Complete Dictionary Recovery Using Nonconvex Optimization

Ju Sun JS4038@COLUMBIA.EDU
Qing Qu QQ2105@COLUMBIA.EDU
John Wright JW2966@COLUMBIA.EDU
Department of Electrical Engineering, Columbia University, New York, NY, USA

Abstract

We consider the problem of recovering a complete (i.e., square and invertible) dictionary $\boldsymbol{A}_0$, from $\boldsymbol{Y} = \boldsymbol{A}_0 \boldsymbol{X}_0$ with $\boldsymbol{Y} \in \mathbb{R}^{n \times p}$. This recovery setting is central to the theoretical understanding of dictionary learning. We give the first efficient algorithm that provably recovers $\boldsymbol{A}_0$ when $\boldsymbol{X}_0$ has $O(n)$ nonzeros per column, under suitable probability model for $\boldsymbol{X}_0$. Prior results provide recovery guarantees when $\boldsymbol{X}_0$ has only $O(\sqrt{n})$ nonzeros per column. Our algorithm is based on nonconvex optimization with a spherical constraint, and hence is naturally phrased in the language of manifold optimization. Our proofs give a geometric characterization of the high-dimensional objective landscape, which shows that with high probability there are no spurious local minima. Experiments with synthetic data corroborate our theory. Full version of this paper is available online: http://arxiv.org/abs/1504.06785.

1. Introduction

Dictionary learning (DL) is the problem of finding a sparse representation for a collection of input signals. Its applications span classical image processing, visual recognition, compressive signal acquisition, and also recent deep architectures for signal classification. Recent surveys on applications and algorithms of DL include Elad (2010) and Mairal et al (2014).

Formally, given a data matrix $\boldsymbol{Y} \in \mathbb{R}^{n \times p}$, DL seeks an approximation $\boldsymbol{Y} \approx \boldsymbol{A}\boldsymbol{X}$, where $\boldsymbol{A}$ lies in a certain admissible set $\mathcal{A}$, and $\boldsymbol{X}$ is as sparse as possible. Typical formulations for DL are nonconvex: the admissible set $\mathcal{A}$ is typically nonconvex, and the observation map $(\boldsymbol{A}, \boldsymbol{X}) \mapsto \boldsymbol{A}\boldsymbol{X}$ is bilinear. There is also an intrinsic symmetry in the problem due to sign-permutation ambiguity, which seems to preclude convexification (Gribonval & Schnass (2010), Geng & Wright (2011)). Thus, despite many empirical successes, relatively little is known about the theoretical properties of DL algorithms.

Towards theoretical understanding, it is natural to start with the dictionary recovery (DR) problem: suppose that the data matrix $\boldsymbol{Y}$ is generated as $\boldsymbol{Y} = \boldsymbol{A}_0 \boldsymbol{X}_0$, where $\boldsymbol{A}_0 \in \mathbb{R}^{n \times m}$ and $\boldsymbol{X}_0 \in \mathbb{R}^{m \times p}$, and try to recover $\boldsymbol{A}_0$ and $\boldsymbol{X}_0$. One might imagine putting favorable structural assumptions on $\boldsymbol{A}_0$ and $\boldsymbol{X}_0$ to make DR well-posed and amenable to efficient algorithms. However, under natural assumptions for $\boldsymbol{A}_0$ and $\boldsymbol{X}_0$, even proving that the target solution is a *local* minimum of certain popular practical DL formulations requires nontrivial analysis (Gribonval & Schnass, 2010; Geng & Wright, 2011; Schnass, 2014a;b; 2015). Obtaining *global* solutions with *efficient* algorithms is a standing challenge.

Existing results on global dictionary recovery pertain only to highly sparse $\boldsymbol{X}_0$. For example, Spielman et al (2012) showed that solving a sequence of certain linear programs can recover a complete dictionary $\boldsymbol{A}_0$ from $\boldsymbol{Y}$, when $\boldsymbol{X}_0$ is a sparse random matrix with $O(\sqrt{n})$ nonzeros per column. Agarwal et al (2013a; 2013b) and Arora et al (2013; 2015) have subsequently given efficient algorithms for the overcomplete setting ($m \geq n$), based on a combination of {clustering or spectral initialization} and local refinement. These algorithms again succeed when $\boldsymbol{X}_0$ has $\widetilde{O}(\sqrt{n})$ nonzeros per column (The $\widetilde{O}$ suppresses some logarithm factors). Barak et al (2014) provides efficient algorithms based on sum-of-square hierarchy that guarantees recovery of complete dictionaries when $\boldsymbol{X}_0$ has $O(n^c)$ nonzeros per column for any $c < 1$. Giving *efficient* algorithms which provably succeeds in linear sparsity regime (i.e., $O(n)$ nonzeros per column) is an open problem. [1]

[1] Recent works, including Arora et al (2014) and Barak et al (2014), contain guarantees for recovery with linear sparsity, but run in super-polynomial (quasipolynomial) time. Aside from efficient recovery, other theoretical work on DL includes results on identifiability (Hillar & Sommer, 2011), generalization bounds (Vainsencher et al., 2011; Mehta & Gray, 2013), and noise

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

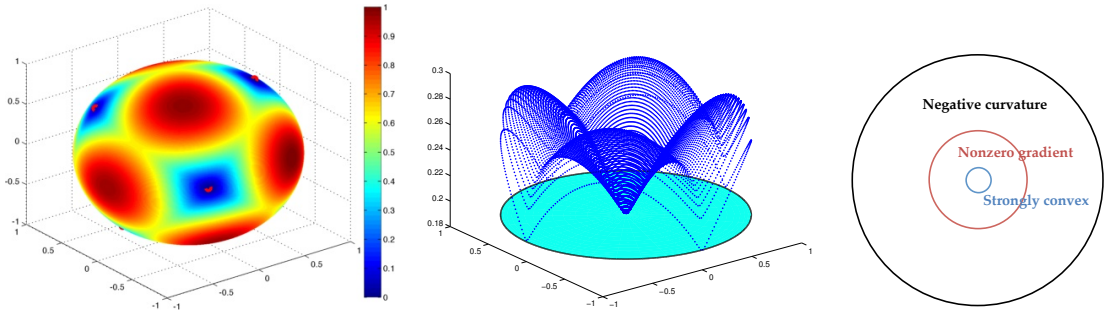

Figure 1. **Why is dictionary learning over $\mathbb{S}^{n-1}$ tractable?** Assume the target dictionary $\boldsymbol{A}_0$ is orthogonal. **Left:** Large sample objective function $\mathbb{E}_{\boldsymbol{X}_0}[f(\boldsymbol{q})]$. The only local minima are the columns of $\boldsymbol{A}_0$ and their 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.

