldsymbol{q}\cos(t) - \boldsymbol{\delta}_0 \sin(t)\rangle + \langle \nabla f(\boldsymbol{q}\cos(t) + \boldsymbol{\delta}_0 \sin(t)), \boldsymbol{q}\rangle\right|$$
$$+ \quad \left|\langle \nabla f(\boldsymbol{q}\cos(t) + \boldsymbol{\delta}_0 \sin(t)), \boldsymbol{q}\rangle - \langle \nabla f(\boldsymbol{q}), \boldsymbol{q}\rangle\right|$$
$$\leq L_{\nabla^2} \left\|\boldsymbol{q}\cos(t) + \boldsymbol{\delta}_0 \sin(t) - \boldsymbol{q}\right\|$$
$$+ M_{\nabla^2} \left\|-\boldsymbol{q}\sin(t) + \boldsymbol{\delta}_0 \cos(t) - \boldsymbol{\delta}_0\right\|$$
$$+ M_{\nabla^2} \left\|-\boldsymbol{q}\sin(t) + \boldsymbol{\delta}_0 \cos(t) - \boldsymbol{\delta}_0\right\|$$
$$+ M_{\nabla} \left\|-\boldsymbol{q}\cos(t) - \boldsymbol{\delta}_0 \sin(t) + \boldsymbol{q}\right\|$$
$$+ L_{\nabla} \left\|\boldsymbol{q}\cos(t) + \boldsymbol{\delta}_0 \sin(t) - \boldsymbol{q}\right\|$$
$$= (L_{\nabla^2} + 2M_{\nabla^2} + M_{\nabla} + L_{\nabla}) \sqrt{(1 - \cos(t))^2 + \sin^2(t)}$$
$$= \eta_f \sqrt{2 - 2\cos t} \leq \eta_f \sqrt{4\sin^2\left(t/2\right)} \leq \eta_f t,$$

where in the final line we have used the fact $1 - \cos x = 2\sin^2\left(x/2\right)$ and that $\sin x \leq x$ for $x \in [0, 1]$, and M_∇, M_{∇^2}, L_∇ and L_{∇^2} are the quantities defined in Lemma 3.3. By the integral form of Taylor's theorem in Lemma A.12 and the result above, we have

$$\left|f\left(\exp_{\boldsymbol{q}}(\boldsymbol{\delta})\right) - \widehat{f}\left(\boldsymbol{q}, \boldsymbol{\delta}\right)\right| = \left|\zeta(t) - \left(\zeta(0) + t\dot{\zeta}(0) + \frac{t^2}{2}\ddot{\zeta}(0)\right)\right|$$
$$= \left|t^2 \int_0^1 (1 - s)\ddot{\zeta}\left(st\right) \, ds - \frac{t^2}{2}\ddot{\zeta}(0)\right|$$
$$= t^2 \left|\int_0^1 (1 - s)\left[\ddot{\zeta}\left(st\right) - \ddot{\zeta}\left(0\right)\right] \, ds\right|$$

73

$$\leq t^2 \int_0^1 (1-s) \, st\eta_f \, ds = \frac{\eta_f t^3}{6},$$

with $t = \|\boldsymbol{\delta}\|$ we obtain the desired result. ∎

Proof [of Lemma 3.5] By the integral form of Taylor's theorem in Lemma A.12, for any $t \in \left[0, \frac{3\Delta}{2\pi\sqrt{n}}\right]$, we have

$$g\left(\boldsymbol{w} - t\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right)$$

$$= g(\boldsymbol{w}) - t\int_0^1 \left\langle \nabla g\left(\boldsymbol{w} - st\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right), \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \right\rangle ds$$

$$= g(\boldsymbol{w}) - t\frac{\boldsymbol{w}^*\nabla g(\boldsymbol{w})}{\|\boldsymbol{w}\|} + t\int_0^1 \left\langle \nabla g(\boldsymbol{w}) - \nabla g\left(\boldsymbol{w} - st\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right), \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \right\rangle ds$$

$$= g(\boldsymbol{w}) - t\frac{\boldsymbol{w}^*\nabla g(\boldsymbol{w})}{\|\boldsymbol{w}\|} + t\int_0^1 \left(\left\langle \nabla g(\boldsymbol{w}), \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \right\rangle - \left\langle \nabla g\left(\boldsymbol{w} - st\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right), \frac{\boldsymbol{w} - st\boldsymbol{w}/\|\boldsymbol{w}\|}{\|\boldsymbol{w} - st\boldsymbol{w}/\|\boldsymbol{w}\|\|} \right\rangle\right) ds$$

$$\leq g(\boldsymbol{w}) - t\frac{\boldsymbol{w}^*\nabla g(\boldsymbol{w})}{\|\boldsymbol{w}\|} + \frac{L_g}{2}t^2 \leq g(\boldsymbol{w}) - t\beta_g + \frac{L_g}{2}t^2.$$

Minimizing this function over $t \in \left[0, \frac{3\Delta}{2\pi\sqrt{n}}\right]$, we obtain that there exists a $\boldsymbol{w}' \in \mathcal{B}\left(\boldsymbol{w}, \frac{3\Delta}{2\pi\sqrt{n}}\right)$ such that

$$g(\boldsymbol{w}') \leq g(\boldsymbol{w}) - \min\left\{\frac{\beta_g^2}{2L_g}, \frac{3\beta_g\Delta}{4\pi\sqrt{n}}\right\}.$$

Given such a $\boldsymbol{w}' \in \mathcal{B}\left(\boldsymbol{w}, \frac{3\Delta}{2\pi\sqrt{n}}\right)$, there must exist some $\boldsymbol{\delta} \in T_{\boldsymbol{q}}\mathbb{S}^{n-1}$ such that $\boldsymbol{q}(\boldsymbol{w}') = \exp_{\boldsymbol{q}}(\boldsymbol{\delta})$. It remains to show that $\|\boldsymbol{\delta}\| \leq \Delta$. By Lemma 7.7, we know that $\|\boldsymbol{q}(\boldsymbol{w}') - \boldsymbol{q}(\boldsymbol{w})\| \leq 2\sqrt{n}\|\boldsymbol{w}' - \boldsymbol{w}\| \leq 3\Delta/\pi$. Hence,

$$\left\|\exp_{\boldsymbol{q}}(\boldsymbol{\delta}) - \boldsymbol{q}\right\|^2 = \left\|\boldsymbol{q}(1 - \cos\|\boldsymbol{\delta}\|) + \frac{\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|}\sin\|\boldsymbol{\delta}\|\right\|^2 = 2 - 2\cos\|\boldsymbol{\delta}\| = 4\sin^2\frac{\|\boldsymbol{\delta}\|}{2} \leq \frac{9\Delta^2}{\pi^2},$$

which means that $\sin(\|\boldsymbol{\delta}\|/2) \leq 3\Delta/(2\pi)$. Because $\sin x \geq \frac{3}{\pi}x$ over $x \in [0, \pi/6]$, it implies that $\|\boldsymbol{\delta}\| \leq \Delta$. Since $g(\boldsymbol{w}) = f(\boldsymbol{q}(\boldsymbol{w}))$, by summarizing all the results, we conclude that there exists a $\boldsymbol{\delta}$ with $\|\boldsymbol{\delta}\| \leq \Delta$, such that

$$f(\exp_{\boldsymbol{q}}(\boldsymbol{\delta})) \leq f(\boldsymbol{q}) - \min\left\{\frac{\beta_g^2}{2L_g}, \frac{3\beta_g\Delta}{4\pi\sqrt{n}}\right\},$$

as claimed. ∎

Proof [of Lemma 3.6] Let $\sigma = \text{sign}(\boldsymbol{w}^*\nabla g(\boldsymbol{w}))$. For any $t \in \left[0, \frac{\Delta}{2\sqrt{n}}\right]$, by integral form of Taylor's theorem in Lemma A.12, we have

$$g\left(\boldsymbol{w} - t\sigma\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right)$$

$$= g(\boldsymbol{w}) - t\sigma \frac{\boldsymbol{w}^* \nabla g(\boldsymbol{w})}{\|\boldsymbol{w}\|} + t^2 \int_0^1 (1-s) \frac{\boldsymbol{w}^* \nabla^2 g\left(\boldsymbol{w} - st\sigma \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right) \boldsymbol{w}}{\|\boldsymbol{w}\|^2} \, ds$$

$$\leq g(\boldsymbol{w}) + \frac{t^2}{2} \frac{\boldsymbol{w}^* \nabla^2 g(\boldsymbol{w}) \boldsymbol{w}}{\|\boldsymbol{w}\|^2} + t^2 \int_0^1 \left[(1-s) \frac{\boldsymbol{w}^* \nabla^2 g\left(\boldsymbol{w} - st\sigma \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right) \boldsymbol{w}}{\|\boldsymbol{w}\|^2} - (1-s) \frac{\boldsymbol{w}^* \nabla^2 g(\boldsymbol{w}) \boldsymbol{w}}{\|\boldsymbol{w}\|^2}\right] ds$$

$$= g(\boldsymbol{w}) + \frac{t^2}{2} \frac{\boldsymbol{w}^* \nabla^2 g(\boldsymbol{w}) \boldsymbol{w}}{\|\boldsymbol{w}\|^2}$$

$$+ t^2 \int_0^1 (1-s) \left[\frac{\left(\boldsymbol{w} - st\sigma \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right)^* \nabla^2 g\left(\boldsymbol{w} - st\sigma \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right)\left(\boldsymbol{w} - st\sigma \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right)}{\left\|\boldsymbol{w} - st\sigma \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right\|^2} - \frac{\boldsymbol{w}^* \nabla^2 g(\boldsymbol{w}) \boldsymbol{w}}{\|\boldsymbol{w}\|^2}\right] ds$$

$$\leq g(\boldsymbol{w}) - \frac{t^2}{2}\beta_\cap + t^2 \int_0^1 (1-s)\, sL_\cap t \, ds \;\leq\; g(\boldsymbol{w}) - \frac{t^2}{2}\beta_\cap + \frac{t^3}{6}L_\cap.$$

Minimizing this function over $t \in \left[0, \frac{3\Delta}{2\pi\sqrt{n}}\right]$, we obtain

$$t_\star = \min\left\{\frac{2\beta_\cap}{L_\cap}, \frac{3\Delta}{2\pi\sqrt{n}}\right\},$$

and there exists a $\boldsymbol{w}' = \boldsymbol{w} - t_\star \sigma \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$ such that

$$g\left(\boldsymbol{w} - t_\star \sigma \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right) \;\leq\; g(\boldsymbol{w}) - \min\left\{\frac{2\beta_\cap^3}{3L_\cap^2}, \frac{3\Delta^2 \beta_\cap}{8\pi^2 n}\right\}.$$

By arguments identical to those used in Lemma 3.5, there exists a tangent vector $\boldsymbol{\delta} \in T_{\boldsymbol{q}}\mathbb{S}^{n-1}$ such that $\boldsymbol{q}(\boldsymbol{w}') = \exp_{\boldsymbol{q}}(\boldsymbol{\delta})$ and $\|\boldsymbol{\delta}\| \leq \Delta$. This completes the proof. ∎

Proof [of Lemma 3.8] For any $t \in \left[0, \frac{\Delta}{\|\operatorname{grad} f(\boldsymbol{q}^{(k)})\|}\right]$, it holds that $\left\|t \ \operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right)\right\| \leq \Delta$, and the quadratic approximation

$$\widehat{f}\left(\boldsymbol{q}^{(k)}, -t \ \operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right)\right) \leq f\left(\boldsymbol{q}^{(k)}\right) - t \left\|\operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right)\right\|^2 + \frac{M_H}{2}t^2 \left\|\operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right)\right\|^2$$

$$= f\left(\boldsymbol{q}^{(k)}\right) - t\left(1 - \frac{1}{2}M_H t\right) \left\|\operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right)\right\|^2.$$

Taking $t_0 = \min\left\{\frac{\Delta}{\|\operatorname{grad} f(\boldsymbol{q}^{(k)})\|}, \frac{1}{M_H}\right\}$, we obtain

$$\widehat{f}\left(\boldsymbol{q}^{(k)}, -t_0 \ \operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right)\right) \leq f\left(\boldsymbol{q}^{(k)}\right) - \frac{1}{2}\min\left\{\frac{\Delta}{\|\operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right)\|}, \frac{1}{M_H}\right\} \left\|\operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right)\right\|^2. \quad (8.1)$$

Now let $\boldsymbol{U}$ be an arbitrary orthonormal basis for $T_{\boldsymbol{q}^{(k)}}\mathbb{S}^{n-1}$. Since the norm constraint is active, by the optimality condition in (3.16), we have

$$\Delta \leq \left\|\left[\boldsymbol{U}^* \operatorname{Hess} f\left(\boldsymbol{q}^{(k)}\right) \boldsymbol{U}\right]^{-1} \boldsymbol{U}^* \operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right)\right\|$$

$$\leq \left\| \left[\boldsymbol{U}^* \operatorname{Hess} f\left(\boldsymbol{q}^{(k)}\right) \boldsymbol{U}\right]^{-1} \right\| \left\| \boldsymbol{U}^* \operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right) \right\| \leq \frac{\left\| \operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right)\right\|}{m_H},$$

which means that $\left\| \operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right)\right\| \geq m_H \Delta$. Substituting this into (8.1), we obtain

$$\widehat{f}\left(\boldsymbol{q}^{(k)}, -t_0 \operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right)\right) \leq f\left(\boldsymbol{q}^{(k)}\right) - \frac{1}{2}\min\left\{ m_H \Delta^2, \frac{m_H^2}{M_H}\Delta^2 \right\} \leq f\left(\boldsymbol{q}^{(k)}\right) - \frac{m_H^2}{2M_H}.$$

By the key comparison result established in proof of Lemma 3.4, we have

$$f\left(\exp_{\boldsymbol{q}^{(k)}}\left(-t_0 \operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right)\right)\right) \leq \widehat{f}\left(\boldsymbol{q}^{(k)}, -t_0 \operatorname{grad} f\left(\boldsymbol{q}^{(k)}\right)\right) + \frac{1}{6}\eta_f \Delta^3$$

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

This completes the proof. ∎

It takes certain delicate work to prove Lemma 3.9. 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 8.1 *For $\tau \in [0, 1]$ and $\|\boldsymbol{\delta}\| \leq 1/2$, we have*

$$\left\| \mathcal{P}_\gamma^{\tau \leftarrow 0} - \boldsymbol{I}\right\| \leq \frac{5}{4}\tau \|\boldsymbol{\delta}\|, \tag{8.2}$$

$$\left\| \mathcal{P}_\gamma^{0 \leftarrow \tau} - \boldsymbol{I}\right\| \leq \frac{3}{2}\tau \|\boldsymbol{\delta}\|. \tag{8.3}$$

Proof By (3.17), we have

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

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

$$\leq 2\sin^2\left(\frac{\tau \|\boldsymbol{\delta}\|}{2}\right) + \sin(\tau \|\boldsymbol{\delta}\|) \leq \frac{1}{4}\tau \|\boldsymbol{\delta}\| + \tau \|\boldsymbol{\delta}\| \leq \frac{5}{4}\tau \|\boldsymbol{\delta}\|,$$

where we have used the fact $\sin(t) \leq t$ and $1 - \cos x = 2\sin^2(x/2)$. Moreover, $\mathcal{P}_\gamma^{0 \leftarrow \tau}$ is in the form of $(\boldsymbol{I} + \boldsymbol{u}\boldsymbol{v}^*)^{-1}$ for some vectors $\boldsymbol{u}$ and $\boldsymbol{v}$. By the Sherman-Morrison matrix inverse formula, i.e., $(\boldsymbol{I} + \boldsymbol{u}\boldsymbol{v}^*)^{-1} = \boldsymbol{I} - \boldsymbol{u}\boldsymbol{v}^*/(1 + \boldsymbol{v}^*\boldsymbol{u})$ (justified as $\left\| (\cos(\tau \|\boldsymbol{\delta}\|) - 1)\frac{\boldsymbol{\delta}\boldsymbol{\delta}^*}{\|\boldsymbol{\delta}\|^2} - \boldsymbol{q}\sin(\tau \|\boldsymbol{\delta}\|)\frac{\boldsymbol{\delta}^*}{\|\boldsymbol{\delta}\|}\right\| \leq 5\tau \|\boldsymbol{\delta}\|/4 \leq 5/8 < 1$ as shown above), we have

$$\left\| \mathcal{P}_\gamma^{0 \leftarrow \tau} - \boldsymbol{I}\right\|$$

$$= \left\| (\cos(\tau \|\boldsymbol{\delta}\|) - 1)\frac{\boldsymbol{\delta}\boldsymbol{\delta}^*}{\|\boldsymbol{\delta}\|^2} - \boldsymbol{q}\sin(\tau \|\boldsymbol{\delta}\|)\frac{\boldsymbol{\delta}^*}{\|\boldsymbol{\delta}\|}\right\| \frac{1}{1 + (\cos(\tau \|\boldsymbol{\delta}\|) - 1)} \quad (\text{as } \boldsymbol{q}^*\boldsymbol{\delta} = 0)$$

$$\leq \frac{5}{4}\tau \left\|\boldsymbol{\delta}\right\| \frac{1}{\cos\left(\tau\left\|\boldsymbol{\delta}\right\|\right)} \leq \frac{5}{4}\tau \left\|\boldsymbol{\delta}\right\| \frac{1}{\cos\left(1/2\right)} \leq \frac{3}{2}\tau \left\|\boldsymbol{\delta}\right\|,$$

completing the proof. ∎

The next lemma establish the "local-Lipschitz" property of the Riemannian Hessian.

