Finding the Sparsest Vectors in a Subspace: Theory, Algorithms, and Applications

Qing Qu[1,*], Zhihui Zhu[2,3,*], Xiao Li[4], Manolis C. Tsakiris[5], John Wright[6], and René Vidal[2,7]

[1]*Center for Data Science, New York University*
[2]*Center for Imaging Science & Mathematical Institute for Data Science, Johns Hopkins University*
[3]*Ritchie School of Engineering and Computer Science, University of Denver*
[4]*Department of Electronic Engineering, Chinese University of Hong Kong*
[5]*School of Information Science and Technology, ShanghaiTech University*
[6]*Department of Electrical Engineering & Data Science Institute, Columbia University*
[7]*Department of Biomedical Engineering, Johns Hopkins University*

Abstract

Finding the sparsest vector (direction) in a low dimensional subspace can be considered a homogeneous variant of the sparse recovery problem, which finds applications in robust subspace recovery, dictionary learning, sparse blind deconvolution, and many other problems in signal processing and machine learning. However, in contrast to the classical sparse recovery problem, the most natural formulation for finding the sparsest vector in a subspace is usually nonconvex. In this tutorial, we provide a comprehensive review of recent advances on global nonconvex optimization theory for solving this problem, ranging from geometric analysis of the optimization landscapes, efficient nonconvex optimization algorithms, to applications in representation learning and imaging sciences. In particular, we show how can we design efficient optimization methods to solve the nonconvex problem to target solutions from a geometric perspective. Finally, we conclude this review by pointing out several interesting open problems for future research.

I. Introduction

Nonconvex optimization problems are *ubiquitous* in signal processing and machine learning [1, 2]. However, for general nonconvex problems, even finding a local minimizer is a NP-hard [3] problem – nevermind the global optimal solutions. While one may consider convex relaxations [4–7] and resort to the rich literature of convex optimization [8, 9], it usually scales poorly with respect to the dimension of the data, and often provably fails for problems with nonlinear models. Nonetheless, recent advances reveal that optimization landscapes of nonconvex problems in practice often have *benign* geometric properties. These examples include phase retrieval [10–13], phase synchronization [14, 15], blind deconvolution [16, 17], dictionary learning [18, 19], matrix factorization [20–28], and tensor decomposition [29], and more. The underlying benign geometric structure ensures fast convergence of iterative algorithms to target solutions:

1) *Benign Local Geometry.* In many cases, there often exists a sufficiently large *basin of attraction* around the target solutions, within which a local-descent algorithm converges rapidly to the solution;
2) *Benign Global Geometry.* Problem specific symmetry structures induce *benign global optimization landscape*, that there are *no* flat saddle points or spurious local minima (see figures above), ensures global convergence of iterative algorithms with even *arbitary* initializations [29–33].

Under this framework, we provide a comprehensive review of recent advances on nonconvex optimization methods for *finding the sparsest vectors in a subspace* [34–37]. Namely, given data $\boldsymbol{Y} \in \mathbb{R}^{n\times p}$ whose rows

* Both authors contribute equally to this work.

(a) Flat saddle points **(b)** Spurious local minimizers

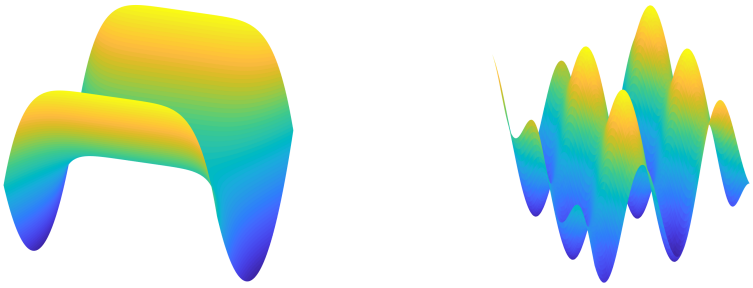


Fig. 1: Worst case scenarios in nonconvex optimization.