In this work, we focus on recovering a complete (i.e., invertible) dictionary $\boldsymbol{A}_0$ from $\boldsymbol{Y} = \boldsymbol{A}_0\boldsymbol{X}_0$. We give the first *polynomial-time* algorithm that provably recovers $\boldsymbol{A}_0$ when $\boldsymbol{X}_0$ has $O(n)$ nonzeros per column, if $\boldsymbol{X}_0$ contains i.i.d. Bernoulli-Gaussian entries. We achieve this by formulating complete DR as a nonconvex program with spherical constraint. Under the probability model for $\boldsymbol{X}_0$, we give a geometric characterization of the high-dimensional objective landscape over the sphere, which shows that with high probability (w.h.p.) there are no "spurious" local minima. In particular, the geometric structure allows us to design a Riemannian trust region algorithm over the sphere that provably converges to one local minimum with an *arbitrary initialization*, despite the presence of saddle points.

This paper is organized as follows. Sec. 2 will motivate our nonconvex formulation of DR and formalize the setting. Sec. 3 and 4 will introduce two integral parts of our algorithmic framework: characterization of the high-dimensional function landscape and a Riemannian trust-region algorithm over the sphere. Main theorems are included in Sec. 5, followed by some numerical verification presented in Sec. 6. Sec. 7 will close this paper by discussing implications of this work. Due to space constraint, we only sketch high-level ideas of the proof in this paper. Detailed proofs can be found in the full version (Sun et al., 2015).

2. A Nonconvex Formulation for DR

We assume $\boldsymbol{Y} = \boldsymbol{A}_0\boldsymbol{X}_0$, where $\boldsymbol{A}_0 \in \mathbb{R}^{n\times n}$ is a complete matrix and $\boldsymbol{X}_0$ follows the Bernoulli-Gaussian (BG) model with rate θ: $[\boldsymbol{X}_0]_{ij} = \Omega_{ij}G_{ij}$, with $\Omega_{ij} \sim \mathrm{Ber}(\theta)$ and $V_{ij} \sim \mathcal{N}(0,1)$. We write compactly $\boldsymbol{X}_0 \sim_{i.i.d.} \mathrm{BG}(\theta)$.

Since $\boldsymbol{Y} = \boldsymbol{A}_0\boldsymbol{X}_0$ and $\boldsymbol{A}_0$ is complete, $\mathrm{row}(\boldsymbol{Y}) = \mathrm{row}(\boldsymbol{X}_0)$ [2] and rows of $\boldsymbol{X}_0$ are sparse vectors in the known subspace $\mathrm{row}(\boldsymbol{Y})$. Following Spielman et al (2012), we 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 $\boldsymbol{X}_0 \sim_{i.i.d.} \mathrm{BG}(\theta)$, rows of $\boldsymbol{X}_0$ are the n *sparsest* vectors (directions) in $\mathrm{row}(\boldsymbol{Y})$ w.h.p. (Spielman et al., 2012). Thus one might try to recover them by solving[3]

$$\min \ \|\boldsymbol{q}^*\boldsymbol{Y}\|_0 \ \ \text{s.t.} \ \ \boldsymbol{q} \neq \boldsymbol{0}. \tag{2.1}$$

The objective is discontinuous, and the domain is an open set. Known convex relaxations (Spielman et al., 2012; Demanet & Hand, 2014) provably break down beyond the aforementioned $O(\sqrt{n})$ sparsity level. Instead, we work with a *nonconvex* alternative: [4]

$$\min \ f(\boldsymbol{q};\widehat{\boldsymbol{Y}}) \doteq \frac{1}{p}\sum_{k=1}^{p} h_\mu(\boldsymbol{q}^*\widehat{\boldsymbol{y}}_k), \ \text{s.t.} \ \|\boldsymbol{q}\|_2 = 1, \tag{2.2}$$

where $\widehat{\boldsymbol{Y}} \in \mathbb{R}^{n\times p}$ is a proxy of $\boldsymbol{Y}$ and k indexes columns of $\widehat{\boldsymbol{Y}}$. Here $h_\mu(\cdot)$ is chosen to be a convex smooth approximation to the $|\cdot|$ function, namely,

$$h_\mu(z) = \mu \log\cosh\left(\frac{z}{\mu}\right), \tag{2.3}$$

which is infinitely differentiable and μ controls the smoothing level. The spherical constraint is nonconvex. Hence, a-priori, it is unclear whether (2.2) admits efficient algorithms that attain one local optimizer (Murty & Kabadi, 1987). Surprisingly, simple descent algorithms for (2.2) exhibit very striking behavior: on many practical numerical examples[5], they appear to produce global solutions. Our next section will uncover interesting geometrical structures underlying the phenomenon.

stability (Gribonval et al., 2014).

[2] $\mathrm{row}(\cdot)$ denotes the row space.

[3] The notation * denotes matrix transposition.

[4] Similar formulation has been proposed in (Zibulevsky & Pearlmutter, 2001) in the context of blind source separation, see also (Qu et al., 2014).

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

3. High-dimensional Geometry

For the moment, suppose $\boldsymbol{A}_0$ is orthogonal, and take $\widehat{\boldsymbol{Y}} = \boldsymbol{Y} = \boldsymbol{A}_0\boldsymbol{X}_0$ in (2.2). Figure. 1 (left) plots $\mathbb{E}_{\boldsymbol{X}_0}\left[f\left(\boldsymbol{q};\boldsymbol{Y}\right)\right]$ over $\boldsymbol{q} \in \mathbb{S}^2$ ($n = 3$). Remarkably, $\mathbb{E}_{\boldsymbol{X}_0}\left[f\left(\boldsymbol{q};\boldsymbol{Y}\right)\right]$ has no spurious local minima. In fact, every local minimum $\widehat{\boldsymbol{q}}$ 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 1 (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. 1 (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 take advantage of this structure without knowing $\boldsymbol{A}_0$ ahead of time. In Sec. 4, we describe a Riemannian trust region method which addresses the later challenge. Now we focus on the first one.