Lemma 8.2 *Let $\gamma(t) = \exp_{\boldsymbol{q}}(t\boldsymbol{\delta})$ denotes a geodesic curve on $\mathbb{S}^{n-1}$. Whenever $\left\|\boldsymbol{\delta}\right\| \leq 1/2$ and $\tau \in [0,1]$,*

$$\left\|\mathcal{P}_\gamma^{0\leftarrow\tau}\operatorname{Hess}f(\gamma(\tau))\mathcal{P}_\gamma^{\tau\leftarrow 0} - \operatorname{Hess}f(\boldsymbol{q})\right\| \leq L_H \cdot \tau \left\|\boldsymbol{\delta}\right\|, \tag{8.4}$$

where $L_H = \frac{5}{2\mu^2}n^{3/2}\left\|\boldsymbol{X}\right\|_\infty^3 + \frac{9}{\mu}n\left\|\boldsymbol{X}\right\|_\infty^2 + 9\sqrt{n}\left\|\boldsymbol{X}\right\|_\infty$.

Proof First of all, by (3.15) and using the fact that the operator norm of a projection operator is unitary bounded, we have

$$\left\|\operatorname{Hess}f(\gamma(\tau)) - \operatorname{Hess}f(\boldsymbol{q})\right\|$$
$$\leq \left\|\mathcal{P}_{T_{\gamma(\tau)}\mathbb{S}^{n-1}}\left[\nabla^2 f\left(\gamma\left(\tau\right)\right) - \nabla^2 f\left(\boldsymbol{q}\right) - \left(\left\langle\nabla f\left(\gamma\left(\tau\right)\right),\gamma\left(\tau\right)\right\rangle - \left\langle\nabla f\left(\boldsymbol{q}\right),\boldsymbol{q}\right\rangle\right)\boldsymbol{I}\right]\mathcal{P}_{T_{\gamma(\tau)}\mathbb{S}^{n-1}}\right\|$$
$$+ \left\|\mathcal{P}_{T_{\gamma(\tau)}\mathbb{S}^{n-1}}\left(\nabla^2 f\left(\boldsymbol{q}\right) - \left\langle\nabla f\left(\boldsymbol{q}\right),\boldsymbol{q}\right\rangle\boldsymbol{I}\right)\mathcal{P}_{T_{\gamma(\tau)}\mathbb{S}^{n-1}}\right.$$
$$\left. - \mathcal{P}_{T_{\boldsymbol{q}}\mathbb{S}^{n-1}}\left(\nabla^2 f\left(\boldsymbol{q}\right) - \left\langle\nabla f\left(\boldsymbol{q}\right),\boldsymbol{q}\right\rangle\boldsymbol{I}\right)\mathcal{P}_{T_{\boldsymbol{q}}\mathbb{S}^{n-1}}\right\|$$
$$\leq \left\|\nabla^2 f\left(\gamma\left(\tau\right)\right) - \nabla^2 f\left(\boldsymbol{q}\right)\right\| + \left|\left\langle\nabla f\left(\gamma\left(\tau\right)\right) - \nabla f\left(\boldsymbol{q}\right),\gamma\left(\tau\right)\right\rangle\right| + \left|\left\langle\nabla f\left(\boldsymbol{q}\right),\gamma\left(\tau\right) - \boldsymbol{q}\right\rangle\right|$$
$$+ \left\|\mathcal{P}_{T_{\gamma(\tau)}\mathbb{S}^{n-1}} - \mathcal{P}_{T_{\boldsymbol{q}}\mathbb{S}^{n-1}}\right\|\left\|\mathcal{P}_{T_{\gamma(\tau)}\mathbb{S}^{n-1}} + \mathcal{P}_{T_{\boldsymbol{q}}\mathbb{S}^{n-1}}\right\|\left\|\nabla^2 f\left(\boldsymbol{q}\right) - \left\langle\nabla f\left(\boldsymbol{q}\right),\boldsymbol{q}\right\rangle\boldsymbol{I}\right\|.$$

By the estimates in Lemma 3.3, we obtain

$$\left\|\operatorname{Hess}f(\gamma(\tau)) - \operatorname{Hess}f(\boldsymbol{q})\right\|$$
$$\leq \frac{2}{\mu^2}n^{3/2}\left\|\boldsymbol{X}\right\|_\infty^3\left\|\gamma\left(\tau\right) - \boldsymbol{q}\right\| + \frac{n}{\mu}\left\|\boldsymbol{X}\right\|_\infty^2\left\|\gamma\left(\tau\right) - \boldsymbol{q}\right\| + \sqrt{n}\left\|\boldsymbol{X}\right\|_\infty\left\|\gamma\left(\tau\right) - \boldsymbol{q}\right\|$$
$$+ 2\left\|\gamma\left(\tau\right)\gamma^*\left(\tau\right) - \boldsymbol{q}\boldsymbol{q}^*\right\|\left(\frac{n}{\mu}\left\|\boldsymbol{X}\right\|_\infty^2 + \sqrt{n}\left\|\boldsymbol{X}\right\|_\infty\right)$$
$$\leq \left(\frac{5}{2\mu^2}n^{3/2}\left\|\boldsymbol{X}\right\|_\infty^3 + \frac{25n}{4\mu}\left\|\boldsymbol{X}\right\|_\infty^2 + \frac{25}{4}\sqrt{n}\left\|\boldsymbol{X}\right\|_\infty\right)\tau\left\|\boldsymbol{\delta}\right\|, \tag{8.5}$$

where at the last line we have used the following estimates:

$$\left\|\gamma\left(\tau\right) - \boldsymbol{q}\right\| = \left\|\boldsymbol{q}\left(\cos\left(\tau\left\|\boldsymbol{\delta}\right\|\right) - 1\right) + \frac{\boldsymbol{\delta}}{\left\|\boldsymbol{\delta}\right\|}\sin\left(\tau\left\|\boldsymbol{\delta}\right\|\right)\right\| \leq \frac{5}{4}\tau\left\|\boldsymbol{\delta}\right\|, \text{ (Proof of Lemma 8.1)}$$

$$\left\|\gamma\left(\tau\right)\gamma^*\left(\tau\right) - \boldsymbol{q}\boldsymbol{q}^*\right\| \leq \left\|\left(\frac{\boldsymbol{\delta}\boldsymbol{\delta}^*}{\left\|\boldsymbol{\delta}\right\|^2} - \boldsymbol{q}\boldsymbol{q}^*\right)\sin^2\left(\tau\left\|\boldsymbol{\delta}\right\|\right)\right\| + 2\sin\left(\tau\left\|\boldsymbol{\delta}\right\|\right)\cos\left(\tau\left\|\boldsymbol{\delta}\right\|\right)$$

$$\leq \sin^2\left(\tau\left\|\boldsymbol{\delta}\right\|\right) + \sin\left(2\tau\left\|\boldsymbol{\delta}\right\|\right) \leq \frac{5}{2}\tau\left\|\boldsymbol{\delta}\right\|.$$

Therefore, by Lemma 8.1, we obtain

$$\left\|\mathcal{P}_\gamma^{0\leftarrow\tau}\operatorname{Hess}f(\gamma(\tau))\mathcal{P}_\gamma^{\tau\leftarrow 0} - \operatorname{Hess}f(\boldsymbol{q})\right\|$$

$$\leq \left\| \mathcal{P}_\gamma^{0\leftarrow\tau} \operatorname{Hess} f(\gamma(\tau))\mathcal{P}_\gamma^{\tau\leftarrow 0} - \operatorname{Hess} f(\gamma(\tau))\mathcal{P}_\gamma^{\tau\leftarrow 0} \right\| + \left\| \operatorname{Hess} f(\gamma(\tau))\mathcal{P}_\gamma^{\tau\leftarrow 0} - \operatorname{Hess} f(\gamma(\tau)) \right\|$$
$$+ \left\| \operatorname{Hess} f(\gamma(\tau)) - \operatorname{Hess} f(\boldsymbol{q}) \right\|$$
$$\leq \left\| \mathcal{P}_\gamma^{0\leftarrow\tau} - \boldsymbol{I} \right\| \left\| \operatorname{Hess} f(\gamma(\tau)) \right\| + \left\| \mathcal{P}_\gamma^{\tau\leftarrow 0} - \boldsymbol{I} \right\| \left\| \operatorname{Hess} f(\gamma(t)) \right\| + \left\| \operatorname{Hess} f(\gamma(t)) - \operatorname{Hess} f(\boldsymbol{q}) \right\|$$
$$\leq \frac{11}{4}\tau \left\| \boldsymbol{\delta} \right\| \left\| \nabla^2 f\left(\gamma\left(\tau\right)\right) - \left\langle \nabla f\left(\gamma\left(\tau\right)\right), \gamma\left(t\right) \right\rangle \boldsymbol{I} \right\| + \left\| \operatorname{Hess} f(\gamma(\tau)) - \operatorname{Hess} f(\boldsymbol{q}) \right\|.$$

By Lemma 3.3 and substituting the estimate in (8.5), we obtain the claimed result. ∎

Proof [of Lemma 3.9] For any given $\boldsymbol{q}$ with $\|\boldsymbol{w}(\boldsymbol{q})\| \leq \mu/(4\sqrt{2})$, assume $\boldsymbol{U}$ is an orthonormal basis for its tangent space $T_{\boldsymbol{q}}\mathbb{S}^{n-1}$. We could compare $\boldsymbol{U}^* \operatorname{Hess} f(\boldsymbol{q})\boldsymbol{U}$ with $\nabla^2_{\boldsymbol{w}} g(\boldsymbol{w})$, and build on the known results for the latter. Instead, we present a direct proof here that yields tighter results as stated in the lemma. Again we first work with the "canonical" section in the vicinity of $\boldsymbol{e}_n$ with the "canonical" reparametrization $\boldsymbol{q}(\boldsymbol{w}) = [\boldsymbol{w}; \sqrt{1 - \|\boldsymbol{w}\|^2}]$.

By definition of the Riemannian Hessian in (3.15), expressions of $\nabla^2 f$ and ∇f in (3.12) and (3.13), and exchange of differential and expectation opeators (justified similarly as in Section 7.1.3), we obtain

$$\boldsymbol{U}^* \operatorname{Hess} \mathbb{E}\left[f(\boldsymbol{q}) \right] \boldsymbol{U} = \mathbb{E}\left[\boldsymbol{U}^* \operatorname{Hess} f(\boldsymbol{q})\boldsymbol{U} \right]$$
$$= \mathbb{E}\left[\boldsymbol{U}^* \nabla^2 f(\boldsymbol{q})\boldsymbol{U} - \left\langle \boldsymbol{q}, \nabla f(\boldsymbol{q}) \right\rangle \boldsymbol{I} \right]$$
$$= \mathbb{E}\left[\frac{1}{\mu}\left\{ 1 - \tanh^2\left(\frac{\boldsymbol{q}^* \boldsymbol{x}_k}{\mu} \right) \right\} \boldsymbol{U}^* \boldsymbol{x}_k \boldsymbol{x}_k^* \boldsymbol{U} \right] - \mathbb{E}\left[\tanh\left(\frac{\boldsymbol{q}^* \boldsymbol{x}_k}{\mu} \right) \boldsymbol{q}^* \boldsymbol{x}_k \right] \boldsymbol{I}.$$

Now because $\boldsymbol{U}^* \boldsymbol{q} = \boldsymbol{0}$, we have

$$\mathbb{E}\left[\frac{1}{\mu}\left\{ 1 - \tanh^2\left(\frac{\boldsymbol{q}^* \boldsymbol{x}_k}{\mu} \right) \right\} \boldsymbol{U}^* \boldsymbol{x}_k \boldsymbol{x}_k^* \boldsymbol{U} \right] = \frac{1}{\mu}\mathbb{E}\left[1 - \tanh^2\left(\frac{\boldsymbol{q}^* \boldsymbol{x}_k}{\mu} \right) \right] \mathbb{E}\left[\boldsymbol{U}^* \boldsymbol{x}_k \boldsymbol{x}_k^* \boldsymbol{U} \right]$$
$$= \frac{\theta}{\mu}\mathbb{E}\left[1 - \tanh^2\left(\frac{\boldsymbol{q}^* \boldsymbol{x}_k}{\mu} \right) \right] \boldsymbol{I}.$$

So it is enough to bound the scalar part:

$$\frac{\theta}{\mu}\mathbb{E}\left[1 - \tanh^2\left(\frac{\boldsymbol{q}^* \boldsymbol{x}_k}{\mu} \right) \right] - \mathbb{E}\left[\tanh\left(\frac{\boldsymbol{q}^* \boldsymbol{x}_k}{\mu} \right) \boldsymbol{q}^* \boldsymbol{x}_k \right]$$
$$= \mathbb{E}_{\mathcal{I}}\left[\frac{\theta}{\mu}\mathbb{E}_{Z\sim\mathcal{N}\left(0,\|\boldsymbol{q}_\mathcal{I}\|^2\right)}\left[1 - \tanh^2(Z/\mu) \right] - \mathbb{E}_{Z\sim\mathcal{N}\left(0,\|\boldsymbol{q}_\mathcal{I}\|^2\right)}\left[\tanh(Z/\mu)Z \right] \right]$$
$$= \mathbb{E}_{\mathcal{I}}\left[\left(\frac{\theta}{\mu} - \frac{\|\boldsymbol{q}_\mathcal{I}\|^2}{\mu} \right) \mathbb{E}_{Z\sim\mathcal{N}\left(0,\|\boldsymbol{q}_\mathcal{I}\|^2\right)}\left[1 - \tanh^2(Z/\mu) \right] \right] \quad \text{(by (B.6))}$$
$$= (1-\theta)\mathbb{E}_{\mathcal{J}}\left[\left(\frac{\theta}{\mu} - \frac{\|\boldsymbol{w}_\mathcal{J}\|^2}{\mu} \right) \mathbb{E}_{Z\sim\mathcal{N}\left(0,\|\boldsymbol{w}_\mathcal{J}\|^2\right)}\left[1 - \tanh^2(Z/\mu) \right] \right]$$
$$+ \theta\mathbb{E}_{\mathcal{J}}\left[\left(\frac{\theta}{\mu} - \frac{\|\boldsymbol{w}_\mathcal{J}\|^2 + q_n^2}{\mu} \right) \mathbb{E}_{Z\sim\mathcal{N}\left(0,\|\boldsymbol{w}_\mathcal{J}\|^2 + q_n^2\right)}\left[1 - \tanh^2(Z/\mu) \right] \right]$$
$$\geq (1-\theta)\left(\frac{\theta}{\mu} - \frac{\mu}{32} \right) \mathbb{E}_{\mathcal{J}}\mathbb{E}_{Z\sim\mathcal{N}\left(0,\|\boldsymbol{w}_\mathcal{J}\|^2\right)}\left[1 - \tanh^2(Z/\mu) \right]$$
$$- \frac{\theta}{\mu}\mathbb{E}_{\mathcal{J}}\mathbb{E}_{Z\sim\mathcal{N}\left(0,\|\boldsymbol{w}_\mathcal{J}\|^2 + q_n^2\right)}\left[1 - \tanh^2(Z/\mu) \right], \tag{8.6}$$

where to obtain the last line above we have used the fact $\|\boldsymbol{w}\| \leq \mu/(4\sqrt{2})$ and that $\|\boldsymbol{w}_{\mathcal{J}}\| + q_n^2 \leq \|\boldsymbol{q}\|^2 = 1$. Now by Lemma A.1,

$$
\begin{aligned}
\mathbb{E}_{Z \sim \mathcal{N}\left(0, \|\boldsymbol{w}_{\mathcal{J}}\|^2\right)} \left[1 - \tanh^2(Z/\mu)\right] &= 2\mathbb{E}_{Z \sim \mathcal{N}\left(0, \|\boldsymbol{w}_{\mathcal{J}}\|^2\right)} \left[\left(1 - \tanh^2(Z/\mu)\right) \mathbb{1}_{Z>0}\right] \\
&\geq 2\mathbb{E}_{Z \sim \mathcal{N}\left(0, \|\boldsymbol{w}_{\mathcal{J}}\|^2\right)} \left[\exp(-2Z/\mu) \mathbb{1}_{Z>0}\right] \\
&= 2 \exp\left(\frac{2\|\boldsymbol{w}_{\mathcal{J}}\|^2}{\mu^2}\right) \Phi^c\left(\frac{2\|\boldsymbol{w}_{\mathcal{J}}\|}{\mu}\right) \quad \text{(by (B.3))} \\
&\geq \frac{1}{\sqrt{2\pi}}\left(\sqrt{\frac{4\|\boldsymbol{w}_{\mathcal{J}}\|^2}{\mu^2} + 4} - \frac{2\|\boldsymbol{w}_{\mathcal{J}}\|}{\mu}\right),
\end{aligned}
$$

where at the last inequality we have applied Gaussian tail lower bound of Type III in Lemma A.5. Also note the function $x \mapsto \sqrt{4+x^2} - x$ is monotonically decreasing and $\|\boldsymbol{w}_{\mathcal{J}}\| \leq \|\boldsymbol{w}\| \leq \mu/\left(4\sqrt{2}\right)$, we obtain

$$
\mathbb{E}_{Z \sim \mathcal{N}\left(0, \|\boldsymbol{w}_{\mathcal{J}}\|^2\right)} \left[1 - \tanh^2(Z/\mu)\right] \geq \frac{1}{\sqrt{2\pi}}\left(\sqrt{\frac{1}{8} + 4} - \frac{2}{4\sqrt{2}}\right) \geq \frac{3}{2\sqrt{2\pi}}. \tag{8.7}
$$