form a n-dimensional subspace $\mathcal{S} \subseteq \mathbb{R}^p$ ($n \ll p$), can we *efficiently* find the *sparsest* nonzero vector in $\mathcal{S}$ (up to scaling)? Mathematically speaking, can we efficiently solve

$$\min_{\boldsymbol{q}} \left\| \boldsymbol{Y}^\top \boldsymbol{q} \right\|_0, \qquad \text{s.t.} \quad \boldsymbol{q} \neq \boldsymbol{0}, \tag{I.1}$$

so that $\boldsymbol{Y}^\top \boldsymbol{q}$ is the sparsest vector[1] in $\mathcal{S} = \mathrm{row}(\boldsymbol{Y})$? Here, the nonzero constraint $\boldsymbol{q} \neq \boldsymbol{0}$ is to avoid the *trivial* sparse solution $\boldsymbol{q} = \boldsymbol{0}$ simply because a subspace $\mathcal{S}$ passes through the origin $\boldsymbol{0}$. In the meanwhile, it should be noted that the problem can also be considered as a homogeneous variant of sparse recovery problem [4, 38], in the sense that the problem (I.1) can be *equivalently* formulated as

$$\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_0, \qquad \text{s.t.} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}, \quad \boldsymbol{x} \neq \boldsymbol{0}, \tag{I.2}$$

where rows of $\boldsymbol{A} \in \mathbb{R}^{(p-n)\times p}$ form a basis of the orthogonal complement of $\mathcal{S}$ so that (I.2) can be reviewed as a *dual* formulation of (I.1). However, in contrast to the classical sparse recovery problem which finds the sparsest vector with $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ and $\boldsymbol{b} \neq \boldsymbol{0}$ [4, 38–41], solving (I.2) has caught less attention and has not been well-studied albeit its importance in many applications in signal processing and machine learning as we discuss in Section II. One major reason is due to our limited understandings on the computational properties of solving the nonconvex problem (I.2). Different from classical sparse recovery problems, where convex relaxations perform near optimally for broad classes of designs of $\boldsymbol{A}$ [42, 43], it has been known for decades that the basic problem (I.2) is NP-hard for an arbitrary subspace $\mathcal{S}$ [44, 45]. Even we relax the ℓ^0 objective with its convex surrogates, the nonzero constraint $\boldsymbol{x} \neq \boldsymbol{0}$ still makes the problem inherently nonconvex. It is only very recently that efficient computational surrogates with nontrivial recovery guarantees have been discovered and studied for specific instances [34, 46–52]. In this paper, we survey several important aspects of recent advances on nonconvex optimization methods for solving the problem of finding the sparsest vector in a subspace, ranging from landscape analysis, efficient optimization methods, to applications.

Paper organization. The rest of the paper is organized as follows. In Section II, we show that several fundamental problems in signal processing and machine learning can be reduced to the task of finding the sparsest vector in a subspace. In Section III, we introduce natural nonconvex relaxations of (I.1) with computational guarantees. In Section IV we provide a systematic overview of geometric analysis on nonconvex optimization landscapes, based on which nonconvex algorithms have recently led to efficient solutions and new performance guarantees that we discuss in Section V. We demonstrate the broad applications in Section VI. Finally, we close this review by discussing several open problems in Section VII.

II. Motivations

Variants of the task of finding the sparsest vector in a subspace take several forms in many applications of modern signal processing and machine learning. In this section, we survey several fundamental problems to

[1]Here, $\mathrm{row}(\boldsymbol{Y})$ denotes the row subspace of $\boldsymbol{Y}$, i.e., the subspace $\mathrm{row}(\boldsymbol{Y})$ is spanned by row vectors of $\boldsymbol{Y}$.

demonstrate its importance, where all the problems can be reduced to solving (I.1), with different structures of the subspace $\mathcal{S} = \mathrm{row}(\boldsymbol{Y})$.