3.1. Geometry for orthogonal $\boldsymbol{A}_0$

In this case, we take $\widehat{\boldsymbol{Y}} = \boldsymbol{Y} = \boldsymbol{A}_0\boldsymbol{X}_0$. Since $f\left(\boldsymbol{q};\boldsymbol{A}_0\boldsymbol{X}_0\right) = f\left(\boldsymbol{A}_0^*\boldsymbol{q};\boldsymbol{X}_0\right)$, the landscape of $f\left(\boldsymbol{q};\boldsymbol{A}_0\boldsymbol{X}_0\right)$ is simply a rotated version of that of $f\left(\boldsymbol{q};\boldsymbol{X}_0\right)$, i.e., when $\boldsymbol{A}_0 = \boldsymbol{I}$. Hence we will focus on the case when $\boldsymbol{A}_0 = \boldsymbol{I}$. Among the $2n$ symmetric sections of $\mathbb{S}^{n-1}$ centered around the signed basis vectors $\pm\boldsymbol{e}_1, \ldots, \pm\boldsymbol{e}_n$, we work with the section around $\boldsymbol{e}_n$ as an example. The result will carry over to all sections with the same argument.

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

$$\boldsymbol{q}\left(\boldsymbol{w}\right) = \left(\boldsymbol{w}, \sqrt{1 - \left\|\boldsymbol{w}\right\|^2}\right),\ \boldsymbol{w} \in \mathbb{R}^{n-1}, \tag{3.1}$$

where $\boldsymbol{w}$ is the new variable in $\boldsymbol{e}_n^\perp$. We study the composition $g\left(\boldsymbol{w};\boldsymbol{Y}\right) \doteq f\left(\boldsymbol{q}\left(\boldsymbol{w}\right);\boldsymbol{Y}\right)$ over the set

$$\Gamma \doteq \left\{\boldsymbol{w} : \left\|\boldsymbol{w}\right\| < \sqrt{\tfrac{4n-1}{4n}}\right\}. \tag{3.2}$$

Our next theorem characterizes the properties of $g\left(\boldsymbol{w}\right)$. In particular, it shows the favorable structure we observed for $n = 3$ persists in high dimensions, w.h.p.[6], even when p is large *yet finite*, for the case $\boldsymbol{A}_0$ is orthogonal.

Theorem 3.1. *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 $\mu < \min\left\{c_a\theta n^{-1}, c_b n^{-5/4}\right\}$, whenever $p \geq \frac{C}{\mu^2\theta^2}n^3\log\frac{n}{\mu\theta}$, the following hold simultaneously w.h.p.:*

$$\begin{aligned}
\nabla^2 g(\boldsymbol{w};\boldsymbol{X}_0) &\succeq \frac{c_\star\theta}{\mu}\boldsymbol{I} \quad \forall\, \boldsymbol{w}\ \|\boldsymbol{w}\| \leq \frac{\mu}{4\sqrt{2}},\\
\frac{\boldsymbol{w}^*\nabla g(\boldsymbol{w};\boldsymbol{X}_0)}{\|\boldsymbol{w}\|} &\geq c_\star\theta \quad \forall\, \boldsymbol{w}\ \frac{\mu}{4\sqrt{2}} \leq \|\boldsymbol{w}\| \leq \frac{1}{20\sqrt{5}}\\
\frac{\boldsymbol{w}^*\nabla^2 g(\boldsymbol{w};\boldsymbol{X}_0)\boldsymbol{w}}{\|\boldsymbol{w}\|^2} &\leq -c_\star\theta \quad \forall\, \boldsymbol{w}\ \frac{1}{20\sqrt{5}} \leq \|\boldsymbol{w}\| \leq \sqrt{\frac{4n-1}{4n}},
\end{aligned}$$

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

$$\left\|\boldsymbol{w}_\star - \boldsymbol{0}\right\| \ \leq\ \min\left\{\frac{c_c\mu}{\theta}\sqrt{\frac{n\log p}{p}}, \frac{\mu}{16}\right\}. \tag{3.3}$$

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

In words, when the samples are numerous enough, one sees the strongly convex, nonzero gradient, and negative curvature regions successively when moving away from target solution $\boldsymbol{0}$, and the local (also global) minimizer of $g\left(\boldsymbol{w};\boldsymbol{Y}\right)$ is next to $\boldsymbol{0}$, within a distance of $O\left(\mu\right)$.

Note that $\boldsymbol{q}(\Gamma)$ contains all points $\boldsymbol{q} \in \mathbb{S}^{n-1}$ such that $\boldsymbol{q}^*\boldsymbol{e}_n = \max_i\left|\boldsymbol{q}^*\boldsymbol{e}_i\right|$. We can characterize the graph of the function $f(\boldsymbol{q};\boldsymbol{X}_0)$ in the vicinity of some 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 3.1 by $2n$), we obtain a characterization of $f(\boldsymbol{q})$ over the entirety of $\mathbb{S}^{n-1}$.

Corollary 3.2. *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 $1 - 8n^2p^{-10} - \theta(np)^{-7} - \exp\left(-0.3\theta np\right) - c_c\exp\left(-c_d p\mu^2\theta^2/n^2\right)$, the function $f\left(\boldsymbol{q};\boldsymbol{X}_0\right)$ 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 $\left\{\pm\boldsymbol{e}_i\right\}_i$, such*

[6] In this work, we say some event occurs with high probability when the failure probability is bounded by an inverse polynomial of n and p.

that the corresponding local minimizer $\boldsymbol{q}_\star$ and $\boldsymbol{b} \in \{\pm \boldsymbol{e}_i\}_i$ satisfy

$$\|\boldsymbol{q}_\star - \boldsymbol{b}\| \le \sqrt{2}\min\left\{\frac{c_c\mu}{\theta}\sqrt{\frac{n\log p}{p}}, \frac{\mu}{16}\right\}. \tag{3.4}$$

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

The proof of Theorem 3.1 is conceptually straightforward: one shows that $\mathbb{E}[f(\boldsymbol{q}; \boldsymbol{X}_0)]$ has the claimed properties, and then proves that each of the quantities of interest concentrates uniformly about its expectation. The detailed calculations are nontrivial.