Similarly we also have

$$
\begin{aligned}
\mathbb{E}_{Z \sim \mathcal{N}\left(0, \|\boldsymbol{w}_{\mathcal{J}}\|^2 + q_n^2\right)} \left[1 - \tanh^2(Z/\mu)\right] &= 2\mathbb{E}_{Z \sim \mathcal{N}\left(0, \|\boldsymbol{w}_{\mathcal{J}}\|^2 + q_n^2\right)} \left[\left(1 - \tanh^2(Z/\mu)\right) \mathbb{1}_{Z>0}\right] \\
&\leq 8\mathbb{E}_{Z \sim \mathcal{N}\left(0, \|\boldsymbol{w}_{\mathcal{J}}\|^2 + q_n^2\right)} \left[\exp(-2Z/\mu) \mathbb{1}_{Z>0}\right] \\
&= 8 \exp\left(\frac{2\|\boldsymbol{w}_{\mathcal{J}}\|^2 + 2q_n^2}{\mu^2}\right) \Phi^c\left(\frac{2\sqrt{\|\boldsymbol{w}_{\mathcal{J}}\|^2 + q_n^2}}{\mu}\right) \quad \text{(by (B.3))} \\
&\leq \frac{4}{\sqrt{2\pi}} \frac{\mu}{\sqrt{\|\boldsymbol{w}_{\mathcal{J}}\|^2 + q_n^2}},
\end{aligned}
$$

where at the last inequality we have applied Gaussian tail upper bound of Type II in Lemma A.5. Since $\|\boldsymbol{w}_{\mathcal{J}}\|^2 + q_n^2 \geq q_n^2 = 1 - \|\boldsymbol{w}\|^2 \geq 1 - \mu^2/32 \geq 31/32$ for $\|\boldsymbol{w}\| \leq \mu/(4\sqrt{2})$ and $\mu \leq 1$, we obtain

$$
\mathbb{E}_{Z \sim \mathcal{N}\left(0, \|\boldsymbol{w}_{\mathcal{J}}\|^2 + q_n^2\right)} \left[1 - \tanh^2(Z/\mu)\right] \leq \frac{4}{\sqrt{2\pi}} \frac{\mu}{\sqrt{31/32}} \leq \frac{4}{\sqrt{2\pi}}\mu. \tag{8.8}
$$

Substituting (8.7) and (8.8) into (8.6), we obtain

$$
\begin{aligned}
&\frac{\theta}{\mu}\mathbb{E}\left[1 - \tanh^2\left(\frac{\boldsymbol{q}^*\boldsymbol{x}_k}{\mu}\right)\right] - \mathbb{E}\left[\tanh\left(\frac{\boldsymbol{q}^*\boldsymbol{x}_k}{\mu}\right)\boldsymbol{q}^*\boldsymbol{x}_k\right] \\
&\geq \frac{3}{2\sqrt{2\pi}}(1 - \theta)\left(\frac{\theta}{\mu} - \frac{\mu}{32}\right) - \frac{4}{\sqrt{2\pi}}\theta \\
&\geq \frac{3}{4\sqrt{2\pi}}\left(\frac{\theta}{\mu} - \frac{1}{10}\frac{\theta}{\mu}\right) - \frac{4}{10\sqrt{2\pi}}\frac{\theta}{\mu} \geq \frac{1}{4\sqrt{2\pi}}\frac{\theta}{\mu},
\end{aligned}
$$

where at the last line we have used the fact $\mu < \theta$ and that $\mu < 1/10$ to simplify. To summarize, we obtain that

$$U^* \operatorname{Hess} \mathbb{E}\left[f(q)\right] U \succeq \frac{1}{4\sqrt{2\pi}} \frac{\theta}{\mu}. \tag{8.9}$$

Next we perform concentration analysis. For any q, we can write

$$U^* \nabla^2 f(q) U = \frac{1}{p} \sum_{k=1}^{p} W_k, \quad \text{with } W_k \doteq \frac{1}{\mu}\left[1 - \tanh^2\left(\frac{q^* x_k}{\mu}\right)\right] U^* x_k x_k^* U.$$

For any integer $m \geq 2$, we have

$$0 \preceq \mathbb{E}\left[W_k^m\right] \preceq \frac{1}{\mu^m} \mathbb{E}\left[(U^* x_k x_k^* U)^m\right] \preceq \frac{1}{\mu^m} \mathbb{E}\left[\|x_k x_k^*\|^m\right] I = \frac{1}{\mu^m} \mathbb{E}\left[\|x_k\|^{2m}\right] I \preceq \frac{1}{\mu^m} \mathbb{E}_{Z \sim \xi^2(n)}\left[Z^m\right] I,$$

where we have used Lemma 7.4 to obtain the last inequality. By Lemma A.7, we obtain

$$0 \preceq \mathbb{E}\left[W_k^m\right] \preceq \frac{1}{\mu^m} \frac{m!}{2} (2n)^m I \preceq \frac{m!}{2}\left(\frac{2n}{\mu}\right)^m I.$$

Taking $R_W = 2n/\mu$, and $\sigma_W^2 = 4n^2/\mu^2 \geq \mathbb{E}\left[W_k^2\right]$, by Lemma A.10, we obtain

$$\mathbb{P}\left[\left\|\frac{1}{p}\sum_{k=1}^{p} W_k - \frac{1}{p}\sum_{k=1}^{p} \mathbb{E}\left[W_k\right]\right\| > t/2\right] \leq 2n \exp\left(-\frac{p\mu^2 t^2}{32n^2 + 8nt}\right) \tag{8.10}$$

for any $t > 0$. Similarly, we write

$$\langle \nabla f(q), q \rangle = \frac{1}{p}\sum_{k=1}^{p} Z_k, \quad \text{with } Z_k \doteq \tanh\left(\frac{q^* x_k}{\mu}\right) q^* x_k.$$

For any integer $m \geq 2$, we have

$$\mathbb{E}\left[|Z_k|^m\right] \leq \mathbb{E}\left[|q^* x_k|^m\right] \leq \mathbb{E}_{Z \sim \mathcal{N}(0,1)}\left[|Z|^m\right] \leq \frac{m!}{2},$$

where at the first inequality we used the fact $|\tanh(\cdot)| \leq 1$, at the second we invoked Lemma 7.4, and at the third we invoked Lemma A.6. Taking $R_Z = \sigma_Z^2 = 1$, by Lemma A.9, we obtain

$$\mathbb{P}\left[\left|\frac{1}{p}\sum_{k=1}^{p} Z_k - \frac{1}{p}\sum_{k=1}^{p} \mathbb{E}\left[Z_k\right]\right| > t/2\right] \leq 2 \exp\left(-pt^2/16\right) \tag{8.11}$$

for any $t > 0$. Gathering (8.10) and (8.11), we obtain that for any $t > 0$,

$$\mathbb{P}\left[\|U^* \operatorname{Hess} \mathbb{E}\left[f(q)\right] U - U^* \operatorname{Hess} f(q) U\| > t\right]$$
$$\leq \mathbb{P}\left[\left\|U^* \nabla^2 f(q) U - \nabla^2 \mathbb{E}\left[f(q)\right]\right\| > t/2\right] + \mathbb{P}\left[|\langle \nabla f(q), q \rangle - \langle \nabla \mathbb{E}\left[f(q)\right], q \rangle| > t/2\right]$$
$$\leq 2n \exp\left(-\frac{p\mu^2 t^2}{32n^2 + 8nt}\right) + 2 \exp\left(-\frac{pt^2}{16}\right) \leq 4n \exp\left(-\frac{p\mu^2 t^2}{32n^2 + 8nt}\right). \tag{8.12}$$

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 ε-net N_ε of size at most $(3\mu/(4\sqrt{2}\varepsilon))^n$ that covers the region $\{\boldsymbol{q} : \|\boldsymbol{w}(\boldsymbol{q})\| \leq \mu/(4\sqrt{2})\}$. By Lemma 8.2, the function $\mathrm{Hess}\, f(\boldsymbol{q})$ is locally Lipschitz within each normal ball of radius

$$\left\| \boldsymbol{q} - \exp_{\boldsymbol{q}}(1/2) \right\| = \sqrt{2 - 2\cos(1/2)} \geq 1/\sqrt{5}$$

with Lipschitz constant L_H (as defined in Lemma 8.2). 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 ε-ball centered around one element of the ε-net. Let

$$\mathcal{E}_\infty \doteq \left\{ 1 \leq \|\boldsymbol{X}_0\|_\infty \leq 4\sqrt{\log(np)} \right\}.$$

From Lemma 7.11, $\mathbb{P}[\mathcal{E}_\infty^c] \leq \theta (np)^{-7} + \exp(-0.3\theta np)$. By Lemma 8.2, with at least the same probability,

$$L_H \leq C_1 \frac{n^{3/2}}{\mu^2} \log^{3/2}(np).$$

Set $\varepsilon = \frac{\theta}{12\sqrt{2\pi}\mu L_H} < \mu/(4\sqrt{2})$, so

$$\#N_\varepsilon \leq \exp\left(n \log \frac{C_2 n^{3/2} \log^{3/2}(np)}{\theta} \right).$$

Let $\mathcal{E}_H$ denote the event that

$$\mathcal{E}_H \doteq \left\{ \max_{\boldsymbol{q} \in N_\varepsilon} \|\boldsymbol{U}^* \mathrm{Hess}\, \mathbb{E}[f(\boldsymbol{q})] \boldsymbol{U} - \boldsymbol{U}^* \mathrm{Hess}\, f(\boldsymbol{q}) \boldsymbol{U}\| \leq \frac{\theta}{12\sqrt{2\pi}\mu} \right\}.$$

On $\mathcal{E}_\infty \cap \mathcal{E}_H$,

$$\sup_{\boldsymbol{q}:\|\boldsymbol{w}(\boldsymbol{q})\| \leq \mu/(4\sqrt{2})} \|\boldsymbol{U}^* \mathrm{Hess}\, \mathbb{E}[f(\boldsymbol{q})] \boldsymbol{U} - \boldsymbol{U}^* \mathrm{Hess}\, f(\boldsymbol{q}) \boldsymbol{U}\| \leq \frac{\theta}{6\sqrt{2\pi}\mu}.$$

So on $\mathcal{E}_\infty \cap \mathcal{E}_H$, we have

$$\boldsymbol{U}^* \mathrm{Hess}\, f(\boldsymbol{q}) \boldsymbol{U} \succeq c_\sharp \frac{\theta}{\mu} \tag{8.13}$$

for any $c_\sharp \leq 1/(12\sqrt{2\pi})$. Setting $t = \frac{\theta}{12\sqrt{2\pi}\mu}$ in (8.12), we obtain that for any fixed $\boldsymbol{q}$ in this region,

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

Taking a union bound, we obtain that

$$\mathbb{P}[\mathcal{E}_H^c] \leq 4n \exp\left(-\frac{p\theta^2}{c_3 n^2 + c_4 n\theta/\mu} + C_5 n \log n + C_6 n \log\log p \right).$$

It is enough to make $p \geq C_7 n^3 \log(n/(\mu\theta))/(\mu\theta^2)$ to make the failure probability small, completing the proof. ∎

Proof [of Lemma 3.11] For a given q, consider the vector $r \doteq q - e_n/q_n$. It is easy to verify that $\langle q, r \rangle = 0$, and hence $r \in T_q\mathbb{S}^{n-1}$. Now, by (3.12) and (3.14), we have

$$
\begin{aligned}
\langle \operatorname{grad} f(q), r \rangle &= \langle (I - qq^*) \nabla f(q), q - e_n/q_n \rangle \\
&= \langle (I - qq^*) \nabla f(q), -e_n/q_n \rangle \\
&= \frac{1}{p} \sum_{k=1}^{p} \left\langle (I - qq^*) \tanh\left(\frac{q^* x_k}{\mu} \right) x_k, -e_n/q_n \right\rangle \\
&= \frac{1}{p} \sum_{k=1}^{p} \tanh\left(\frac{q^* x_k}{\mu} \right) \left(-\frac{x_k(n)}{q_n} + q^* x_k \right) \\
&= \frac{1}{p} \sum_{k=1}^{p} \tanh\left(\frac{q^* x_k}{\mu} \right) \left(w^*(q) \overline{x}_k - \frac{x_k(n)}{q_n} \|w(q)\|^2 \right) \\
&= w^*(q) \nabla g(w),
\end{aligned}
$$

where to get the last line we have used (7.1). Thus,

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

where

$$
\frac{\|r\|^2}{\|w\|^2} = \frac{\|w\|^2 + \left(q_n - \frac{1}{q_n} \right)^2}{\|w\|^2} = \frac{\|w\|^2 + \|w\|^4/q_n^2}{\|w\|^2} = \frac{1}{q_n^2} = \frac{1}{1 - \|w\|^2} \leq \frac{1}{1 - \frac{1}{2000}} = \frac{2000}{1999},
$$

where we have invoked our assumption that $\|w\| \leq \frac{1}{20\sqrt{5}}$. Therefore we obtain

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

completing the proof. ∎

Proof of Lemma 3.12 combines the local Lipschitz property of Hess $f(q)$ in Lemma 8.2, and the Taylor's theorem (manifold version, Lemma 7.4.7 of [AMS09]).

Proof [of Lemma 3.12] Let $\gamma(t)$ be the unique geodesic that satisfies $\gamma(0) = q^{(k)}$, $\gamma(1) = q^{(k+1)}$, and its directional derivative $\dot{\gamma}(0) = \delta_\star$. Since the parallel translation defined by the Riemannian connection is an isometry, then $\left\| \operatorname{grad} f(q^{(k+1)}) \right\| = \left\| \mathcal{P}_\gamma^{0 \leftarrow 1} \operatorname{grad} f(q^{(k+1)}) \right\|$. Moreover, since $\|\delta_\star\| \leq \Delta$, the unconstrained optimality condition in (3.16) implies that $\operatorname{grad} f(q^{(k)}) + \operatorname{Hess} f(q^{(k)}) \delta_\star = \mathbf{0}_{q^{(k)}}$. Thus, by using Taylor's theorem in [AMS09], we have

$$
\begin{aligned}
\left\| \operatorname{grad} f(q^{(k+1)}) \right\| &= \left\| \mathcal{P}_\gamma^{0 \leftarrow 1} \operatorname{grad} f\left(q^{(k+1)} \right) - \operatorname{grad} f\left(q^{(k)} \right) - \operatorname{Hess} f\left(q^{(k)} \right) \delta_\star \right\| \\
&= \left\| \int_0^1 \left[\mathcal{P}_\gamma^{0 \leftarrow t} \operatorname{Hess} f(\gamma(t)) [\dot{\gamma}(t)] - \operatorname{Hess} f\left(q^{(k)} \right) \delta_\star \right] dt \right\| \quad \text{(Taylor's theorem)}
\end{aligned}
$$

$$= \left\| \int_0^1 \left(\mathcal{P}_\gamma^{0 \leftarrow t} \operatorname{Hess} f \left(\gamma \left(t \right) \right) \mathcal{P}_\gamma^{t \leftarrow 0} \boldsymbol{\delta}_\star - \operatorname{Hess} f \left(\boldsymbol{q}^{(k)} \right) \boldsymbol{\delta}_\star \right) \, dt \right\|$$

$$\leq \left\| \boldsymbol{\delta}_\star \right\| \int_0^1 \left\| \mathcal{P}_\gamma^{0 \leftarrow t} \operatorname{Hess} f \left(\gamma \left(t \right) \right) \mathcal{P}_\gamma^{t \leftarrow 0} - \operatorname{Hess} f \left(\boldsymbol{q}^{(k)} \right) \right\| \, dt.$$

From the Lipschitz bound in Lemma 8.2 and the optimality condition in (3.16), we obtain

$$\left\| \operatorname{grad} f \left(\boldsymbol{q}^{(k+1)} \right) \right\| \leq \frac{1}{2} \left\| \boldsymbol{\delta}_\star \right\|^2 L_H = \frac{L_H}{2 m_H^2} \left\| \operatorname{grad} f \left(\boldsymbol{q}^{(k)} \right) \right\|^2.$$

This completes the proof. ∎

Proof [of Lemma 3.14] By invoking Taylor's theorem in [AMS09], we have

$$\mathcal{P}_\gamma^{0 \leftarrow \tau} \operatorname{grad} f \left(\gamma \left(\tau \right) \right) = \int_0^\tau \mathcal{P}_\gamma^{0 \leftarrow t} \operatorname{Hess} f \left(\gamma \left(t \right) \right) \left[\dot{\gamma} \left(t \right) \right] dt.$$

Hence, we have

$$\left\langle \mathcal{P}_\gamma^{0 \leftarrow \tau} \operatorname{grad} f \left(\gamma \left(\tau \right) \right), \boldsymbol{\delta} \right\rangle = \int_0^\tau \left\langle \mathcal{P}_\gamma^{0 \leftarrow t} \operatorname{Hess} f \left(\gamma \left(t \right) \right) \left[\dot{\gamma} \left(t \right) \right], \boldsymbol{\delta} \right\rangle \, dt$$

$$= \int_0^\tau \left\langle \mathcal{P}_\gamma^{0 \leftarrow t} \operatorname{Hess} f \left(\gamma \left(t \right) \right) \left[\dot{\gamma} \left(t \right) \right], \mathcal{P}_\gamma^{0 \leftarrow t} \dot{\gamma} \left(t \right) \right\rangle \, dt$$

$$= \int_0^\tau \left\langle \operatorname{Hess} f \left(\gamma \left(t \right) \right) \left[\dot{\gamma} \left(t \right) \right], \dot{\gamma} \left(t \right) \right\rangle \, dt$$

$$\geq m_H \int_0^\tau \left\| \dot{\gamma} \left(t \right) \right\|^2 \, dt \geq m_H \tau \left\| \boldsymbol{\delta} \right\|^2,$$

where we have used the fact that the parallel transport $\mathcal{P}_\gamma^{0 \leftarrow t}$ defined by the Riemannian connection is an isometry. On the other hand, we have

$$\left\langle \mathcal{P}_\gamma^{0 \leftarrow \tau} \operatorname{grad} f \left(\gamma \left(\tau \right) \right), \boldsymbol{\delta} \right\rangle \leq \left\| \mathcal{P}_\gamma^{0 \leftarrow \tau} \operatorname{grad} f \left(\gamma \left(\tau \right) \right) \right\| \left\| \boldsymbol{\delta} \right\| = \left\| \operatorname{grad} f \left(\gamma \left(\tau \right) \right) \right\| \left\| \boldsymbol{\delta} \right\|,$$

where again used the isometry property of the operator $\mathcal{P}_\gamma^{0 \leftarrow \tau}$. Combining the two bounds above, we obtain

$$\left\| \operatorname{grad} f \left(\gamma \left(\tau \right) \right) \right\| \left\| \boldsymbol{\delta} \right\| \geq m_H \tau \left\| \boldsymbol{\delta} \right\|^2,$$

which implies the claimed result. ∎

9 Proofs of Technical Results for Section 4

We need one technical lemma to prove Lemma 4.2 and the relevant lemma for complete dictionaries.