a) Robust subspace recovery [6, 53–56]. Fitting a linear subspace to dataset corrupted by outliers is a fundamental problem in machine learning and statistics, primarily known as robust principal component analysis (PCA) [57], or robust subspace recovery (RSR) [56]. Given the dataset $\boldsymbol{Y}$ of form the form

$$\underset{\text{data}}{\boldsymbol{Y}} = \Big[\underset{\text{inliers}}{\boldsymbol{X}} \quad \underset{\text{outliers}}{\boldsymbol{O}}\Big] \underset{\text{permutation}}{\boldsymbol{\Gamma}} \in \mathbb{R}^{n\times p}, \tag{II.1}$$

where the columns of $\boldsymbol{X} \in \mathbb{R}^{n\times p_1}$ form inlier points spanning a subspace $\mathcal{S}_{\boldsymbol{X}}$, the columns of $\boldsymbol{O} \in \mathbb{R}^{n\times p_2}$ are outlier points with no linear structure, and $\boldsymbol{\Gamma}$ is an unknown permutation, the goal is to recover the inlier subspace $\mathcal{S}_{\boldsymbol{X}}$, or equivalently to cluster the points into inliers and outliers. It is well-known that the presence of outliers can severely affect the quality of the solutions obtained by the classical PCA approach [57]. This challenge can be conquered by finding the sparsest vector in[2] $\mathcal{S} = \mathrm{row}(\boldsymbol{Y})$ via solving (I.1), which returns a normal vector[3] of the subspace $\mathcal{S}_{\boldsymbol{X}}$ [51], producing a hyperplane containing all columns of $\boldsymbol{X}$. This approach is called *dual principal component pursuit* (DPCP) [51, 52, 58], which can be reviewed as a *dual* method of classical ways of solving robust subspace recovery problems [56]. The DPCP has led to new recovery guarantees, which can deal with more number of outliers than traditional methods [51, 52, 58].

b) Learning sparsely-used dictionaries [39, 59–61]. Dictionary learning (DL) aims to learn the underlying compact representation from the data $\boldsymbol{Y}$, which finds many applications in signal/imaging processing, machine learning, and computer vision [60–64]. Mathematically speaking, the problem is to factorize the data

$$\underset{\text{data}}{\boldsymbol{Y}} = \underset{\text{dictionary}}{\boldsymbol{A}} \; \underset{\text{sparse code}}{\boldsymbol{X}}. \tag{II.2}$$

into a compact representation dictionary $\boldsymbol{A}$ and sparse coefficient matrix $\boldsymbol{X}$. Such representations naturally allow signal compression [60], and also facilitate efficient signal acquisition [65], denoising [62], and classification [66] (see relevant discussion in [64]). In particular, when the dictionary $\boldsymbol{A}$ is *complete*[4], the authors [34, 46, 47] showed that the DL problem can be reduced to finding the sparsest vector in the subspace $\mathcal{S} = \mathrm{row}(\boldsymbol{Y})$: by solving (I.1), presumably the solution $\boldsymbol{Y}^\top \boldsymbol{q}$ returns one row of the sparse matrix $\boldsymbol{X}$. Based on this, the full matrices $(\boldsymbol{A}, \boldsymbol{X})$ can be recovered via extra techniques such as deflation [46]. For complete DL, this approach has led to new theoretical and algorithmic advances [46, 47, 67–69].

c) Sparse blind deconvolution [70–74]. Sparse blind deconvolution is a classical inverse problem that ubiquitously appears in various areas of digital communication [75], signal/image processing [76, 77], neuroscience [78, 79], geophysics [80], and more. Given multiple measurements $\{\boldsymbol{y}_i\}_{i=1}^p$ in the form of the circulant convolution

$$\underset{\text{measurements}}{\boldsymbol{y}_i} = \underset{\text{kernel}}{\boldsymbol{a}_0} \circledast \underset{\text{sparse signal}}{\boldsymbol{x}_i}, \qquad 1 \le i \le m, \tag{II.3}$$