3.2. Geometry for complete $\boldsymbol{A}_0$

For general complete dictionaries $\boldsymbol{A}_0$, we hope that the function f retains the nice geometric structure discussed above. We can ensure this by "preconditioning" $\boldsymbol{Y}$ such that the output looks as if being generated from a certain orthogonal matrix, possibly plus a small perturbation. We can then argue that the perturbation does not significantly affect the properties of the graph of the objective function. Write

$$\overline{\boldsymbol{Y}} = \left(\tfrac{1}{p\theta}\boldsymbol{Y}\boldsymbol{Y}^*\right)^{-1/2}\boldsymbol{Y}. \tag{3.5}$$

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 $\frac{1}{p\theta}\boldsymbol{Y}\boldsymbol{Y}^* = \frac{1}{p\theta}\boldsymbol{A}_0\boldsymbol{X}_0\boldsymbol{X}_0^*\boldsymbol{A}_0^*$ to behave roughly like $\boldsymbol{A}_0\boldsymbol{A}_0^*$ and hence $\overline{\boldsymbol{Y}}$ to behave like

$$\left(\boldsymbol{A}_0\boldsymbol{A}_0^*\right)^{-1/2}\boldsymbol{A}_0\boldsymbol{X}_0 = \boldsymbol{U}\boldsymbol{V}^*\boldsymbol{X}_0 \tag{3.6}$$

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

$$\overline{\boldsymbol{Y}} = \boldsymbol{U}\boldsymbol{V}^*\boldsymbol{X}_0 + \boldsymbol{\Xi}\boldsymbol{X}_0, \tag{3.7}$$

where $\boldsymbol{\Xi}$ is a matrix with small magnitude. Perturbation analysis combines with the the above results for the orthogonal case yields:

Theorem 3.3. *Suppose $\boldsymbol{A}_0$ is complete with its condition number $\kappa(\boldsymbol{A}_0)$. 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 $p \ge \frac{C}{c_\star^2\theta}\max\left\{\frac{n^4}{\mu^4}, \frac{n^5}{\mu^2}\right\}\kappa^8(\boldsymbol{A}_0)\log^4\left(\frac{\kappa(\boldsymbol{A}_0)n}{\mu\theta}\right)$ and $\overline{\boldsymbol{Y}} \doteq \sqrt{p\theta}\left(\boldsymbol{Y}\boldsymbol{Y}^*\right)^{-1/2}\boldsymbol{Y}$, $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = \mathtt{SVD}(\boldsymbol{A}_0)$, and let $\widetilde{\boldsymbol{Y}} = \boldsymbol{V}\boldsymbol{U}^*\overline{\boldsymbol{Y}}$, then the following hold simultaneously w.h.p.:*

$$\nabla^2 g(\boldsymbol{w}; \widetilde{\boldsymbol{Y}}) \succeq \frac{c_\star\theta}{2\mu}\boldsymbol{I} \qquad \forall \boldsymbol{w}\, \|\boldsymbol{w}\| \le \frac{\mu}{4\sqrt{2}},$$

$$\frac{\boldsymbol{w}^*\nabla g(\boldsymbol{w}; \widetilde{\boldsymbol{Y}})}{\|\boldsymbol{w}\|} \ge \frac{1}{2}c_\star\theta \qquad \forall \boldsymbol{w}\, \frac{\mu}{4\sqrt{2}} \le \|\boldsymbol{w}\| \le \frac{1}{20\sqrt{5}}$$

$$\frac{\boldsymbol{w}^*\nabla^2 g(\boldsymbol{w}; \widetilde{\boldsymbol{Y}})\boldsymbol{w}}{\|\boldsymbol{w}\|^2} \le -\frac{1}{2}c_\star\theta \quad \forall \boldsymbol{w}\, \frac{1}{20\sqrt{5}} \le \|\boldsymbol{w}\| \le \sqrt{\frac{4n-1}{4n}}$$

and the function $g(\boldsymbol{w}; \widetilde{\boldsymbol{Y}})$ 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

$$\|\boldsymbol{w}_\star - \boldsymbol{0}\| \le \frac{\mu}{7}. \tag{3.8}$$

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 3.4. *Suppose $\boldsymbol{A}_0$ is complete with its condition number $\kappa(\boldsymbol{A}_0)$. 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 $p \ge \frac{C}{c_\star^2\theta}\max\left\{\frac{n^4}{\mu^4}, \frac{n^5}{\mu^2}\right\}\kappa^8(\boldsymbol{A}_0)\log^4\left(\frac{\kappa(\boldsymbol{A}_0)n}{\mu\theta}\right)$ and $\overline{\boldsymbol{Y}} \doteq \sqrt{p\theta}\left(\boldsymbol{Y}\boldsymbol{Y}^*\right)^{-1/2}\boldsymbol{Y}$, $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = \mathtt{SVD}(\boldsymbol{A}_0)$, with probability at least $1 - 8n^2p^{-10} - \theta(np)^{-7} - \exp(-0.3\theta np) - p^{-8} - c_d\exp\left(-c_e p\mu^2\theta^2/n^2\right)$, the function $f\left(\boldsymbol{q}; \boldsymbol{V}\boldsymbol{U}^*\overline{\boldsymbol{Y}}\right)$ 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*

$$\|\boldsymbol{q}_\star - \boldsymbol{b}\| \le \frac{\sqrt{2}\mu}{7}. \tag{3.9}$$

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

4. Riemannian Trust Region Algorithm

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. 1). Therefore, certain form of second-order information is needed to help escape the saddle points. Based on these considerations, we describe a Riemannian trust region method (TRM) (Absil et al., 2007; 2009) over the sphere for this purpose.

4.1. Trust region method for Euclidean spaces

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

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}), \tag{4.1}$$

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

$$\widehat{f}\left(\boldsymbol{\delta};\boldsymbol{x}_{k-1}\right) \doteq f\left(\boldsymbol{x}_{k-1}\right)+\nabla^{*} f\left(\boldsymbol{x}_{k-1}\right)\boldsymbol{\delta} + \tfrac{1}{2}\boldsymbol{\delta}^{*}\boldsymbol{Q}\left(\boldsymbol{x}_{k-1}\right)\boldsymbol{\delta}, \quad (4.2)$$

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 iterate is determined by seeking a minimum of $\widehat{f}\left(\boldsymbol{\delta};\boldsymbol{x}_{k-1}\right)$ over a small region, normally an ℓ^2 ball, commonly known as the trust region. Thus, the well known trust region subproblem takes the form

$$\boldsymbol{\delta}_k \doteq \underset{\boldsymbol{\delta}\in\mathbb{R}^n,\|\boldsymbol{x}\|\le\Delta}{\arg\min}\ \widehat{f}\left(\boldsymbol{\delta};\boldsymbol{x}_{k-1}\right), \quad (4.3)$$

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

$$\rho_k \doteq \frac{f\left(\boldsymbol{x}_{k-1}\right)-f\left(\boldsymbol{x}_{k-1}+\boldsymbol{\delta}_k\right)}{\widehat{f}\left(\boldsymbol{0}\right)-\widehat{f}\left(\boldsymbol{\delta}_{k-1}\right)} \quad (4.4)$$

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. If the progress is satisfactory, the next iterate is (perhaps plus some line search improvement)

$$\boldsymbol{x}_k = \boldsymbol{x}_{k-1}+\boldsymbol{\delta}_k. \quad (4.5)$$

Detailed introduction to the classical TRM can be found in the texts (Conn et al., 2000; Nocedal & Wright, 2006).