Lemma 9.1 *There exists a positive constant C, such that for all integer $n_1 \in \mathbb{N}$, $\theta \in (0, 1/3)$, and $n_2 \in \mathbb{N}$ with $n_2 \geq C n_1 \log \left(n_1 / \theta \right) / \theta^2$, any random matrix $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2} \sim_{i.i.d.} \mathrm{BG}(\theta)$ obeys the following. For any fixed index set $\mathcal{I} \subset [n_2]$ with $|\mathcal{I}| \leq \frac{9}{8} \theta n_2$, it holds that*

$$\left\| \boldsymbol{v}^* \boldsymbol{M}_{\mathcal{I}^c} \right\|_1 - \left\| \boldsymbol{v}^* \boldsymbol{M}_{\mathcal{I}} \right\|_1 \geq \frac{n_2}{6} \sqrt{\frac{2}{\pi}} \theta \left\| \boldsymbol{v} \right\| \quad \textit{for all } \boldsymbol{v} \in \mathbb{R}^{n_1},$$

with probability at least $1 - n_2^{-10} - \theta \left(n_1 n_2 \right)^{-7} - \exp \left(-0.3 \theta n_1 n_2 \right)$.

Proof By homogeneity, it is sufficient to consider all $v \in \mathbb{S}^{n_1}$. 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

$$T(v) \doteq \|v^* M_{\mathcal{I}^c}\|_1 - \|v^* M_{\mathcal{I}}\|_1 = \sum_{i \in \mathcal{I}^c} |v^* m_i| - \sum_{i \in \mathcal{I}} |v^* m_i|,$$

namely as a sum of independent random variables. Since $|\mathcal{I}| \leq 9n_2\theta/8$, we have

$$\mathbb{E}[T(v)] \geq \left(n_2 - \frac{9}{8}\theta n_2 - \frac{9}{8}\theta n_2\right) \mathbb{E}[|v^* m_1|] = \left(1 - \frac{9}{4}\theta\right) n_2 \mathbb{E}[|v^* m_1|] \geq \frac{1}{4} n_2 \mathbb{E}[|v^* m_1|],$$

where the expectation $\mathbb{E}[|v^* m_1|]$ can be lower bounded as

$$\mathbb{E}[|v^* m_1|] = \sum_{k=0}^{n_1} \theta^k (1-\theta)^{n_1-k} \sum_{\mathcal{J} \in \binom{[n_1]}{k}} \mathbb{E}_{g \sim \mathcal{N}(0,I)}[|v_{\mathcal{J}}^* g|]$$

$$= \sum_{k=0}^{n_1} \theta^k (1-\theta)^{n_1-k} \sum_{\mathcal{J} \in \binom{[n_1]}{k}} \sqrt{\frac{2}{\pi}} \|v_{\mathcal{J}}\| \geq \sqrt{\frac{2}{\pi}} \|\mathbb{E}_{\mathcal{J}}[v_{\mathcal{J}}]\| = \sqrt{\frac{2}{\pi}}\theta.$$

Moreover, by Lemma 7.4 and Lemma A.6, for any $i \in [n_2]$ and any integer $m \geq 2$,

$$\mathbb{E}[|v^* m_i|^m] \leq \mathbb{E}_{Z \sim \mathcal{N}(0,1)}[|Z|^m] \leq (m-1)!! \leq \frac{m!}{2}.$$

So invoking the moment-control Bernstein's inequality in Lemma A.9, we obtain

$$\mathbb{P}\left[T(v) < \frac{n_2}{4}\sqrt{\frac{2}{\pi}}\theta - t\right] \leq \mathbb{P}[T(v) < \mathbb{E}[T(v)] - t] \leq \exp\left(-\frac{t^2}{2n_2 + 2t}\right).$$

Taking $t = \frac{n_2}{20}\sqrt{\frac{2}{\pi}}\theta$ and simplifying, we obtain that

$$\mathbb{P}\left[T(v) < \frac{n_2}{5}\sqrt{\frac{2}{\pi}}\theta\right] \leq \exp\left(-c_1\theta^2 n_2\right) \tag{9.1}$$

for some positive constant c_1. Fix $\varepsilon = \sqrt{\frac{2}{\pi}}\frac{\theta}{120}[n_1 \log(n_1 n_2)]^{-1/2} < 1$. The unit sphere $\mathbb{S}^{n_1}$ has an ε-net N_ε of cardinality at most $(3/\varepsilon)^{n_1}$. Consider the event

$$\mathcal{E}_{bg} \doteq \left\{T(v) \geq \frac{n_2}{5}\sqrt{\frac{2}{\pi}}\theta \ \forall \ v \in N_\varepsilon\right\}.$$

A simple union bound implies

$$\mathbb{P}[\mathcal{E}_{bg}^c] \leq \exp\left(-c_1\theta^2 n_2 + n_1 \log\left(\frac{3}{\varepsilon}\right)\right) \leq \exp\left(-c_1\theta^2 n_2 + c_2 n_1 \log\frac{n_1 \log n_2}{\theta}\right), \tag{9.2}$$

where $c_2 > 0$ is numerical. Conditioned on $\mathcal{E}_{bg}$, we have that any $z \in \mathbb{S}^{n_1-1}$ can be written as $z = v + e$ for some $v \in N_\varepsilon$ and $\|e\| \leq \varepsilon$. Moreover,

$$T(z) = \|(v+e)^* M_{\mathcal{I}^c}\|_1 - \|(v+e)^* M_{\mathcal{I}}\|_1 \geq T(v) - \|e^* M_{\mathcal{I}^c}\|_1 - \|e^* M_{\mathcal{I}}\|_1$$

$$= \frac{n_2}{5}\sqrt{\frac{2}{\pi}}\theta - \|\boldsymbol{e}^*\boldsymbol{M}\|_1 = \frac{n_2}{5}\sqrt{\frac{2}{\pi}}\theta - \sum_{k=1}^{n_2}|\boldsymbol{e}^*\boldsymbol{m}_k|$$

$$\geq \frac{n_2}{5}\sqrt{\frac{2}{\pi}}\theta - \varepsilon\sum_{k=1}^{n_2}\|\boldsymbol{m}_k\|.$$

By Lemma 7.11, with probability at least $1 - \theta(n_1 n_2)^{-7} - \exp(-0.3\theta n_1 n_2)$, $\|\boldsymbol{M}\|_\infty \leq 4\sqrt{\log(n_1 n_2)}$. Thus,

$$T(\boldsymbol{z}) \geq \frac{n_2}{5}\sqrt{\frac{2}{\pi}}\theta - \sqrt{\frac{2}{\pi}}\frac{\theta}{120}\frac{n_2\sqrt{n_1}4\sqrt{\log(n_1 n_2)}}{\sqrt{n_1}\sqrt{\log(n_1 n_2)}} = \frac{n_2}{6}\sqrt{\frac{2}{\pi}}\theta. \tag{9.3}$$

Thus, by (9.2), it is enough to take $n_2 > Cn_1\log(n_1/\theta)/\theta^2$ for sufficiently large $C > 0$ to make the overall failure probability small enough so that the lower bound (9.3) holds. $\blacksquare$

Proof [Proof of Lemma 4.2] The proof is similar to that of [QSW14]. First, let us assume the dictionary $\boldsymbol{A}_0 = \boldsymbol{I}$. Wlog, suppose that the Riemannian TRM algorithm returns a solution $\hat{\boldsymbol{q}}$, to which $\boldsymbol{e}_n$ is the nearest signed basis vector. Thus, the rounding LP (4.1) takes the form:

$$\operatorname{minimize}_{\boldsymbol{q}} \|\boldsymbol{q}^*\boldsymbol{X}_0\|_1, \quad \text{subject to} \quad \langle\boldsymbol{r},\boldsymbol{q}\rangle = 1. \tag{9.4}$$

where the vector $\boldsymbol{r} = \hat{\boldsymbol{q}}$. Next, We will show whenever $\hat{\boldsymbol{q}}$ is close enough to $\boldsymbol{e}_n$, w.h.p., the above linear program returns $\boldsymbol{e}_n$. Let $\boldsymbol{X}_0 = [\overline{\boldsymbol{X}}; \boldsymbol{x}_n^*]$, where $\overline{\boldsymbol{X}} \in \mathbb{R}^{(n-1)\times p}$ and $\boldsymbol{x}_n^*$ is the last row of $\boldsymbol{X}_0$. Set $\boldsymbol{q} = [\overline{\boldsymbol{q}}, q_n]$, where $\overline{\boldsymbol{q}}$ denotes the first $n-1$ coordinates of $\boldsymbol{q}$ and q_n is the last coordinate; similarly for $\boldsymbol{r}$. Let us consider a relaxation of the problem (9.4),

$$\operatorname{minimize}_{\boldsymbol{q}} \|\boldsymbol{q}^*\boldsymbol{X}_0\|_1, \quad \text{subject to} \quad q_n r_n + \langle\overline{\boldsymbol{q}},\overline{\boldsymbol{r}}\rangle \geq 1, \tag{9.5}$$

It is obvious that the feasible set of (9.5) contains that of (9.4). So if $\boldsymbol{e}_n$ is the unique optimal solution (UOS) of (9.5), it is the UOS of (9.4). Suppose $\mathcal{I} = \operatorname{supp}(\boldsymbol{x}_n)$ and define an event $\mathcal{E}_0 = \{|\mathcal{I}| \leq \frac{9}{8}\theta p\}$. By Hoeffding's inequality, we know that $\mathbb{P}[\mathcal{E}_0^c] \leq \exp(-\theta^2 p/2)$. Now conditioned on $\mathcal{E}_0$ and consider a fixed support $\mathcal{I}$. (9.5) can be further relaxed as

$$\operatorname{minimize}_{\boldsymbol{q}} \|\boldsymbol{x}_n\|_1 |q_n| - \|\overline{\boldsymbol{q}}^*\overline{\boldsymbol{X}}_{\mathcal{I}}\|_1 + \|\overline{\boldsymbol{q}}^*\overline{\boldsymbol{X}}_{\mathcal{I}^c}\|_1, \quad \text{subject to} \quad q_n r_n + \|\overline{\boldsymbol{q}}\|\,\|\overline{\boldsymbol{r}}\| \geq 1. \tag{9.6}$$

The objective value of (9.6) lower bounds that of (9.5), and are equal when $\boldsymbol{q} = \boldsymbol{e}_n$. So if $\boldsymbol{q} = \boldsymbol{e}_n$ is UOS of (9.6), it is UOS of (9.4). By Lemma 9.1, we know that

$$\left\|\overline{\boldsymbol{q}}^*\overline{\boldsymbol{X}}_{\mathcal{I}^c}\right\|_1 - \left\|\overline{\boldsymbol{q}}^*\overline{\boldsymbol{X}}_{\mathcal{I}}\right\|_1 \geq \frac{p}{6}\sqrt{\frac{2}{\pi}}\theta\|\overline{\boldsymbol{q}}\|$$

holds w.h.p. when $p \geq C_1(n-1)\log((n-1)/\theta)/\theta^2$. Let $\zeta = \frac{p}{6}\sqrt{\frac{2}{\pi}}\theta$, thus we can further lower bound the objective value in (9.6) by

$$\operatorname{minimize}_{\boldsymbol{q}} \|\boldsymbol{x}_n\|_1 |q_n| + \zeta\|\overline{\boldsymbol{q}}\|, \quad \text{subject to} \quad q_n r_n + \|\overline{\boldsymbol{q}}\|\,\|\overline{\boldsymbol{r}}\| \geq 1. \tag{9.7}$$

By similar arguments, if $\boldsymbol{e}_n$ is the UOS of (9.7), it is also the UOS of (9.4). For the optimal solution of (9.7), notice that it is necessary to have $\operatorname{sign}(q_n) = \operatorname{sign}(r_n)$ and $q_n r_n + \|\overline{\boldsymbol{q}}\|\,\|\overline{\boldsymbol{r}}\| = 1$. Therefore, the problem (9.7) is equivalent to

$$\operatorname{minimize}_{q_n} \|\boldsymbol{x}_n\|_1 |q_n| + \zeta\frac{1 - |r_n|\,|q_n|}{\|\overline{\boldsymbol{r}}\|}, \quad \text{subject to} \quad |q_n| \leq \frac{1}{|r_n|}. \tag{9.8}$$

Notice that the problem (9.8) 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, $\boldsymbol{q} = \boldsymbol{e}_n$ is the UOS of (9.8) if and only if

$$\frac{1}{|r_n|} \|\boldsymbol{x}_n\|_1 < \frac{\zeta}{\|\overline{\boldsymbol{r}}\|}. \tag{9.9}$$

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

$$\mathbb{P}\left[\|\boldsymbol{x}_n\|_1 \geq \frac{9}{8}\sqrt{\frac{2}{\pi}}\theta p + t\right] \;\leq\; \mathbb{P}\left[\|\boldsymbol{x}_n\|_1 \geq \mathbb{E}\left[\|\boldsymbol{x}_n\|_1\right] + t\right] \;\leq\; \exp\left(-\frac{t^2}{2p}\right),$$

which means that

$$\mathbb{P}\left[\|\boldsymbol{x}_n\|_1 \geq \frac{5}{4}\sqrt{\frac{2}{\pi}}\theta p\right] \;\leq\; \exp\left(-\frac{\theta^2 p}{64\pi}\right). \tag{9.10}$$

Therefore, by (9.9) and (9.10), for $\boldsymbol{q} = \boldsymbol{e}_n$ to be the UOS of (9.4) w.h.p., it is sufficient to have

$$\frac{5}{4\,|r_n|}\sqrt{\frac{2}{\pi}}\theta p \;<\; \frac{\theta p}{6\sqrt{1 - |r_n|^2}}\sqrt{\frac{2}{\pi}}, \tag{9.11}$$

which is implied 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 $\boldsymbol{A}_0$ is an arbitrary orthogonal matrix, it only rotates the row subspace of $\boldsymbol{X}_0$. Thus, wlog, suppose the TRM algorithm returns a solution $\widehat{\boldsymbol{q}}$, to which $\boldsymbol{A}_0 \boldsymbol{q}_\star$ is the nearest "target" with $\boldsymbol{q}_\star$ a signed basis vector. By a change of variable $\tilde{\boldsymbol{q}} = \boldsymbol{A}_0^* \boldsymbol{q}$, the problem (9.4) is of the form

$$\text{minimize}_{\tilde{\boldsymbol{q}}} \|\widetilde{\boldsymbol{q}}^* \boldsymbol{X}_0\|_1, \quad \text{subject to} \quad \langle \boldsymbol{A}_0^* \boldsymbol{r}, \tilde{\boldsymbol{q}} \rangle = 1,$$

obviously our target solution for $\tilde{\boldsymbol{q}}$ is again the standard basis $\boldsymbol{q}_\star$. By a similar argument above, we only need $\langle \boldsymbol{A}_0^* \boldsymbol{r}, \boldsymbol{e}_n \rangle > 249/250$ to exactly recover the target, which is equivalent to $\langle \boldsymbol{r}, \widehat{\boldsymbol{q}}_\star \rangle > 249/250$. This implies that our rounding (4.1) is invariant to change of basis, completing the proof. ∎

Proof [of Lemma 4.4] Define $\widetilde{\boldsymbol{q}} \doteq (\boldsymbol{U}\boldsymbol{V}^* + \boldsymbol{\Xi})^* \boldsymbol{q}$. By Lemma 2.14, and in particular (2.15), when $p \geq \frac{C}{c_\star^2 \theta} \max\left\{\frac{n^4}{\mu^4}, \frac{n^5}{\mu^2}\right\} \kappa^8\left(\boldsymbol{A}_0\right) \log^4\left(\frac{\kappa(\boldsymbol{A}_0)n}{\mu\theta}\right)$, $\|\boldsymbol{\Xi}\| \leq 1/2$ so that $\boldsymbol{U}\boldsymbol{V}^* + \boldsymbol{\Xi}$ is invertible. Then the LP rounding can be written as

$$\text{minimize}_{\widetilde{\boldsymbol{q}}} \|\widetilde{\boldsymbol{q}}^* \boldsymbol{X}_0\|_1, \quad \text{subject to} \quad \langle (\boldsymbol{U}\boldsymbol{V}^* + \boldsymbol{\Xi})^{-1} \boldsymbol{r}, \widetilde{\boldsymbol{q}} \rangle = 1.$$

By Lemma 4.2, to obtain $\widetilde{\boldsymbol{q}} = \boldsymbol{e}_n$ from this LP, it is enough to have

$$\langle (\boldsymbol{U}\boldsymbol{V}^* + \boldsymbol{\Xi})^{-1} \boldsymbol{r}, \boldsymbol{e}_n \rangle \geq 249/250,$$

and $p \geq Cn^2 \log(n/\theta)/\theta$ for some large enough C. This implies that to obtain $q_\star$ for the original LP, such that $(UV^* + \Xi)^* q_\star = e_n$, it is enough that

$$\left\langle (UV^* + \Xi)^{-1} r, (UV^* + \Xi)^* q_\star \right\rangle = \langle r, q_\star \rangle \geq 249/250,$$

completing the proof. ∎

Proof [of Lemma 4.5] Note that $[q_\star^1, \ldots, q_\star^\ell] = (Q^* + \Xi^*)^{-1}[e_1, \ldots, e_\ell]$, we have

$$U^*(Q + \Xi)X_0 = U^*(Q^* + \Xi^*)^{-1}(Q + \Xi)^*(Q + \Xi)X_0$$
$$= U^* \left[q_\star^1, \ldots, q_\star^\ell \mid \widehat{V} \right] (I + \Delta_1)X_0,$$

where $\widehat{V} \doteq (Q^* + \Xi^*)^{-1}[e_{\ell+1}, \ldots, e_n]$, and the matrix $\Delta_1 = Q^* \Xi + \Xi^* Q + \Xi^* \Xi$ so that $\|\Delta_1\| \leq 3 \|\Xi\|$. Since $U^* \left[q_\star^1, \ldots, q_\star^\ell \mid \widehat{V} \right] = \left[0 \mid U^* \widehat{V} \right]$, we have

$$U^*(Q + \Xi)X_0 = \left[0 \mid U^* \widehat{V} \right] X_0 + \left[0 \mid U^* \widehat{V} \right] \Delta_1 X_0 = U^* \widehat{V} X_0^{[n-\ell]} + \Delta_2 X_0, \tag{9.12}$$

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

$$\|\Delta_2\| \leq \frac{\|\Delta_1\|}{\sigma_{\min}(Q + \Xi)} \leq \frac{3 \|\Xi\|}{\sigma_{\min}(Q + \Xi)} \leq \frac{3\delta}{1 - \delta}. \tag{9.13}$$

Since the matrix $\widehat{V}$ is near orthogonal, it can be decomposed as $\widehat{V} = V + \Delta_3$, where V is orthogonal, and Δ_3 is a small perturbation. Obviously, $V = UR$ for some orthogonal matrix R, so that spans the same subspace as that of U. Next, we control the spectral norm of Δ_3 so that it is sufficiently small,

$$\|\Delta_3\| = \min_{R \in O_\ell} \left\| UR - \widehat{V} \right\| \leq \min_{R \in O_\ell} \left\| UR - Q_{[n-\ell]} \right\| + \left\| Q_{[n-\ell]} - \widehat{V} \right\|, \tag{9.14}$$

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

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

where we have used perturbation bound for matrix inverse (see, e.g., Theorem 2.5 of Chapter III in [SS90]). To bound the first term, from Lemma B.4, it is enough to upper bound the largest principal angle θ_1 between the subspaces $\text{span}([q_\star^1, \ldots, q_\star^\ell])$, and that spanned by $Q[e_1, \ldots, e_\ell]$. Write $I_{[\ell]} \doteq [e_1, \ldots, e_\ell]$ for short, we bound $\sin \theta_1$ as

$$\sin \theta_1 \leq \left\| QI_{[\ell]} I_{[\ell]}^* Q^* - (Q^* + \Xi^*)^{-1} I_{[\ell]} \left(I_{[\ell]}^* (Q + \Xi)^{-1} (Q^* + \Xi^*)^{-1} I_{[\ell]} \right)^{-1} I_{[\ell]}^* (Q + \Xi)^{-1} \right\|$$

$$= \left\| QI_{[\ell]} I_{[\ell]}^* Q^* - (Q^* + \Xi^*)^{-1} I_{[\ell]} \left(I_{[\ell]}^* (I + \Delta_1)^{-1} I_{[\ell]} \right)^{-1} I_{[\ell]}^* (Q + \Xi)^{-1} \right\|$$

$$\leq \left\| QI_{[\ell]} I_{[\ell]}^* Q^* - (Q^* + \Xi^*)^{-1} I_{[\ell]} I_{[\ell]}^* (Q + \Xi)^{-1} \right\|$$

87

$$+ \left\| (Q^* + \Xi^*)^{-1} I_{[\ell]} \left[I - \left(I_{[\ell]}^* (I + \Delta_1)^{-1} I_{[\ell]} \right)^{-1} \right] I_{[\ell]}^* (Q + \Xi)^{-1} \right\|$$

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

$$\leq \left(1 + \frac{1}{1 - \delta} \right) \frac{\delta}{1 - \delta} + \frac{1}{(1 - \delta)^2} \frac{\left\| I_{[\ell]}^* (I + \Delta_1)^{-1} I_{[\ell]} - I \right\|}{1 - \left\| I_{[\ell]}^* (I + \Delta_1)^{-1} I_{[\ell]} - I \right\|}$$

$$\leq \left(1 + \frac{1}{1 - \delta} \right) \frac{\delta}{1 - \delta} + \frac{1}{(1 - \delta)^2} \frac{\| \Delta_1 \|}{1 - 2 \| \Delta_1 \|},$$

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

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

For $\delta < 1/20$, the upper bound is nontrivial. By Lemma B.4,

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

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

$$\| \Delta_3 \| \leq \delta/(1 - \delta) + 8\sqrt{2}\delta \leq 12.5\delta. \tag{9.15}$$

Therefore, by (9.12), we obtain

$$U^*(Q + \Xi)X_0 = U^*V X_0^{[n-\ell]} + \Delta, \quad \text{with} \quad \Delta \doteq U^*\Delta_3 X_0^{[n-\ell]} + \Delta_2 X_0. \tag{9.16}$$

By using the results in (9.13) and (9.15), we get the desired result. ∎

Appendices

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 A.1 (Derivates and Lipschitz Properties of $h_\mu(z)$) *For the sparsity surrogate*

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

the first two derivatives are

$$\dot{h}_\mu(z) = \tanh\left(\frac{z}{\mu}\right), \quad \ddot{h}_\mu(z) = \frac{1}{\mu}\left[1 - \tanh^2\left(\frac{z}{\mu}\right)\right]. \tag{A.1}$$

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

$$\frac{1}{2}\left(1 - \exp\left(-\frac{2z}{\mu}\right)\right) \leq \tanh\left(\frac{z}{\mu}\right) \leq 1 - \exp\left(-\frac{2z}{\mu}\right), \tag{A.2}$$

$$\exp\left(-\frac{2z}{\mu}\right) \leq 1 - \tanh^2\left(\frac{z}{\mu}\right) \leq 4\exp\left(-\frac{2z}{\mu}\right). \tag{A.3}$$

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

$$\left|\dot{h}_\mu(z) - \dot{h}_\mu(z')\right| \leq \frac{1}{\mu}\left|z - z'\right|, \quad \left|\ddot{h}_\mu(z) - \ddot{h}_\mu(z')\right| \leq \frac{2}{\mu^2}\left|z - z'\right| \tag{A.4}$$

Lemma A.2 (Chebyshev's Association Inequality) *Let X denote a real-valued random variable, and $f, g : \mathbb{R} \mapsto \mathbb{R}$ nondecreasing (nonincreasing) functions of X with $\mathbb{E}[f(X)] < \infty$ and $\mathbb{E}[g(X)] < \infty$. Then*

$$\mathbb{E}[f(X)g(X)] \geq \mathbb{E}[f(X)]\mathbb{E}[g(X)]. \tag{A.5}$$

If f is nondecreasing (nonincreasing) and g is nonincreasing (nondecreasing), we have

$$\mathbb{E}[f(X)g(X)] \leq \mathbb{E}[f(X)]\mathbb{E}[g(X)]. \tag{A.6}$$

Proof Consider Y, an independent copy of X. Then it is easy to see

$$\mathbb{E}[(f(X) - f(Y))(g(X) - g(Y))] \geq 0.$$

Expanding the expectation and noticing $\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(Y)g(X)] = \mathbb{E}[f(X)]\mathbb{E}[g(X)]$ and also $\mathbb{E}[f(X)g(X)] = \mathbb{E}[f(Y)g(Y)]$ yields the result. Similarly, we can prove the second one. ∎

This lemma implies the following lemma.

Lemma A.3 (Harris' Inequality, [Har60], see also Theorem 2.15 of [BLM13]) *Let $X_1, \ldots, X_n$ be independent, real-valued random variables and $f, g : \mathbb{R}^n \mapsto \mathbb{R}$ be nonincreasing (nondecreasing) w.r.t. any one variable while fixing the others. Define a random vector $\boldsymbol{X} = (X_1, \cdots, X_n) \in \mathbb{R}^n$, then we have*

$$\mathbb{E}[f(\boldsymbol{X})g(\boldsymbol{X})] \geq \mathbb{E}[f(\boldsymbol{X})]\mathbb{E}[g(\boldsymbol{X})]. \tag{A.7}$$

Similarly, if f is nondecreasing (nonincreasing) and g is nonincreasing (nondecreasing) coordinatewise in the above sense, we have

$$\mathbb{E}[f(\boldsymbol{X})g(\boldsymbol{X})] \leq \mathbb{E}[f(\boldsymbol{X})]\mathbb{E}[g(\boldsymbol{X})]. \tag{A.8}$$

Proof Again, it suffices to prove the first equality, which can be shown by induction. For $n = 1$, it reduces to Lemma A.2. Suppose the claim is true for any $m < n$. Since both g and f are nondecreasing functions in X_n given $\widehat{\boldsymbol{X}} = (X_1, \cdots, X_{n-1})$, then

$$\mathbb{E}[f(\boldsymbol{X})g(\boldsymbol{X})] = \mathbb{E}\left[\mathbb{E}\left[f(\boldsymbol{X})g(\boldsymbol{X}) \mid \widehat{\boldsymbol{X}}\right]\right] \geq \mathbb{E}\left[\mathbb{E}\left[f(\boldsymbol{X}) \mid \widehat{\boldsymbol{X}}\right]\mathbb{E}\left[g(\boldsymbol{X}) \mid \widehat{\boldsymbol{X}}\right]\right]$$

Now, it follows by independence that $f'\left(\widehat{\boldsymbol{X}}\right) = \mathbb{E}\left[f(\boldsymbol{X}) \mid \widehat{\boldsymbol{X}}\right]$ and $g'\left(\widehat{\boldsymbol{X}}\right) = \mathbb{E}\left[g(\boldsymbol{X}) \mid \widehat{\boldsymbol{X}}\right]$ are both nondecreasing functions, then by the induction hypothesis, we have

$$\mathbb{E}\left[f\left(\boldsymbol{X}\right) g\left(\boldsymbol{X}\right)\right] \geq \mathbb{E}\left[f'\left(\widehat{\boldsymbol{X}}\right)\right] \mathbb{E}\left[g'\left(\widehat{\boldsymbol{X}}\right)\right] = \mathbb{E}\left[f(\boldsymbol{X})\right] \mathbb{E}\left[g(\boldsymbol{X})\right],$$

as desired. ∎

Lemma A.4 (Differentiation under the Integral Sign) *Consider a function $F : \mathbb{R}^n \times \mathbb{R} \mapsto \mathbb{R}$ such that $\frac{\partial F(\boldsymbol{x},s)}{\partial s}$ is well defined and measurable over $\mathcal{U} \times (0, t_0)$ for some open subset $\mathcal{U} \subset \mathbb{R}^n$ and some $t_0 > 0$. For any probability measure μ on $\mathbb{R}^n$ and any $t \in (0, t_0)$ such that $\int_0^t \int_{\mathcal{U}} \left| \frac{\partial F(\boldsymbol{x},s)}{\partial s} \right| \mu\left(d\boldsymbol{x}\right) ds < \infty$, it holds that*

$$\frac{d}{dt} \int_{\mathcal{U}} F\left(\boldsymbol{x}, t\right) \mu\left(d\boldsymbol{x}\right) = \int_{\mathcal{U}} \frac{\partial F\left(\boldsymbol{x}, t\right)}{\partial t} \mu\left(d\boldsymbol{x}\right), \ or \ \frac{d}{dt} \mathbb{E}_{\boldsymbol{x}}\left[F\left(\boldsymbol{x}, t\right) \mathbb{1}_{\mathcal{U}}\right] = \mathbb{E}_{\boldsymbol{x}}\left[\frac{\partial F\left(\boldsymbol{x}, t\right)}{\partial t} \mathbb{1}_{\mathcal{U}}\right]. \quad (A.9)$$

Proof We have

$$\begin{aligned}
\int_{\mathcal{U}} \frac{\partial F\left(\boldsymbol{x}, t\right)}{\partial t} \mu\left(d\boldsymbol{x}\right) &= \frac{d}{dt} \int_0^t \int_{\mathcal{U}} \frac{\partial F\left(\boldsymbol{x}, s\right)}{\partial s} \mu\left(d\boldsymbol{x}\right) ds \\
&= \frac{d}{dt} \int_{\mathcal{U}} \int_0^t \frac{\partial F\left(\boldsymbol{x}, s\right)}{\partial s} \ ds \ \mu\left(d\boldsymbol{x}\right) \\
&= \frac{d}{dt} \int_{\mathcal{U}} \left(F\left(\boldsymbol{x}, t\right) - F\left(\boldsymbol{x}, 0\right)\right) \ \mu\left(d\boldsymbol{x}\right) \\
&= \frac{d}{dt} \int_{\mathcal{U}} F\left(\boldsymbol{x}, t\right) \ \mu\left(d\boldsymbol{x}\right),
\end{aligned}$$

where we have used the fundamental theorem of calculus for the first and third equalities, and measure-theoretic Fubini's theorem (see, e.g., Theorem 2.37 of [Fol99]) for the second equality (as justified by our integrability assumption). ∎

Lemma A.5 (Gaussian Tail Estimates) *Let $X \sim \mathcal{N}\left(0, 1\right)$ and $\Phi\left(x\right)$ be CDF of X. For any $x \geq 0$, we have the following estimates for $\Phi^c\left(x\right) \doteq 1 - \Phi\left(x\right)$:*

$$\left(\frac{1}{x} - \frac{1}{x^3}\right) \frac{\exp\left(-x^2/2\right)}{\sqrt{2\pi}} \leq \Phi^c\left(x\right) \leq \left(\frac{1}{x} - \frac{1}{x^3} + \frac{3}{x^5}\right) \frac{\exp\left(-x^2/2\right)}{\sqrt{2\pi}}, \quad (\textit{Type I}) \quad (A.10)$$

$$\frac{x}{x^2 + 1} \frac{\exp\left(-x^2/2\right)}{\sqrt{2\pi}} \leq \Phi^c\left(x\right) \leq \frac{1}{x} \frac{\exp\left(-x^2/2\right)}{\sqrt{2\pi}}, \quad (\textit{Type II}) \quad (A.11)$$

$$\frac{\sqrt{x^2 + 4} - x}{2} \frac{\exp\left(-x^2/2\right)}{\sqrt{2\pi}} \leq \Phi^c\left(x\right) \leq \left(\sqrt{2 + x^2} - x\right) \frac{\exp\left(-x^2/2\right)}{\sqrt{2\pi}} \quad (\textit{Type III}). \quad (A.12)$$

Proof Type I bounds can be obtained by integration by parts with proper truncations. Type II upper bound can again be obtained via integration by parts, and the lower bound can be obtained via considering the function $f\left(x\right) \doteq \Phi^c\left(x\right) - \frac{x}{x^2 + 1} \frac{\exp\left(-x^2/2\right)}{\sqrt{2\pi}}$ and noticing it is always nonnegative. Type III bounds are mentioned in [Due10] and reproduced by the systematic approach developed therein (section 2). ∎

Lemma A.6 (Moments of the Gaussian Random Variables) *If $X \sim \mathcal{N}\left(0, \sigma^2\right)$, then it holds for all integer $p \geq 1$ that*

$$\mathbb{E}\left[|X|^p\right] = \sigma^p \left(p-1\right)!! \left[\sqrt{\frac{2}{\pi}} \mathbb{1}_{p \text{ odd}} + \mathbb{1}_{p \text{ even}}\right] \leq \sigma^p \left(p-1\right)!!. \tag{A.13}$$

Lemma A.7 (Moments of the χ^2 Random Variables) *If $X \sim \chi^2\left(n\right)$, then it holds for all integer $p \geq 1$,*

$$\mathbb{E}\left[X^p\right] = 2^p \frac{\Gamma\left(p + n/2\right)}{\Gamma\left(n/2\right)} = \prod_{k=1}^{p}\left(n + 2k - 2\right) \leq \frac{p!}{2}\left(2n\right)^p. \tag{A.14}$$

Lemma A.8 (Moments of the χ Random Variables) *If $X \sim \chi\left(n\right)$, then it holds for all integer $p \geq 1$,*

$$\mathbb{E}\left[X^p\right] = 2^{p/2} \frac{\Gamma\left(p/2 + n/2\right)}{\Gamma\left(n/2\right)} \leq p! n^{p/2}. \tag{A.15}$$

Lemma A.9 (Moment-Control Bernstein's Inequality for Scalar RVs, Theorem 2.10 of [FR13]) *Let $X_1, \ldots, X_p$ be i.i.d. real-valued random variables. Suppose that there exist some positive number R and σ^2 such that*

$$\mathbb{E}\left[|X_k|^m\right] \leq \frac{m!}{2}\sigma^2 R^{m-2}, \ \text{for all integers } m \geq 2.$$

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

$$\mathbb{P}\left[|S - \mathbb{E}\left[S\right]| \geq t\right] \leq 2\exp\left(-\frac{pt^2}{2\sigma^2 + 2Rt}\right). \tag{A.16}$$

Lemma A.10 (Moment-Control Bernstein's Inequality for Matrix RVs, Theorem 6.2 of [Tro12]) *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p \in \mathbb{R}^{d \times d}$ be i.i.d. random, symmetric matrices. Suppose there exist some positive number R and σ^2 such that*

$$\mathbb{E}\left[\boldsymbol{X}_k^m\right] \preceq \frac{m!}{2}\sigma^2 R^{m-2}\boldsymbol{I} \text{ and } -\mathbb{E}\left[\boldsymbol{X}_k^m\right] \preceq \frac{m!}{2}\sigma^2 R^{m-2}\boldsymbol{I}, \text{ for all integers } m \geq 2.$$

Let $\boldsymbol{S} \doteq \frac{1}{p}\sum_{k=1}^{p}\boldsymbol{X}_k$, then for all $t > 0$, it holds that

$$\mathbb{P}\left[\|\boldsymbol{S} - \mathbb{E}\left[\boldsymbol{S}\right]\| \geq t\right] \leq 2d\exp\left(-\frac{pt^2}{2\sigma^2 + 2Rt}\right). \tag{A.17}$$

Proving this lemma requires some modification to the original proof of Theorem 6.2 in [Tro12]. We record it here for the sake of completeness.