4.2. Trust region method over the sphere

To generalize the idea to smooth manifolds, one natural choice is to form the approximation over the tangent spaces (Absil et al., 2007; 2009). Specific to our spherical manifold, for which the tangent space at an iterate $\boldsymbol{q}_k \in \mathbb{S}^{n-1}$ is $T_{\boldsymbol{q}_k}\mathbb{S}^{n-1} \doteq \{\boldsymbol{v} : \boldsymbol{v}^*\boldsymbol{q}_k = 0\}$, we work with the quadratic approximation $\widehat{f} : T_{\boldsymbol{q}_k}\mathbb{S}^{n-1} \mapsto \mathbb{R}$ defined as

$$\widehat{f}(\boldsymbol{q}_k,\boldsymbol{\delta}) \doteq f(\boldsymbol{q}_k)+\langle\nabla f(\boldsymbol{q}_k),\boldsymbol{\delta}\rangle + \frac{1}{2}\boldsymbol{\delta}^{*}\left(\nabla^2 f(\boldsymbol{q}_k)-\langle\nabla f(\boldsymbol{q}_k),\boldsymbol{q}_k\rangle\,\boldsymbol{I}\right)\boldsymbol{\delta}. \quad (4.6)$$

To interpret this approximation, let $\mathcal{P}_{T_{\boldsymbol{q}_k}\mathbb{S}^{n-1}} \doteq (\boldsymbol{I}-\boldsymbol{q}_k\boldsymbol{q}_k^{*})$ be the orthoprojector onto $T_{\boldsymbol{q}_k}\mathbb{S}^{n-1}$ and write (4.6) into an equivalent form:

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

The two terms

$$\mathrm{grad} f\left(\boldsymbol{q}_k\right) \doteq \mathcal{P}_{T_{\boldsymbol{q}_k}\mathbb{S}^{n-1}}\nabla f(\boldsymbol{q}_k),$$

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

are the Riemannian gradient and Riemannian Hessian of f w.r.t. $\mathbb{S}^{n-1}$, respectively (Absil et al., 2007; 2009), turning (4.6) into the form of familiar quadratic approximation, as described in (4.2).

Then the Riemannian trust-region subproblem is

$$\min_{\boldsymbol{\delta}\in T_{\boldsymbol{q}_k}\mathbb{S}^{n-1},\ \|\boldsymbol{\delta}\|\le\Delta}\ \widehat{f}\left(\boldsymbol{q}_k,\boldsymbol{\delta}\right), \quad (4.7)$$

where $\Delta > 0$ is the familiar trust-region parameter. Taking any orthonormal basis $\boldsymbol{U}_{\boldsymbol{q}_k}$ for $T_{\boldsymbol{q}_k}\mathbb{S}^{n-1}$, we can transform (4.7) into a classical trust-region subproblem:

$$\min_{\|\boldsymbol{\xi}\|\le\Delta}\ \widehat{f}\left(\boldsymbol{q}_k,\boldsymbol{U}_{\boldsymbol{q}_k}\boldsymbol{\xi}\right), \quad (4.8)$$

for which very efficient numerical algorithms exist (Moré & Sorensen, 1983; Hazan & Koren, 2014). Once we obtain the minimizer $\boldsymbol{\xi}_\star$, we set $\boldsymbol{\delta}_\star = \boldsymbol{U}\boldsymbol{\xi}_\star$, which solves (4.7).

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 manifold. To resolve this, we resort to the natural exponential map:

$$\boldsymbol{q}_{k+1} \doteq \exp_{\boldsymbol{q}_k}\boldsymbol{\delta}_\star = \boldsymbol{q}_k\cos\|\boldsymbol{\delta}_\star\| + \frac{\boldsymbol{\delta}_\star}{\|\boldsymbol{\delta}_\star\|}\sin\|\boldsymbol{\delta}_\star\|, \quad (4.9)$$

which move the sequence to the next iterate "along the direction"[7] of $\boldsymbol{\delta}_\star$ while staying over the sphere.

There are many variants of (Riemannian) TRM that allow one to solve the subproblem 4.8 only approximately while still guarantee convergence. For simplicity, we avoid the extra burden caused thereof for analysis by solving the subproblem exactly via SDP relaxation: introduce

$$\tilde{\boldsymbol{\xi}} = [\boldsymbol{\xi}^{*},1]^{*},\ \boldsymbol{\Xi} = \tilde{\boldsymbol{\xi}}\tilde{\boldsymbol{\xi}}^{*},\ \boldsymbol{M} = \begin{bmatrix}\boldsymbol{A} & \boldsymbol{b}\\ \boldsymbol{b}^{*} & 0\end{bmatrix} \quad (4.10)$$

where $\boldsymbol{A} = \boldsymbol{U}^{*}\left(\nabla^2 f(\boldsymbol{q})-\langle\nabla f(\boldsymbol{q}),\boldsymbol{q}\rangle\,\boldsymbol{I}\right)\boldsymbol{U}$ and $\boldsymbol{b} = \boldsymbol{U}^{*}\nabla f(\boldsymbol{q})$ for any orthobasis $\boldsymbol{U}$ of $T_{\boldsymbol{q}_{k-1}}\mathbb{S}^{n-1}$. The subproblem is known to be equivalent to the SDP problem (Fortin & Wolkowicz, 2004):

$$\min_{\boldsymbol{\Xi}}\ \langle\boldsymbol{M},\boldsymbol{\Xi}\rangle,\quad \text{s.t.}\,\mathrm{tr}(\boldsymbol{\Xi}) \le \Delta^2+1,\ \langle\boldsymbol{E}_{n+1},\boldsymbol{\Xi}\rangle = 1,\ \boldsymbol{\Xi}\succeq\boldsymbol{0}, \quad (4.11)$$

where $\boldsymbol{E}_{n+1} = \boldsymbol{e}_{n+1}\boldsymbol{e}_{n+1}^{*}$. The detailed trust region algorithm is presented in Algorithm 1.