Proof Let us define $\boldsymbol{S}_p = \sum_{k=1}^{p}\boldsymbol{X}_k$, by Proposition 3.1 of [Tro12], we have

$$\mathbb{P}\left[\lambda_{\max}\left(\boldsymbol{S}_p - \mathbb{E}\left[\boldsymbol{S}_p\right]\right) \geq t\right] \leq \inf_{t>0} e^{-\theta t}\mathbb{E}\left[\operatorname{tr}\exp\left(\theta\boldsymbol{S}_p - \theta\mathbb{E}\left[\boldsymbol{S}_p\right]\right)\right], \tag{A.18}$$

To proceed, notice that

$$\mathbb{E}\left[\operatorname{tr}\exp\left(\theta\boldsymbol{S}_p - \theta\mathbb{E}\left[\boldsymbol{S}_p\right]\right)\right]$$

$$\begin{aligned}
&= \mathbb{E}_{\boldsymbol{S}_{p-1}} \mathbb{E}_{\boldsymbol{X}_p} \left[\operatorname{tr} \exp \left(\theta \left(\boldsymbol{S}_{p-1} - \mathbb{E}\left[\boldsymbol{S}_{p-1}\right] \right) + \theta \boldsymbol{X}_p - \theta \mathbb{E}\left[\boldsymbol{X}_p\right] \right) \right] \\
&\leq \mathbb{E}_{\boldsymbol{S}_{p-1}} \left[\operatorname{tr} \exp \left(\theta (\boldsymbol{S}_{p-1} - \mathbb{E}\left[\boldsymbol{S}_{p-1}\right]) + \log \left(\mathbb{E}\left[e^{\theta \boldsymbol{X}_p} \right] \right) - \theta \mathbb{E}\left[\boldsymbol{X}_p\right] \right) \right] \\
&\leq \mathbb{E}_{\boldsymbol{S}_{p-1}} \left[\operatorname{tr} \exp \left(\theta (\boldsymbol{S}_{p-1} - \mathbb{E}\left[\boldsymbol{S}_{p-1}\right]) + \mathbb{E}\left[e^{\theta \boldsymbol{X}_p} \right] - \boldsymbol{I} - \theta \mathbb{E}\left[\boldsymbol{X}_p\right] \right) \right] \\
&= \mathbb{E}_{\boldsymbol{S}_{p-1}} \left[\operatorname{tr} \exp \left(\theta (\boldsymbol{S}_{p-1} - \mathbb{E}\left[\boldsymbol{S}_{p-1}\right]) + \sum_{\ell=2}^{\infty} \frac{\theta^\ell \mathbb{E}\left[\boldsymbol{X}_k^\ell \right]}{\ell!} \right) \right]
\end{aligned}$$

where at the third line we have used the result of Corollary 3.3 of [Tro12], i.e., $\mathbb{E}\left[\operatorname{tr} \exp\left(\boldsymbol{H} + \boldsymbol{X}\right)\right] \leq \operatorname{tr} \exp\left(\boldsymbol{H} + \log\left(\mathbb{E}\left[e^{\boldsymbol{X}}\right]\right)\right)$ for any fixed $\boldsymbol{H}$ and random, symmetric $\boldsymbol{X}$, at the fourth we have used the fact that $\log \boldsymbol{X} \preceq \boldsymbol{X} - \boldsymbol{I}$ for any $\boldsymbol{X} \succ \boldsymbol{0}$ (as $\log u \leq u - 1$ for any $u > 0$ and transfer rule applies here), and the last line relies on exchange of infinite summation and expectation, justified as $\boldsymbol{X}_p$ has a bounded spectral radius. By repeating the argument backwards for $\boldsymbol{X}_{p-1}, \cdots, \boldsymbol{X}_1$, we get

$$\begin{aligned}
&\mathbb{E}\left[\operatorname{tr} \exp\left(\theta \boldsymbol{S}_p - \theta \mathbb{E}\left[\boldsymbol{S}_p\right]\right)\right] \\
&\leq \operatorname{tr} \exp \left(p \sum_{\ell=2}^{\infty} \frac{\theta^\ell \mathbb{E}\left[\boldsymbol{X}_k^\ell\right]}{\ell!} \right) \leq \operatorname{tr} \exp \left(p \sum_{\ell=2}^{p} \frac{\theta^\ell \sigma^2 R^{\ell-2}}{2} \boldsymbol{I} \right) \\
&\leq d \left\| \exp \left(p \sum_{\ell=2}^{p} \frac{\theta^\ell \sigma^2 R^{\ell-2}}{2} \boldsymbol{I} \right) \right\| \leq d \exp \left(\frac{p \theta^2 \sigma^2}{2(1 - \theta R)} \right),
\end{aligned} \tag{A.19}$$

where we used the fact that $\mathbb{E}\left[\boldsymbol{X}_i^m\right] \preceq \frac{m!}{2} \sigma^2 R^{m-2} \boldsymbol{I}$ in (A.17) and restrict $\theta < \frac{1}{R}$. Combining the results in (A.18) and (A.19), we have

$$\mathbb{P}\left[\lambda_{\max}\left(\boldsymbol{S}_p - \mathbb{E}\left[\boldsymbol{S}_p\right]\right) \geq t\right] \leq d \inf_{\theta < 1/R} \exp \left(\frac{p \theta^2 \sigma^2}{2(1 - \theta R)} - \theta t \right) \tag{A.20}$$

by taking $\theta = t/(p\sigma^2 + Rt) < 1/R$, we obtain

$$\mathbb{P}\left[\lambda_{\max}\left(\boldsymbol{S}_p - \mathbb{E}\left[\boldsymbol{S}_p\right]\right) \geq t\right] \leq d \exp \left(-\frac{t^2}{2p\sigma^2 + 2Rt} \right). \tag{A.21}$$

Considering $\boldsymbol{X}_k' = -\boldsymbol{X}_k$ and repeating the above argument, we can similarly obtain

$$\mathbb{P}\left[\lambda_{\min}\left(\boldsymbol{S}_p - \mathbb{E}\left[\boldsymbol{S}_p\right]\right) \leq -t\right] \leq d \exp \left(-\frac{t^2}{2p\sigma^2 + 2Rt} \right). \tag{A.22}$$

Putting the above bounds together, we have

$$\mathbb{P}\left[\|\boldsymbol{S}_p - \mathbb{E}\left[\boldsymbol{S}_p\right]\| \geq t\right] \leq 2d \exp \left(-\frac{t^2}{2p\sigma^2 + 2Rt} \right). \tag{A.23}$$

We obtain the claimed bound by substituting $\boldsymbol{S}_p = p\boldsymbol{S}$ and simplifying the resulting expressions. ∎

Corollary A.11 (Moment-Control Bernstein's Inequality for Vector RVs) *Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p \in \mathbb{R}^d$ be i.i.d. random vectors. Suppose there exist some positive number R and σ^2 such that*

$$\mathbb{E}\left[\|\boldsymbol{x}_k\|^m\right] \leq \frac{m!}{2} \sigma^2 R^{m-2}, \quad \text{for all integers } m \geq 2.$$

Let $s = \frac{1}{p}\sum_{k=1}^{p} x_k$, then for any $t > 0$, it holds that

$$\mathbb{P}\left[\|s - \mathbb{E}\left[s\right]\| \geq t\right] \leq 2(d+1)\exp\left(-\frac{pt^2}{2\sigma^2 + 2Rt}\right). \tag{A.24}$$

Proof To obtain the result, we apply the matrix Bernstein inequality in Lemma A.10 to a suitable embedding of the random vectors $\{x_k\}_{k=1}^{p}$. For any $k \in [p]$, define the symmetric matrix

$$X_k = \left[\begin{array}{cc} 0 & x_k^* \\ x_k & 0 \end{array}\right] \in \mathbb{R}^{(d+1)\times(d+1)}.$$

Then it holds that

$$X_k^{2\ell+1} = \|x_k\|_2^{2\ell}\left[\begin{array}{cc} 0 & x_k^* \\ x_k & 0 \end{array}\right], \quad X_k^{2\ell+2} = \|x_k\|^{2\ell}\left[\begin{array}{cc} \|x_k\|^2 & 0 \\ 0 & x_k x_k^* \end{array}\right], \quad \text{for all integers } \ell \geq 0.$$

Using the fact that

$$x_k x_k^* \preceq \|x_k\|^2\,I, \quad \|X_k\| = \sqrt{\|X_k^2\|} = \|x_k\| \implies -\|x_k\|\,I \preceq X_k \preceq \|x_k\|\,I,$$

and combining the above expressions for $X_k^{2\ell+1}$ and $X_k^{2\ell+2}$, we obtain

$$\mathbb{E}\left[X_k^m\right], -\mathbb{E}\left[X_k^m\right] \preceq \mathbb{E}\left[\|x_k\|_2^m\right]I \preceq \frac{m!}{2}\sigma^2 R^{m-2}I, \quad \text{for all integers } m \geq 2, \tag{A.25}$$

Let $S = \frac{1}{p}\sum_{k=1}^{p} X_k$, noting that

$$\|S - \mathbb{E}\left[S\right]\| = \|s - \mathbb{E}\left[s\right]\|, \tag{A.26}$$

and applying Lemma A.10, we complete the proof. ∎

Lemma A.12 (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*

$$f(x + ty) = f(x) + t\int_0^1 \langle \nabla f(x + sty), y\rangle\,ds, \tag{A.27}$$

$$f(x + ty) = f(x) + t\langle\nabla f(x), y\rangle + t^2\int_0^1 (1-s)\langle\nabla^2 f(x + sty)y, y\rangle\,ds. \tag{A.28}$$

Proof By the fundamental theorem of calculus, since f is continuous differentiable, it is obvious that

$$f(x + ty) = f(x) + \int_0^t \langle\nabla f(x + \tau y), y\rangle\,d\tau. \tag{A.29}$$

If f is twice continuously differentiable, by using integral by parts, we obtain

$$f(x + ty) = f(x) + [(\tau - t)\langle\nabla f(x + \tau y), y\rangle]|_0^t - \int_0^t (\tau - t)\,d\langle\nabla f(x + \tau y), y\rangle$$

$$= f(x) + t\langle\nabla f(x + \tau y), y\rangle + \int_0^t (t - \tau)\langle\nabla^2 f(x + \tau y)y, y\rangle\,d\tau. \tag{A.30}$$

By a change of variable $\tau = st$ ($0 \leq s \leq 1$) for (A.29) and (A.30), we get the desired results. ∎

B Auxillary Results for Proofs

Lemma B.1 *Let $X \sim \mathcal{N}(0, \sigma_X^2)$ and $Y \sim \mathcal{N}(0, \sigma_Y^2)$ be independent random variables and $\Phi^c(t) = \frac{1}{\sqrt{2\pi}} \int_t^\infty \exp\left(-x^2/2\right) dx$ be the complementary cumulative distribution function of the standard normal. For any $a > 0$, we have*

$$\mathbb{E}\left[X\mathbb{1}_{X>0}\right] = \frac{\sigma_X}{\sqrt{2\pi}}, \tag{B.1}$$

$$\mathbb{E}\left[\exp\left(-aX\right)X\mathbb{1}_{X>0}\right] = \frac{\sigma_X}{\sqrt{2\pi}} - a\sigma_X^2 \exp\left(\frac{a^2\sigma_X^2}{2}\right)\Phi^c\left(a\sigma_X\right), \tag{B.2}$$

$$\mathbb{E}\left[\exp\left(-aX\right)\mathbb{1}_{X>0}\right] = \exp\left(\frac{a^2\sigma_X^2}{2}\right)\Phi^c\left(a\sigma_X\right), \tag{B.3}$$

$$\mathbb{E}\left[\exp\left(-a(X+Y)\right)X^2\mathbb{1}_{X+Y>0}\right] = \sigma_X^2\left(1+a^2\sigma_X^2\right)\exp\left(\frac{a^2\sigma_X^2 + a^2\sigma_Y^2}{2}\right)\Phi^c\left(a\sqrt{\sigma_X^2 + \sigma_Y^2}\right)$$
$$- \frac{a\sigma_X^4}{\sqrt{2\pi}\sqrt{\sigma_X^2 + \sigma_Y^2}}, \tag{B.4}$$

$$\mathbb{E}\left[\exp\left(-a(X+Y)\right)XY\mathbb{1}_{X+Y>0}\right] = a^2\sigma_X^2\sigma_Y^2 \exp\left(\frac{a^2\sigma_X^2 + a^2\sigma_Y^2}{2}\right)\Phi^c\left(a\sqrt{\sigma_X^2 + \sigma_Y^2}\right)$$
$$- \frac{a\sigma_X^2\sigma_Y^2}{\sqrt{2\pi}\sqrt{\sigma_X^2 + \sigma_Y^2}}, \tag{B.5}$$

$$\mathbb{E}\left[\tanh\left(aX\right)X\right] = a\sigma_X^2\mathbb{E}\left[1 - \tanh^2\left(aX\right)\right], \tag{B.6}$$

$$\mathbb{E}\left[\tanh\left(a(X+Y)\right)X\right] = a\sigma_X^2\mathbb{E}\left[1 - \tanh^2\left(a(X+Y)\right)\right]. \tag{B.7}$$

Proof Equalities (B.1), (B.2), (B.3), (B.4) and (B.5) can be obtained by direct integrations. Equalities (B.6) and (B.7) can be derived using integration by part. ∎

Proof [of Lemma 7.1] Indeed $\frac{1}{(1+\beta t)^2} = \sum_{k=0}^\infty (-1)^k(k+1)\beta^k t^k$, as

$$\sum_{k=0}^\infty (-1)^k(k+1)\beta^k t^k = \sum_{k=0}^\infty (-\beta t)^k + \sum_{k=0}^\infty k(-\beta t)^k = \frac{1}{1+\beta t} + \frac{-\beta t}{(1+\beta t)^2} = \frac{1}{(1+\beta t)^2}.$$

The magnitude of the coefficient vector is

$$\|\boldsymbol{b}\|_{\ell^1} = \sum_{k=0}^\infty \beta^k(1+k) = \sum_{k=0}^\infty \beta^k + \sum_{k=0}^\infty k\beta^k = \frac{1}{1-\beta} + \frac{\beta}{(1-\beta)^2} = \frac{1}{(1-\beta)^2} = T.$$

Observing that $\frac{1}{(1+\beta t)^2} > \frac{1}{(1+t)^2}$ for $t \in [0,1]$ when $0 < \beta < 1$, we obtain

$$\|p - f\|_{L^1[0,1]} = \int_0^1 |p(t) - f(t)| \, dt = \int_0^1 \left[\frac{1}{(1+\beta t)^2} - \frac{1}{(1+t)^2}\right] dt = \frac{1-\beta}{2(1+\beta)} \leq \frac{1}{2\sqrt{T}}. \tag{B.8}$$

Moreover, we have

$$\|f - p\|_{L^\infty[0,1]} = \max_{t\in[0,1]} p(t) - f(t) = \max_{t\in[0,1]} \frac{t(1-\beta)(2+t(1+\beta))}{(1+t)^2(1+\beta t)^2} \leq 1 - \beta = \frac{1}{\sqrt{T}}. \tag{B.9}$$

Finally, notice that

$$\sum_{k=0}^{\infty} \frac{b_k}{(1+k)^3} = \sum_{k=0}^{\infty} \frac{(-\beta)^k}{(1+k)^2} = \sum_{i=0}^{\infty} \left[\frac{\beta^{2i}}{(1+2i)^2} - \frac{\beta^{2i+1}}{(2i+2)^2} \right]$$

$$= \sum_{i=0}^{\infty} \beta^{2i} \frac{(2i+2)^2 - \beta(2i+1)^2}{(2i+2)^2(2i+1)^2} > 0, \qquad (\text{B.10})$$

where at the second equality we have grouped consecutive even-odd pair of summands. In addition, we have

$$\sum_{k=0}^{n} \frac{b_k}{(1+k)^3} \leq \sum_{k=0}^{n} \frac{|b_k|}{(1+k)^3} = \sum_{k=0}^{n} \frac{\beta^k}{(1+k)^2} \leq 1 + \sum_{k=1}^{n} \frac{1}{(1+k)k} = 2 - \frac{1}{n+1}, \qquad (\text{B.11})$$

which converges to 2 when $n \to \infty$, completing the proof. ∎

Proof [of Lemma 7.4] The first inequality is obviously true for $v = 0$. When $v \neq 0$, we have

$$\mathbb{E}\left[|v^* z|^m\right] = \sum_{\ell=0}^{n} \theta^\ell (1-\theta)^{n-\ell} \sum_{\mathcal{J} \in \binom{[n]}{\ell}} \mathbb{E}_{Z \sim \mathcal{N}(0, \|v_{\mathcal{J}}\|^2)} \left[|Z|^m\right]$$

$$\leq \sum_{\ell=0}^{n} \theta^\ell (1-\theta)^{n-\ell} \sum_{\mathcal{J} \in \binom{[n]}{\ell}} \mathbb{E}_{Z \sim \mathcal{N}(0, \|v\|^2)} \left[|Z|^m\right]$$

$$= \mathbb{E}_{Z \sim \mathcal{N}(0, \|v\|^2)} \left[|Z|^m\right] \sum_{\ell=0}^{n} \theta^\ell (1-\theta)^{n-\ell} \binom{n}{\ell}$$

$$= \mathbb{E}_{Z \sim \mathcal{N}(0, \|v\|^2)} \left[|Z|^m\right],$$

where the second line relies on the fact $\|v_{\mathcal{J}}\| \leq \|v\|$ and that for a fixed order, central moment of Gaussian is monotonically increasing w.r.t. its variance. Similarly, to see the second inequality,

$$\mathbb{E}\left[\|z\|^m\right] = \sum_{\ell=0}^{n} \theta^\ell (1-\theta)^{n-\ell} \sum_{\mathcal{J} \in \binom{[n]}{\ell}} \mathbb{E}\left[\|z'_{\mathcal{J}}\|^m\right]$$

$$\leq \mathbb{E}\left[\|z'\|^m\right] \sum_{\ell=0}^{n} \theta^\ell (1-\theta)^{n-\ell} \binom{n}{\ell} = \mathbb{E}\left[\|z'\|^m\right],$$

as desired. ∎

Proof [of Lemma 7.11] Consider one component of X, i.e., $X_{ij} = B_{ij} V_{ij}$ for $i \in [n]$ and $j \in [p]$, where $B_{ij} \sim \mathrm{Ber}(\theta)$) and $V_{ij} \sim \mathcal{N}(0, 1)$. We have

$$\mathbb{P}\left[|X_{ij}| > 4\sqrt{\log(np)}\right] \leq \theta \mathbb{P}\left[|V_{ij}| > 4\sqrt{\log(np)}\right] \leq \theta \exp\left(-8 \log(np)\right) = \theta(np)^{-8}.$$

And also

$$\mathbb{P}\left[|X_{ij}| < 1\right] = 1 - \theta + \theta \mathbb{P}\left[|V_{ij}| < 1\right] \leq 1 - 0.3\theta.$$

Applying a union bound as

$$\mathbb{P}\left[\|\boldsymbol{X}\|_{\infty} \le 1 \text{ or } \|\boldsymbol{X}\|_{\infty} \ge 4\sqrt{\log(np)}\right] \le (1 - 0.3\theta)^{np} + np\theta(np)^{-8} \le \exp(-0.3\theta np) + \theta(np)^{-7},$$

we complete the proof. $\blacksquare$

Lemma B.2 *Suppose $\boldsymbol{A} \succ \boldsymbol{0}$. Then for any symmetric perturbation matrix $\boldsymbol{\Delta}$ with $\|\boldsymbol{\Delta}\| \le \frac{\sigma_{\min}(\boldsymbol{A})}{2}$, it holds that*

$$\left\|(\boldsymbol{A} + \boldsymbol{\Delta})^{-1/2} - \boldsymbol{A}^{-1/2}\right\| \le \frac{2\|\boldsymbol{A}\|^{1/2}\|\boldsymbol{\Delta}\|}{\sigma_{\min}^2(\boldsymbol{A})}. \tag{B.12}$$

Proof First note that

$$\left\|(\boldsymbol{A} + \boldsymbol{\Delta})^{-1/2} - \boldsymbol{A}^{-1/2}\right\| \le \frac{\left\|(\boldsymbol{A} + \boldsymbol{\Delta})^{-1} - \boldsymbol{A}^{-1}\right\|}{\sigma_{\min}^{1/2}(\boldsymbol{A}^{-1})}$$

as by our assumption $\boldsymbol{A} + \boldsymbol{\Delta} \succ \boldsymbol{0}$ and the fact (Theorem 6.2 in [Hig08]) that $\left\|\boldsymbol{X}^{1/2} - \boldsymbol{Y}^{1/2}\right\| \le \|\boldsymbol{X} - \boldsymbol{Y}\| / \left(\sigma_{\min}^{1/2}(\boldsymbol{X}) + \sigma_{\min}^{1/2}(\boldsymbol{Y})\right)$ for any $\boldsymbol{X}, \boldsymbol{Y} \succ \boldsymbol{0}$ applies. Moreover, using the fact

$$\left\|(\boldsymbol{X} + \boldsymbol{\Delta})^{-1} - \boldsymbol{X}^{-1}\right\| \le \frac{\|\boldsymbol{X}^{-1}\|\|\boldsymbol{X}^{-1}\boldsymbol{\Delta}\|}{1 - \|\boldsymbol{X}^{-1}\boldsymbol{\Delta}\|} \le \frac{\|\boldsymbol{\Delta}\|\|\boldsymbol{X}^{-1}\|^2}{1 - \|\boldsymbol{X}^{-1}\|\|\boldsymbol{\Delta}\|}$$

for nonsingular $\boldsymbol{X}$ and perturbation $\boldsymbol{\Delta}$ with $\|\boldsymbol{X}^{-1}\|\|\boldsymbol{\Delta}\| < 1$ (see, e.g., Theorem 2.5 of Chapter III in [SS90]), we obtain

$$\frac{1}{\sigma_{\min}^{1/2}(\boldsymbol{A}^{-1})}\left\|(\boldsymbol{A} + \boldsymbol{\Delta})^{-1} - \boldsymbol{A}^{-1}\right\| \le \|\boldsymbol{A}\|^{1/2}\frac{\|\boldsymbol{\Delta}\|\|\boldsymbol{A}^{-1}\|^2}{1 - \|\boldsymbol{A}^{-1}\|\|\boldsymbol{\Delta}\|} \le \frac{2\|\boldsymbol{A}\|^{1/2}\|\boldsymbol{\Delta}\|}{\sigma_{\min}^2(\boldsymbol{A})},$$

where we have used the fact $\|\boldsymbol{A}^{-1}\|\|\boldsymbol{\Delta}\| \le 1/2$ to simplify at the last inequality. $\blacksquare$

Lemma B.3 *There exists a positive constant C such that for any $\theta \in (0, 1/2)$ and $n_2 > Cn_1^2 \log n_1$, the random matrix $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$ with $\boldsymbol{X} \sim_{i.i.d.} \text{BG}(\theta)$ obeys*

$$\left\|\frac{1}{n_2\theta}\boldsymbol{X}\boldsymbol{X}^* - \boldsymbol{I}\right\| \le 10\sqrt{\frac{\theta n_1 \log n_2}{n_2}} \tag{B.13}$$

with probability at least $1 - n_2^{-8}$.

Proof Observe that $\mathbb{E}\left[\frac{1}{\theta}\boldsymbol{x}_k\boldsymbol{x}_k^*\right] = \boldsymbol{I}$ for any column $\boldsymbol{x}_k$ of $\boldsymbol{X}$ and so $\frac{1}{n_2\theta}\boldsymbol{X}\boldsymbol{X}^*$ can be considered as a normalize sum of independent random matrices. Moreover, for any integer $m \ge 2$,

$$\mathbb{E}\left[\left(\frac{1}{\theta}\boldsymbol{x}_k\boldsymbol{x}_k^*\right)^m\right] = \frac{1}{\theta^m}\mathbb{E}\left[\|\boldsymbol{x}_k\|^{2m-2}\boldsymbol{x}_k\boldsymbol{x}_k^*\right].$$

Now $\mathbb{E}\left[\|\boldsymbol{x}_k\|^{2m-2}\boldsymbol{x}_k\boldsymbol{x}_k^*\right]$ is a diagonal matrix (as $\mathbb{E}\left[\|\boldsymbol{x}_k\|^2 x_k(i) x_k(j)\right] = -\mathbb{E}\left[\|\boldsymbol{x}_k\|^2 x_k(i) x_k(j)\right]$ for any $i \neq j$ by symmetry of the distribution) in the form $\mathbb{E}\left[\|\boldsymbol{x}_k\|^{2m-2}\boldsymbol{x}_k\boldsymbol{x}_k^*\right] = \mathbb{E}\left[\|\boldsymbol{x}\|^{2m-2} x(1)^2\right]\boldsymbol{I}$ for $\boldsymbol{x} \sim_{i.i.d.} \mathrm{BG}\left(\theta\right)$ with $\boldsymbol{x} \in \mathbb{R}^{n_1}$. Let $t^2\left(\boldsymbol{x}\right) = \|\boldsymbol{x}\|^2 - x(1)^2$. Then if $m = 2$,

$$
\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{x}\|^2 x(1)^2\right] &= \mathbb{E}\left[x(1)^4\right] + \mathbb{E}\left[t^2\left(\boldsymbol{x}\right)\right]\mathbb{E}\left[x(1)^2\right] \\
&= \mathbb{E}\left[x(1)^4\right] + (n_1 - 1)\left(\mathbb{E}\left[x(1)^2\right]\right)^2 = 3\theta + (n_1 - 1)\theta^2 \leq 3n_1\theta,
\end{aligned}
$$

where for the last simplification we use the assumption $\theta \leq 1/2$. For $m \geq 3$,

$$
\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{x}\|^{2m-2} x(1)^2\right] &= \sum_{k=0}^{m-1}\binom{m-1}{k}\mathbb{E}\left[t^{2k}\left(\boldsymbol{x}\right) x(1)^{2m-2k}\right] = \sum_{k=0}^{m-1}\binom{m-1}{k}\mathbb{E}\left[t^{2k}\left(\boldsymbol{x}\right)\right]\mathbb{E}\left[x(1)^{2m-2k}\right] \\
&\leq \sum_{k=0}^{m-1}\binom{m-1}{k}\mathbb{E}_{Z\sim\chi^2(n_1-1)}\left[Z^k\right]\theta\mathbb{E}_{W\sim\mathcal{N}(0,1)}\left[W^{2m-2k}\right] \\
&\leq \theta\sum_{k=0}^{m-1}\binom{m-1}{k}\frac{k!}{2}(2n_1 - 2)^k (2m - 2k)!! \\
&\leq \theta 2^m\frac{m!}{2}\sum_{k=0}^{m-1}\binom{m-1}{k}(n_1 - 1)^k \\
&\leq \frac{m!}{2}n_1^{m-1}2^{m-1},
\end{aligned}
$$

where we have used the moment estimates for Gaussian and χ^2 random variables from Lemma A.6 and Lemma A.7, and also $\theta \leq 1/2$. Taking $\sigma^2 = 3n_1\theta$ and $R = 2n_1$, and invoking the matrix Bernstein in Lemma A.10, we obtain

$$
\mathbb{E}\left[\left\|\frac{1}{p\theta}\sum_{k=1}^{p}\boldsymbol{x}_k\boldsymbol{x}_k^* - \boldsymbol{I}\right\| > t\right] \leq \exp\left(-\frac{n_2 t^2}{6n_1\theta + 4n_1 t} + 2\log n_1\right) \tag{B.14}
$$

for any $t \geq 0$. Taking $t = 10\sqrt{\theta n_1 \log\left(n_2\right)/n_2}$ gives the claimed result. ∎

Lemma B.4 *Consider two linear subspaces $\mathcal{U}, \mathcal{V}$ of dimension k in $\mathbb{R}^n$ ($k \in [n]$) spanned by orthonormal bases $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively. Suppose $\pi/2 \geq \theta_1 \geq \theta_2 \cdots \geq \theta_k \geq 0$ are the principal angles between $\mathcal{U}$ and $\mathcal{V}$. Then it holds that*
i) $\min_{\boldsymbol{Q}\in O_k}\|\boldsymbol{U} - \boldsymbol{V}\boldsymbol{Q}\| \leq \sqrt{2 - 2\cos\theta_1}$;
ii) $\sin\theta_1 = \|\boldsymbol{U}\boldsymbol{U}^* - \boldsymbol{V}\boldsymbol{V}^*\|$;
iii) Let $\mathcal{U}^\perp$ and $\mathcal{V}^\perp$ be the orthogonal complement of $\mathcal{U}$ and $\mathcal{V}$, respectively. Then $\theta_1(\mathcal{U},\mathcal{V}) = \theta_1(\mathcal{U}^\perp,\mathcal{V}^\perp)$.

Proof Proof to i) is similar to that of II. Theorem 4.11 in [SS90]. For $2k \leq n$, w.l.o.g., we can assume $\boldsymbol{U}$ and $\boldsymbol{V}$ are the canonical bases for $\mathcal{U}$ and $\mathcal{V}$, respectively. Then

$$
\min_{\boldsymbol{Q}\in O_k}\left\|\begin{bmatrix}\boldsymbol{I} - \boldsymbol{\Gamma}\boldsymbol{Q} \\ -\boldsymbol{\Sigma}\boldsymbol{Q} \\ \boldsymbol{0}\end{bmatrix}\right\| \leq \left\|\begin{bmatrix}\boldsymbol{I} - \boldsymbol{\Gamma} \\ -\boldsymbol{\Sigma} \\ \boldsymbol{0}\end{bmatrix}\right\| \leq \left\|\begin{bmatrix}\boldsymbol{I} - \boldsymbol{\Gamma} \\ -\boldsymbol{\Sigma}\end{bmatrix}\right\|.
$$

Now by definition

$$\left\| \begin{bmatrix} \boldsymbol{I} - \boldsymbol{\Gamma} \\ -\boldsymbol{\Sigma} \end{bmatrix} \right\|^2 = \max_{\|\boldsymbol{x}\|=1} \left\| \begin{bmatrix} \boldsymbol{I} - \boldsymbol{\Gamma} \\ -\boldsymbol{\Sigma} \end{bmatrix} \boldsymbol{x} \right\|^2 = \max_{\|\boldsymbol{x}\|=1} \sum_{i=1}^{k} (1 - \cos\theta_i)^2 x_i^2 + \sin^2\theta_i x_i^2$$

$$= \max_{\|\boldsymbol{x}\|=1} \sum_{i=1}^{k} (2 - 2\cos\theta_i) x_i^2 \leq 2 - 2\cos\theta_1.$$

Note that the upper bound is achieved by taking $\boldsymbol{x} = \boldsymbol{e}_1$. When $2k > n$, by the results from CS decomposition (see, e.g., I Theorem 5.2 of [SS90]).

$$\min_{\boldsymbol{Q} \in O_k} \left\| \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\Gamma} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{bmatrix} \right\| \leq \left\| \begin{bmatrix} \boldsymbol{I} - \boldsymbol{\Gamma} \\ -\boldsymbol{\Sigma} \end{bmatrix} \right\|,$$

and the same argument then carries through. To prove ii), note the fact that $\sin\theta_1 = \|\boldsymbol{U}\boldsymbol{U}^* - \boldsymbol{V}\boldsymbol{V}^*\|$ (see, e.g., Theorem 4.5 and Corollary 4.6 of [SS90]). Obviously one also has

$$\sin\theta_1 = \|\boldsymbol{U}\boldsymbol{U}^* - \boldsymbol{V}\boldsymbol{V}^*\| = \|(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^*) - (\boldsymbol{I} - \boldsymbol{V}\boldsymbol{V}^*)\|,$$

while $\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^*$ and $\boldsymbol{I} - \boldsymbol{V}\boldsymbol{V}^*$ are projectors onto $\mathcal{U}^\perp$ and $\mathcal{V}^\perp$, respectively. This completes the proof. ∎

References

[AAJ+13] Alekh Agarwal, Animashree Anandkumar, Prateek Jain, Praneeth Netrapalli, and Rashish Tandon, *Learning sparsely used overcomplete dictionaries via alternating minimization*, arXiv preprint arXiv:1310.7991 (2013).

[AAN13] Alekh Agarwal, Animashree Anandkumar, and Praneeth Netrapalli, *Exact recovery of sparsely used overcomplete dictionaries*, arXiv preprint arXiv:1309.1952 (2013).

[ABG07] P-A Absil, Christopher G Baker, and Kyle A Gallivan, *Trust-region methods on riemannian manifolds*, Foundations of Computational Mathematics **7** (2007), no. 3, 303–330.

[ABGM13] Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma, *Provable bounds for learning some deep representations*, arXiv preprint arXiv:1310.6343 (2013).

[ABGM14] ———, *More algorithms for provable dictionary learning*, arXiv preprint arXiv:1401.0579 (2014).

[ABRS10] Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran, *Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the kurdyka-lojasiewicz inequality*, Mathematics of Operations Research **35** (2010), no. 2, 438–457.

[AEB06] Michal Aharon, Michael Elad, and Alfred M Bruckstein, *On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them*, Linear algebra and its applications **416** (2006), no. 1, 48–67.

[AGJ14] Animashree Anandkumar, Rong Ge, and Majid Janzamin, *Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates*, arXiv preprint arXiv:1402.5180 (2014).

[AGKM12] Sanjeev Arora, Rong Ge, Ravindran Kannan, and Ankur Moitra, *Computing a nonnegative matrix factorization–provably*, Proceedings of the forty-fourth annual ACM symposium on Theory of computing, ACM, 2012, pp. 145–162.

[AGM13]   Sanjeev Arora, Rong Ge, and Ankur Moitra, *New algorithms for learning incoherent and overcomplete dictionaries*, arXiv preprint arXiv:1308.6273 (2013).

[AGMM15]   Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra, *Simple, efficient, and neural algorithms for sparse coding*, arXiv preprint arXiv:1503.00778 (2015).

[AGMS12]   Sanjeev Arora, Rong Ge, Ankur Moitra, and Sushant Sachdeva, *Provable ica with unknown gaussian noise, with implications for gaussian mixtures and autoencoders*, Advances in Neural Information Processing Systems, 2012, pp. 2375–2383.