4.3. Algorithmic results

Using the geometric characterization in Theorem 3.1 and Theorem 3.3, we prove that when the parameter Δ is suf-

[7] Technically, moving along a curve on the manifold of which $\boldsymbol{\delta}_\star$ is the initial tangent vector in certain canonical way.

Algorithm 1 Trust Region Method for Finding a Single Sparse Vector

Input: data matrix $\boldsymbol{Y} \in \mathbb{R}^{n\times p}$, smoothing parameter μ and parameters $\eta_{vs} = 0.9$, $\eta_s = 0.1$, $\gamma_i = 2$, $\gamma_d = 1/2$, $\Delta_{\max} = 1$, and $\Delta_{\min} = 10^{-16}$.

Output: $\widehat{\boldsymbol{q}} \in \mathbb{S}^{n-1}$

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

$$\widehat{\boldsymbol{\xi}} = \mathop{\arg\min}_{\|\boldsymbol{\xi}\| \le \Delta^{(k-1)}} \widehat{f}\left(\boldsymbol{q}^{(k-1)}, \boldsymbol{U}\boldsymbol{\xi}\right) \tag{4.12}$$

  Set

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

  Set

$$\rho_k \leftarrow \frac{f(\boldsymbol{q}^{(k-1)}) - f(\widehat{\boldsymbol{q}})}{f(\boldsymbol{q}^{(k-1)}) - \widehat{f}(\boldsymbol{q}^{(k-1)}, \widehat{\boldsymbol{\delta}})} \tag{4.13}$$

  if $\rho_k \ge \eta_{vs}$ **then**
    Set $\boldsymbol{q}^{(k)} \leftarrow \widehat{\boldsymbol{q}}$, $\Delta^{(k)} \leftarrow \min\left(\gamma_i \Delta^{(k-1)}, \Delta_{\max}\right)$.
  else if $\rho_k \ge \eta_s$ **then**
    Set $\boldsymbol{q}^{(k)} \leftarrow \widehat{\boldsymbol{q}}$, $\Delta^{(k)} \leftarrow \Delta^{(k-1)}$.
  else
    Set $\boldsymbol{q}^{(k)} \leftarrow \boldsymbol{q}^{(k-1)}$,
    $\Delta^{(k)} \leftarrow \max(\gamma_d \Delta^{(k-1)}, \Delta_{\min})$.
  end if
  Set $k = k + 1$.
end while

ficiently small[8], (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 global minima with an asymptotic quadratic rate. In particular, the geometry implies that from *any initialization*, the iterate sequence converges to a close approximation to one local minimizer in a polynomial number of steps.

The following two theorems collect these results, for orthogonal and general complete $\boldsymbol{A}_0$, respectively.

Theorem 4.1 (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 $\mu < \min\left\{c_a\theta n^{-1}, c_b n^{-5/4}\right\}$, whenever $\exp(n) \ge p \ge Cn^3 \log\frac{n}{\mu\theta}/(\mu^2\theta^2)$, with probability at least $1 - 8n^2p^{-10} - \theta(np)^{-7} - \exp\left(-0.3\theta np\right) - p^{-10} - c_c\exp\left(-c_d p\mu^2\theta^2/n^2\right)$, 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*

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

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., $\|\widehat{\boldsymbol{q}} - \boldsymbol{q}_\star\| \le \varepsilon$) in

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

iterations. Here $c_\star$ is as defined in Theorem 3.1, and c_a, c_b are the same numerical constants as defined in Theorem 3.1, c_c to c_i are other positive numerical constants.

Proofs for the complete case basically follows from that the slight perturbation of structure parameters as summarized in Theorem 3.3 (vs. Theorem 3.1) change all algorithm parameter by at most small multiplicative constants.

Theorem 4.2 (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 $\mu < \min\left\{c_a\theta n^{-1}, c_b n^{-5/4}\right\}$, whenever $\exp(n) \ge p \ge \frac{C}{c_\star^2\theta}\max\left\{\frac{n^4}{\mu^4}, \frac{n^5}{\mu^2}\right\}\kappa^8(\boldsymbol{A}_0)\log^4\left(\frac{\kappa(\boldsymbol{A}_0)n}{\mu\theta}\right)$, with probability at least $1 - 8n^2p^{-10} - \theta(np)^{-7} - \exp\left(-0.3\theta np\right) - 2p^{-8} - c_c\exp\left(-c_d p\mu^2\theta^2/n^2\right)$, the Riemannian trust-region algorithm with input data matrix $\overline{\boldsymbol{Y}} \doteq \sqrt{p\theta}\left(\boldsymbol{Y}\boldsymbol{Y}^*\right)^{-1/2}\boldsymbol{Y}$ where $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = \mathtt{SVD}\left(\boldsymbol{A}_0\right)$, any initialization $\boldsymbol{q}^{(0)}$ on the sphere and a step size satisfying*

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

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., $\|\widehat{\boldsymbol{q}} - \boldsymbol{q}_\star\| \le \varepsilon$) in

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

iterations. Here $c_\star$ is as defined in Theorem 3.1, and c_a, c_b are the same numerical constants as defined in Theorem 3.1, c_c to c_i are other positive numerical constants.

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

5. Main Results

For orthogonal dictionaries, from Theorem 3.1 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 **??**, 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. The same procedure can be used to recover complete dictionaries, though the analysis is slightly more complicated. 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 3.1 (resp. Theorem 3.3) and Theorem 4.1 (resp. Theorem 4.2), 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:

$$\min_{\boldsymbol{q}} \|\boldsymbol{q}^* \widehat{\boldsymbol{Y}}\|_1, \quad \text{s.t.} \quad \langle \boldsymbol{r}, \boldsymbol{q} \rangle = 1, \qquad (5.1)$$

   with $\boldsymbol{r} = \widehat{\boldsymbol{q}}$. We show 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 (5.1) 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 $\boldsymbol{X}_0$ ($1 \le \ell \le n-2$) have been recovered, say, by unit vectors $\boldsymbol{q}_\star^1, \ldots, \boldsymbol{q}_\star^\ell$, one takes an orthonormal basis $\boldsymbol{U}$ for $[\mathrm{span}\left(\boldsymbol{q}_\star^1, \ldots, \boldsymbol{q}_\star^\ell\right)]^\perp$, 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 (5.1) 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 $\boldsymbol{A}_0$.** By solving the linear system $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}_0$, one can obtain the dictionary $\boldsymbol{A}_0 = \boldsymbol{Y}\boldsymbol{X}_0^* \left(\boldsymbol{X}_0 \boldsymbol{X}_0^*\right)^{-1}$.