[ALMT14]   Dennis Amelunxen, Martin Lotz, Michael B McCoy, and Joel A Tropp, *Living on the edge: Phase transitions in convex programs with random data*, Information and Inference (2014), iau005.

[AMS09]   Pierre-Antoine Absil, Robert Mahoney, and Rodolphe Sepulchre, *Optimization algorithms on matrix manifolds*, Princeton University Press, 2009.

[ARR14]   Ali Ahmed, Benjamin Recht, and Justin Romberg, *Blind deconvolution using convex programming*, Information Theory, IEEE Transactions on **60** (2014), no. 3, 1711–1732.

[BCJ13]   Chenglong Bao, Jian-Feng Cai, and Hui Ji, *Fast sparsity-based orthogonal dictionary learning for image restoration*, Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE, 2013, pp. 3384–3391.

[BJQS14]   Chenglong Bao, Hui Ji, Yuhui Quan, and Zuowei Shen, *L0 norm based dictionary learning by proximal methods with global convergence*, Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE, 2014, pp. 3858–3865.

[BJS14]   Chenglong Bao, Hui Ji, and Zuowei Shen, *Convergence analysis for iterative data-driven tight frame construction scheme*, Applied and Computational Harmonic Analysis (2014).

[BKS13]   Afonso S Bandeira, Christopher Kennedy, and Amit Singer, *Approximating the little grothendieck problem over the orthogonal and unitary groups*, arXiv preprint arXiv:1308.5207 (2013).

[BKS14]   Boaz Barak, Jonathan A Kelner, and David Steurer, *Dictionary learning and tensor decomposition via the sum-of-squares method*, arXiv preprint arXiv:1407.1543 (2014).

[BLM13]   Stéphane Boucheron, Gábor Lugosi, and Pascal Massart, *Concentration inequalities: A nonasymptotic theory of independence*, Oxford University Press, 2013.

[BQJ14]   Chenglong Bao, Yuhui Quan, and Hui Ji, *A convergent incoherent dictionary learning algorithm for sparse coding*, Computer Vision–ECCV 2014, Springer, 2014, pp. 302–316.

[BR14]   Jop Briët and Oded Regev, *Tight hardness of the non-commutative grothendieck problem*, arXiv preprint arXiv:1412.4413 (2014).

[BST14]   Jérôme Bolte, Shoham Sabach, and Marc Teboulle, *Proximal alternating linearized minimization for nonconvex and nonsmooth problems*, Mathematical Programming **146** (2014), no. 1-2, 459–494.

[BT89]   Dimitri P Bertsekas and John N Tsitsiklis, *Parallel and distributed computation: numerical methods*, vol. 23, Prentice hall Englewood Cliffs, NJ, 1989.

[BV04]   Stephen Boyd and Lieven Vandenberghe, *Convex optimization*, Cambridge University Press, New York, NY, USA, 2004.

[BWY14]   Sivaraman Balakrishnan, Martin J Wainwright, and Bin Yu, *Statistical guarantees for the em algorithm: From population to sample-based analysis*, arXiv preprint arXiv:1408.2156 (2014).

[Can02]   Emmanuel J. Candès, *New ties between computational harmonic analysis and approximation theory*, Approximation Theory X (2002), 87–153.

[Can14]   Emmanuel J. Candès, *Mathematics of sparsity (and few other things)*, Proceedings of the International Congress of Mathematicians, Seoul, South Korea, 2014.

[CGT00a] Andrew R. Conn, Nicholas I. M. Gould, and Philippe L. Toint, *Trust-region methods*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2000.

[CGT00b] Andrew R Conn, Nicholas IM Gould, and Ph L Toint, *Trust region methods*, vol. 1, Siam, 2000.

[CLMW11] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright, *Robust principal component analysis?*, Journal of the ACM (JACM) **58** (2011), no. 3, 11.

[CLS14] Emmanuel Candès, Xiaodong Li, and Mahdi Soltanolkotabi, *Phase retrieval via wirtinger flow: Theory and algorithms*, arXiv preprint arXiv:1407.1065 (2014).

[CM14] Sunav Choudhary and Urbashi Mitra, *Identifiability scaling laws in bilinear inverse problems*, arXiv preprint arXiv:1402.2637 (2014).

[Com94] Pierre Comon, *Independent component analysis, a new concept?*, Signal processing **36** (1994), no. 3, 287–314.

[CRPW12] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky, *The convex geometry of linear inverse problems*, Foundations of Computational mathematics **12** (2012), no. 6, 805–849.

[CSV13] Emmanuel J Candès, Thomas Strohmer, and Vladislav Voroninski, *Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming*, Communications on Pure and Applied Mathematics **66** (2013), no. 8, 1241–1274.

[DeV98] Ronald A. DeVore, *Nonlinear approximation*, Acta numerica **7** (1998), 51–150.

[DeV09] Ronald A DeVore, *Nonlinear approximation and its applications*, Multiscale, Nonlinear and Adaptive Approximation, Springer, 2009, pp. 169–201.

[DGM13] David L Donoho, Matan Gavish, and Andrea Montanari, *The phase transition of matrix recovery from gaussian measurements matches the minimax mse of matrix denoising*, Proceedings of the National Academy of Sciences **110** (2013), no. 21, 8405–8410.

[DH14] Laurent Demanet and Paul Hand, *Scaling law for recovering the sparsest element in a subspace*, Information and Inference **3** (2014), no. 4, 295–309.

[DT09] David Donoho and Jared Tanner, *Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing*, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **367** (2009), no. 1906, 4273–4293.

[Due10] Lutz Duembgen, *Bounding standard gaussian tail probabilities*, arXiv preprint arXiv:1012.2063 (2010).

[DVDD98] David L. Donoho, Martin Vetterli, Ronald A. DeVore, and Ingrid Daubechies, *Data compression and harmonic analysis*, Information Theory, IEEE Transactions on **44** (1998), no. 6, 2435–2476.

[EAS98] Alan Edelman, Tomás A Arias, and Steven T Smith, *The geometry of algorithms with orthogonality constraints*, SIAM journal on Matrix Analysis and Applications **20** (1998), no. 2, 303–353.

[Ela10] Michael Elad, *Sparse and redundant representations: from theory to applications in signal and image processing*, Springer, 2010.

[FJK96] Alan Frieze, Mark Jerrum, and Ravi Kannan, *Learning linear transformations*, 2013 IEEE 54th Annual Symposium on Foundations of Computer Science, IEEE Computer Society, 1996, pp. 359–359.

[Fol99] Gerald B Folland, *Real analysis: Modern techniques and their applications*, 2nd ed., John Wiley & Sons, 1999.

[FR13]   Simon Foucart and Holger Rauhut, *A mathematical introduction to compressive sensing*, Springer, 2013.

[FW04]   Charles Fortin and Henry Wolkowicz, *The trust region subproblem and semidefinite programming**, Optimization methods and software **19** (2004), no. 1, 41–67.

[GHJY15]   Rong Ge, Furong Huang, Chi Jin, and Yang Yuan, *Escaping from saddle points—online stochastic gradient for tensor decomposition*, arXiv preprint arXiv:1503.02101 (2015).

[GJB⁺13]   Remi Gribonval, Rodolphe Jenatton, Francis Bach, Martin Kleinsteuber, and Matthias Seibert, *Sample complexity of dictionary learning and other matrix factorizations*, arXiv preprint arXiv:1312.3790 (2013).

[GJB14]   Rémi Gribonval, Rodolphe Jenatton, and Francis Bach, *Sparse and spurious: dictionary learning with noise and outliers*, arXiv preprint arXiv:1407.5155 (2014).

[GN10]   Lee-Ad Gottlieb and Tyler Neylon, *Matrix sparsification and the sparse null space problem*, Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, Springer, 2010, pp. 205–218.

[GS10]   Rémi Gribonval and Karin Schnass, *Dictionary identification - sparse matrix-factorization via ℓ^1-minimization*, IEEE Transactions on Information Theory **56** (2010), no. 7, 3523–3539.

[GW11]   Quan Geng and John Wright, *On the local correctness of ℓ^1-minimization for dictionary learning*, *Submitted to* IEEE Transactions on Information Theory (2011), Preprint: http://www.columbia.edu/~jw2966.

[Har60]   Theodore E. Harris, *A lower bound for the critical probability in a certain percolation process*, Mathematical Proceedings of the Cambridge Philosophical Society, vol. 56, Cambridge Univ Press, 1960, pp. 13–20.

[Har14]   Moritz Hardt, *Understanding alternating minimization for matrix completion*, Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on, IEEE, 2014, pp. 651–660.

[Hig08]   Nicholas J. Higham, *Functions of matrices*, Society for Industrial and Applied Mathematics, 2008.

[HK14]   Elad Hazan and Tomer Koren, *A linear-time algorithm for trust region problems*, arXiv preprint arXiv:1401.6757 (2014).

[HMG94]   Uwe Helmke, John B Moore, and Würzburg Germany, *Optimization and dynamical systems*.

[HO00]   Aapo Hyvärinen and Erkki Oja, *Independent component analysis: algorithms and applications*, Neural networks **13** (2000), no. 4, 411–430.

[HO01]   Karhunen J. Hyvärinen, A. and E. Oja, *Independent component analysis*, John Wiley and Sons., 2001.

[HS11]   Christopher Hillar and Friedrich T Sommer, *When can dictionary learning uniquely recover sparse data from subsamples?*, arXiv preprint arXiv:1106.3616 (2011).

[HW14]   Moritz Hardt and Mary Wootters, *Fast matrix completion without the condition number*, Proceedings of The 27th Conference on Learning Theory, 2014, pp. 638–678.

[Hyv99]   Aapo Hyvarinen, *Fast and robust fixed-point algorithms for independent component analysis*, IEEE Trans. Neural Networks **10** (1999), no. 3, 626–634.

[JN14]   Prateek Jain and Praneeth Netrapalli, *Fast exact matrix completion with finite samples*, arXiv preprint arXiv:1411.1087 (2014).

[JNS13]   Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi, *Low-rank matrix completion using alternating minimization*, Proceedings of the forty-fifth annual ACM symposium on Theory of Computing, ACM, 2013, pp. 665–674.

[JO14]   Prateek Jain and Sewoong Oh, *Provable tensor factorization with missing data*, Advances in Neural Information Processing Systems, 2014, pp. 1431–1439.

[KMO10]   Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh, *Matrix completion from a few entries*, Information Theory, IEEE Transactions on **56** (2010), no. 6, 2980–2998.

[LGBB05]   Sylvain Lesage, Rémi Gribonval, Frédéric Bimbot, and Laurent Benaroya, *Learning unions of orthonormal bases with thresholded singular value decomposition*, Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 5, IEEE, 2005, pp. v–293.

[Loh15]   Po-Ling Loh, *Statistical consistency and asymptotic normality for high-dimensional robust m-estimators*, arXiv preprint arXiv:1501.00312 (2015).

[LSSS14]   Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir, *On the computational efficiency of training neural networks*, Advances in Neural Information Processing Systems, 2014, pp. 855–863.

[LV15]   Kyle Luh and Van Vu, *Dictionary learning with few samples and matrix concentration*, arXiv preprint arXiv:1503.08854 (2015).

[LW11]   Po-Ling Loh and Martin J Wainwright, *High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity*, Advances in Neural Information Processing Systems, 2011, pp. 2726–2734.

[LW13]   ———, *Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima*, Advances in Neural Information Processing Systems, 2013, pp. 476–484.

[LW14]   ———, *Support recovery without incoherence: A case for nonconvex regularization*, arXiv preprint arXiv:1412.5632 (2014).

[LWB13]   Kiryung Lee, Yihong Wu, and Yoram Bresler, *Near optimal compressed sensing of sparse rank-one matrices via sparse power factorization*, arXiv preprint arXiv:1312.0525 (2013).

[MBP14]   Julien Mairal, Francis Bach, and Jean Ponce, *Sparse modeling for image and vision processing*, Foundations and Trends in Computer Graphics and Vision **8** (2014), no. 2-3, 85–283.

[MG13]   Nishant Mehta and Alexander G. Gray, *Sparsity-based generalization bounds for predictive sparse coding*, Proceedings of the 30th International Conference on Machine Learning (ICML-13) **28** (2013), no. 1, 36–44.

[MHWG13] Cun Mu, Bo Huang, John Wright, and Donald Goldfarb, *Square deal: Lower bounds and improved relaxations for tensor recovery*, arXiv preprint arXiv:1307.5870 (2013).

[MK87]   Katta G Murty and Santosh N Kabadi, *Some np-complete problems in quadratic and nonlinear programming*, Mathematical programming **39** (1987), no. 2, 117–129.

[MP10a]   Jianwei Ma and Gerlind Plonka, *A review of curvelets and recent applications*, IEEE Signal Processing Magazine **27** (2010), no. 2, 118–133.

[MP10b]   Andreas Maurer and Massimiliano Pontil, *K-dimensional coding schemes in hilbert spaces*, Information Theory, IEEE Transactions on **56** (2010), no. 11, 5839–5846.

[MS83]   Jorge J Moré and Danny C Sorensen, *Computing a trust region step*, SIAM Journal on Scientific and Statistical Computing **4** (1983), no. 3, 553–572.

[MT14]   Michael B McCoy and Joel A Tropp, *Sharp recovery bounds for convex demixing, with applications*, Foundations of Computational Mathematics **14** (2014), no. 3, 503–567.

[NJS13]   Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi, *Phase retrieval using alternating minimization*, Advances in Neural Information Processing Systems, 2013, pp. 2796–2804.

[NNS+14]   Praneeth Netrapalli, UN Niranjan, Sujay Sanghavi, Animashree Anandkumar, and Prateek Jain, *Non-convex robust pca*, Advances in Neural Information Processing Systems, 2014, pp. 1107–1115.

[NP13]   Behnam Neyshabur and Rina Panigrahy, *Sparse matrix factorization*, arXiv preprint arXiv:1311.3315 (2013).

[NW06]   Jorge Nocedal and Stephen Wright, *Numerical optimization*, Springer, 2006.

[OF96]   Bruno A. Olshausen and David J. Field, *Emergence of simple-cell receptive field properties by learning a sparse code for natural images*, Nature **381** (1996), no. 6583, 607–609.

[OF97]   _____ , *Sparse coding with an overcomplete basis set: A strategy employed by v1?*, Vision research **37** (1997), no. 23, 3311–3325.

[OH10]   Samet Oymak and Babak Hassibi, *New null space results and recovery thresholds for matrix rank minimization*, arXiv preprint arXiv:1011.6326 (2010).

[QSW14]   Qing Qu, Ju Sun, and John Wright, *Finding a sparse vector in a subspace: Linear sparsity using alternating directions*, Advances in Neural Information Processing Systems, 2014, pp. 3401–3409.

[RW97]   Franz Rendl and Henry Wolkowicz, *A semidefinite framework for trust region subproblems with applications to large scale minimization*, Mathematical Programming **77** (1997), no. 1, 273–299.

[SA14]   Hanie Sedghi and Anima Anandkumar, *Provable methods for training neural networks with sparse connectivity*, arXiv preprint arXiv:1412.2693 (2014).

[Sch14a]   Karin Schnass, *Local identification of overcomplete dictionaries*, arXiv preprint arXiv:1401.6354 (2014).

[Sch14b]   _____ , *On the identifiability of overcomplete dictionaries via the minimisation principle underlying k-svd*, Applied and Computational Harmonic Analysis **37** (2014), no. 3, 464–491.

[Sch15]   _____ , *Convergence radius and sample complexity of itkm algorithms for dictionary learning*, arXiv preprint arXiv:1503.07027 (2015).

[SLLC15]   Wei Sun, Junwei Lu, Han Liu, and Guang Cheng, *Provable sparse tensor decomposition*, arXiv preprint arXiv:1502.01425 (2015).

[SS90]   Gilbert W Stewart and Ji-guang Sun, *Matrix perturbation theory*, Academic press, 1990.

[SWW12]   Daniel A Spielman, Huan Wang, and John Wright, *Exact recovery of sparsely-used dictionaries*, Proceedings of the 25th Annual Conference on Learning Theory, 2012.

[Tem03]   Vladimir N Temlyakov, *Nonlinear methods of approximation*, Foundations of Computational Mathematics **3** (2003), no. 1, 33–107.

[Tro12]   Joel A. Tropp, *User-friendly tail bounds for sums of random matrices*, Foundations of Computational Mathematics **12** (2012), no. 4, 389–434.

[Tse01]   Paul Tseng, *Convergence of a block coordinate descent method for nondifferentiable minimization*, Journal of optimization theory and applications **109** (2001), no. 3, 475–494.

[Udr94]   Constantin Udriste, *Convex functions and optimization methods on riemannian manifolds*, vol. 297, Springer Science & Business Media, 1994.

[VMB11]   Daniel Vainsencher, Shie Mannor, and Alfred M. Bruckstein, *The sample complexity of dictionary learning*, Journal of Machine Learning Research **12** (2011), no. 23, 3259–3281.

[WGNL14]   Zhaoran Wang, Quanquan Gu, Yang Ning, and Han Liu, *High dimensional expectation-maximization algorithm: Statistical optimization and asymptotic normality*, arXiv preprint arXiv:1412.8729 (2014).

[WLL14]   Zhaoran Wang, Huanran Lu, and Han Liu, *Nonconvex statistical optimization: minimax-optimal sparse pca in polynomial time*, arXiv preprint arXiv:1408.5352 (2014).

[YCS13] Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi, *Alternating minimization for mixed linear regression*, arXiv preprint arXiv:1310.3745 (2013).

[YZ03] Yinyu Ye and Shuzhong Zhang, *New results on quadratic minimization*, SIAM Journal on Optimization **14** (2003), no. 1, 245–267.

[ZP01] Michael Zibulevsky and Barak Pearlmutter, *Blind source separation by sparse decomposition in a signal dictionary*, Neural computation **13** (2001), no. 4, 863–882.