Formally, we have the following results:

Theorem 5.1 (Orthogonal Dictionary)**.** *Assume the dictionary $\boldsymbol{A}_0$ is orthogonal and we take $\widehat{\boldsymbol{Y}} = \boldsymbol{Y}$. Suppose $\theta \in (0, 1/3)$, $\mu_\star < \min\left\{c_a \theta n^{-1}, c_b n^{-5/4}\right\}$, and $p \ge C n^3 \log \frac{n}{\mu_\star \theta} / \left(\mu_\star^2 \theta^2\right)$. The above algorithmic pipeline with parameter setting*

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

recovers the dictionary $\boldsymbol{A}_0$ and $\boldsymbol{X}_0$ in polynomial time, with failure probability bounded by $c_e p^{-6}$. Here $c_\star$ is as defined in Theorem 3.1, and c_a through c_e, and C are all positive numerical constants.

Theorem 5.2 (Complete Dictionary)**.** *Assume the dictionary $\boldsymbol{A}_0$ is complete with condition number $\kappa(\boldsymbol{A}_0)$ and we take $\widehat{\boldsymbol{Y}} = \overline{\boldsymbol{Y}}$. Suppose $\theta \in (0, 1/3)$, $\mu_\star < \min\left\{c_a \theta n^{-1}, c_b n^{-5/4}\right\}$, and $p \ge \frac{C}{c_\star^2 \theta} \max\left\{\frac{n^4}{\mu^4}, \frac{n^5}{\mu^2}\right\} \kappa^8(\boldsymbol{A}_0) \log^4\left(\frac{\kappa(\boldsymbol{A}_0) n}{\mu\theta}\right)$. The algorithmic pipeline with parameter setting*

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

recovers the dictionary $\boldsymbol{A}_0$ and $\boldsymbol{X}_0$ in polynomial time, with failure probability bounded by $c_e p^{-6}$. Here $c_\star$ is as defined in Theorem 3.1, and c_a through c_f, and C are all positive numerical constants.

6. Numerical Results

To corroborate our theory, we experiment with dictionary recovery on simulated data [9] . 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 $p = 5n^2 \log(n)$, 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 (2.3), 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

[9] The code is available online: https://github.com/sunju/dl_focm

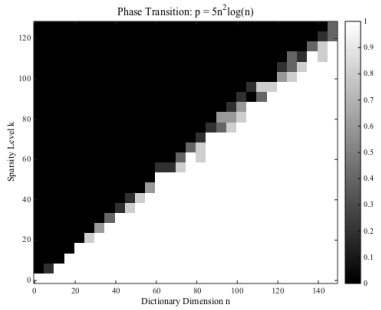

Figure 2. Phase transition for recovering a single sparse vector under the dictionary learning model with $p = 5n^2 \log n$.

dictionary dimension n and the sparsity k both between 1 and 150; 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 $\texttt{RE} = \min_{1\le i\le n} \left(\|\widehat{\boldsymbol{q}} - \boldsymbol{e}_i\|, \|\widehat{\boldsymbol{q}} + \boldsymbol{e}_i\|\right)$. The trial is determined to be a success once $\texttt{RE} \le \mu$, 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 2 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 $p \in O(n^2 \log n)$. Our analysis is tight up to logarithm factors, and also the polynomial dependency on $1/\mu$, which under the theory is polynomial in n.

7. Discussion

For recovery of complete dictionaries, the LP program approach in (Spielman et al., 2012) that works with $\theta \le O(1/\sqrt{n})$ only demands $p \ge \Omega(n^2 \log n^2)$, which is recently improved to $p \ge \Omega(n \log^4 n)$ (Luh & Vu, 2015), almost matching the lower bound $\Omega(n \log n)$ (i.e., when $\theta \sim 1/n$). The sample complexity stated in Theorem 5.2 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 $p \sim O(n^2 \log n)$ 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 $\boldsymbol{q}$ 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 (Edelman et al., 1998; Absil et al., 2009). 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. [10]

Under the i.i.d. BG coefficient model, our recovery problem is also an instance of the ICA problem. It is interesting to ask what is vital in making the problem tractable: sparsity or independence. The full version (Sun et al., 2015) includes an experimental study in this direction, which underlines the importance of the sparsity prior. In fact, the preliminary experiments there suggest the independence assumption we made here likely can be removed without losing the favorable geometric structures. In addition, the connection to ICA also suggests the possibility of adapting our geometric characterization and algorithms to the ICA problem. This likely will provide new theoretical insights and computational schemes to ICA.

In the surge of theoretical understanding of nonconvex heuristics (Keshavan et al., 2010; Jain et al., 2013; Hardt, 2014; Hardt & Wootters, 2014; Netrapalli et al., 2014; Jain & Netrapalli, 2014; Netrapalli et al., 2013; Candes et al., 2014; Jain & Oh, 2014; Anandkumar et al., 2014; Yi et al., 2013; Lee et al., 2013; Qu et al., 2014; Lee et al., 2013; Agarwal et al., 2013a;b; Arora et al., 2013; 2015; 2014), 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.

[10] This recent work (Arora et al., 2015) 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.

Acknowledgments

This work was partially supported by grants ONR N00014-13-1-0492, NSF 1343282, and funding from the Moore and Sloan Foundations and the Wei Family Private Foundation. We thank the area chair and the anonymous reviewers for making painstaking effort to read our long proofs and providing insightful feedback. We also thank Cun Mu and Henry Kuo for discussions related to this project.

References

Absil, P.-A., Baker, C. G., and Gallivan, K. A. Trust-region methods on riemannian manifolds. *Foundations of Computational Mathematics*, 7(3):303–330, 2007.

Absil, Pierre-Antoine, Mahoney, Robert, and Sepulchre, Rodolphe. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

Agarwal, Alekh, Anandkumar, Animashree, Jain, Prateek, Netrapalli, Praneeth, and Tandon, Rashish. Learning sparsely used overcomplete dictionaries via alternating minimization. *arXiv preprint arXiv:1310.7991*, 2013a.

Agarwal, Alekh, Anandkumar, Animashree, and Netrapalli, Praneeth. Exact recovery of sparsely used overcomplete dictionaries. *arXiv preprint arXiv:1309.1952*, 2013b.

Anandkumar, Animashree, Ge, Rong, and Janzamin, Majid. Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates. *arXiv preprint arXiv:1402.5180*, 2014.

Arora, Sanjeev, Ge, Rong, and Moitra, Ankur. New algorithms for learning incoherent and overcomplete dictionaries. *arXiv preprint arXiv:1308.6273*, 2013.

Arora, Sanjeev, Bhaskara, Aditya, Ge, Rong, and Ma, Tengyu. More algorithms for provable dictionary learning. *arXiv preprint arXiv:1401.0579*, 2014.

Arora, Sanjeev, Ge, Rong, Ma, Tengyu, and Moitra, Ankur. Simple, efficient, and neural algorithms for sparse coding. *arXiv preprint arXiv:1503.00778*, 2015.

Barak, Boaz, Kelner, Jonathan A, and Steurer, David. Dictionary learning and tensor decomposition via the sum-of-squares method. *arXiv preprint arXiv:1407.1543*, 2014.

Candes, Emmanuel, Li, Xiaodong, and Soltanolkotabi, Mahdi. Phase retrieval via wirtinger flow: Theory and algorithms. *arXiv preprint arXiv:1407.1065*, 2014.

Conn, Andrew R., Gould, Nicholas I. M., and Toint, Philippe L. *Trust-region Methods*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2000. ISBN 0-89871-460-5.

Demanet, Laurent and Hand, Paul. Scaling law for recovering the sparsest element in a subspace. *Information and Inference*, 3(4):295–309, 2014.

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

Elad, Michael. *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer, 2010.

Fortin, Charles and Wolkowicz, Henry. The trust region subproblem and semidefinite programming*. *Optimization methods and software*, 19(1):41–67, 2004.

Geng, Quan and Wright, John. 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.

Gribonval, Rémi and Schnass, Karin. Dictionary identification - sparse matrix-factorization via ℓ^1-minimization. *IEEE Transactions on Information Theory*, 56(7):3523–3539, 2010.

Gribonval, Rémi, Jenatton, Rodolphe, and Bach, Francis. Sparse and spurious: dictionary learning with noise and outliers. *arXiv preprint arXiv:1407.5155*, 2014.

Hardt, Moritz. Understanding alternating minimization for matrix completion. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pp. 651–660. IEEE, 2014.

Hardt, Moritz and Wootters, Mary. Fast matrix completion without the condition number. In *Proceedings of The 27th Conference on Learning Theory*, pp. 638–678, 2014.

Hazan, Elad and Koren, Tomer. A linear-time algorithm for trust region problems. *arXiv preprint arXiv:1401.6757*, 2014.

Hillar, Christopher and Sommer, Friedrich T. When can dictionary learning uniquely recover sparse data from subsamples? *arXiv preprint arXiv:1106.3616*, 2011.

Jain, Prateek and Netrapalli, Praneeth. Fast exact matrix completion with finite samples. *arXiv preprint arXiv:1411.1087*, 2014.

Jain, Prateek and Oh, Sewoong. Provable tensor factorization with missing data. In *Advances in Neural Information Processing Systems*, pp. 1431–1439, 2014.

Jain, Prateek, Netrapalli, Praneeth, and Sanghavi, Sujay. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of Computing*, pp. 665–674. ACM, 2013.

Keshavan, Raghunandan H, Montanari, Andrea, and Oh, Sewoong. Matrix completion from a few entries. *Information Theory, IEEE Transactions on*, 56(6):2980–2998, 2010.

Lee, Kiryung, Wu, Yihong, and Bresler, Yoram. Near optimal compressed sensing of sparse rank-one matrices via sparse power factorization. *arXiv preprint arXiv:1312.0525*, 2013.

Luh, Kyle and Vu, Van. Dictionary learning with few samples and matrix concentration. *arXiv preprint arXiv:1503.08854*, 2015.

Mairal, Julien, Bach, Francis, and Ponce, Jean. Sparse modeling for image and vision processing. *Foundations and Trends in Computer Graphics and Vision*, 8(2-3), 2014.

Mehta, Nishant and Gray, Alexander G. Sparsity-based generalization bounds for predictive sparse coding. *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 28(1):36–44, 2013.

Moré, J. J. and Sorensen, D. C. Computing a trust region step. *SIAM J. Scientific and Statistical Computing*, 4: 553–572, 1983.

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

Netrapalli, Praneeth, Jain, Prateek, and Sanghavi, Sujay. Phase retrieval using alternating minimization. In *Advances in Neural Information Processing Systems*, pp. 2796–2804, 2013.

Netrapalli, Praneeth, Niranjan, UN, Sanghavi, Sujay, Anandkumar, Animashree, and Jain, Prateek. Non-convex robust pca. In *Advances in Neural Information Processing Systems*, pp. 1107–1115, 2014.

Nocedal, Jorge and Wright, Stephen. *Numerical Optimization*. Springer, 2006.

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

Schnass, Karin. On the identifiability of overcomplete dictionaries via the minimisation principle underlying k-svd. *Applied and Computational Harmonic Analysis*, 37(3): 464–491, 2014a.

Schnass, Karin. Local identification of overcomplete dictionaries. *arXiv preprint arXiv:1401.6354*, 2014b.

Schnass, Karin. Convergence radius and sample complexity of itkm algorithms for dictionary learning. *arXiv preprint arXiv:1503.07027*, 2015.

Spielman, Daniel A, Wang, Huan, and Wright, John. Exact recovery of sparsely-used dictionaries. In *Proceedings of the 25th Annual Conference on Learning Theory*, 2012.

Sun, Ju, Qu, Qing, and Wright, John. Complete dictionary recovery over the sphere. *arXiv preprint arXiv:1504.06785*, 2015.

Vainsencher, Daniel, Mannor, Shie, and Bruckstein, Alfred M. The sample complexity of dictionary learning. *J. Mach. Learn. Res.*, 12:3259–3281, November 2011. ISSN 1532-4435.

Yi, Xinyang, Caramanis, Constantine, and Sanghavi, Sujay. Alternating minimization for mixed linear regression. *arXiv preprint arXiv:1310.3745*, 2013.

Zibulevsky, Michael and Pearlmutter, Barak. Blind source separation by sparse decomposition in a signal dictionary. *Neural computation*, 13(4):863–882, 2001.