the *multichannel sparse blind deconvolution* (MCS-BD) problem [48, 50, 74, 81] aims to simultaneously recover the unknown kernel $\boldsymbol{a}_0$ and sparse signals $\{\boldsymbol{x}_i\}_{i=1}^p$. Notice that the circulant convolution (II.3) can be rewritten in the matrix-vector form with[5] $\boldsymbol{C}_{\boldsymbol{y}_i} = \boldsymbol{C}_{\boldsymbol{a}} \boldsymbol{C}_{\boldsymbol{x}_i}$. Thus, by concatenating all the measurements, we can write the problem in a similar form of complete DL in the sense that

$$\underbrace{\big[\boldsymbol{C}_{\boldsymbol{y}_1} \quad \cdots \quad \boldsymbol{C}_{\boldsymbol{y}_p}\big]}_{\boldsymbol{Y}} = \boldsymbol{C}_{\boldsymbol{a}_0} \underbrace{\big[\boldsymbol{C}_{\boldsymbol{x}_1} \quad \cdots \quad \boldsymbol{C}_{\boldsymbol{x}_p}\big]}_{\boldsymbol{X}}.$$

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

[3]A normal vector of a subspace is a nonzero vector that is orthogonal to all points in the subspace.

[4]Complete means that the dictionary $\boldsymbol{A}$ is square and invertible. For a proper conditioned dictionary, the complete DL can be approximately reduced to orthogonal DL via preconditioning or whitening of the data [47].

[5]Here, any vector $\boldsymbol{v} \in \mathbb{R}^n$, we use $\boldsymbol{C}_v \in \mathbb{R}^{n\times n}$ to denote corresponding circulant matrix generated from $\boldsymbol{v}$.

TABLE I: Summary of Convex Surrogates $\varphi(\cdot)$ for ℓ^0-norm

Name	Objective $\varphi(\cdot)$	(Sub)gradient $\nabla\varphi(\cdot)$	Smoothness
ℓ_1-norm [52, 69]	$\sum_k \lvert z_k\rvert$	$\operatorname{sign}(\boldsymbol{z})$	nonsmooth
Huber loss[7] [81]	$\min_{\boldsymbol{v}} \frac{1}{2}\lVert\boldsymbol{z}-\boldsymbol{v}\rVert^2 + \mu\lVert\boldsymbol{v}\rVert_1$	$\boldsymbol{z}/\mu\mathbb{1}_{\lvert\boldsymbol{z}\rvert<\mu} + \operatorname{sign}(\boldsymbol{z})\mathbb{1}_{\lvert\boldsymbol{z}\rvert\geq\mu}$	$\mathcal{C}^1$ smooth
pseudo-Huber [71]	$\mu\sum_k \sqrt{1+(z_k/\mu)^2}$	$\boldsymbol{z}/\sqrt{\boldsymbol{z}+\mu^2}$	$\mathcal{C}^\infty$ smooth
Logcosh [47, 74]	$\sum_k \mu\log\cosh(z_k/\mu)$	$\tanh(\boldsymbol{z}/\mu)$	$\mathcal{C}^\infty$ smooth

When the kernel $\boldsymbol{a}_0$ is invertible[6], per our discussion for complete DL, this implies that we can solve the MCS-BD problem by finding the sparsest vector in $\mathcal{S} = \operatorname{row}(\boldsymbol{Y})$ as well. This discovery has recently led to new guaranteed, efficient methods for solving MCS-BD under general settings [50, 74, 81].

d) Other problems. Variants and generalizations of finding the sparsest vectors in a subspace problem also appear in orthogonal ℓ_1 regression [82], sparse PCA [83, 84], numerical linear algebra [45, 85, 86], applications regarding control and optimization [87], nonrigid structure from motion [88], spectral estimation and Prony's method [89], blind source separation [90], graphical model learning [91], and sparse coding on manifolds [92]. Nonetheless, we believe the potential of seeking sparse/structured element in a subspace is still largely unexplored, in spite of the cases we discussed here. We hope this review will bring more attention to this problem and inspire further application ideas of recent theoretical and algorithmic advances.

III. Problem Formulation

Per our discussion in Section I, to find the sparsest vectors in $\mathcal{S} = \operatorname{row}(\boldsymbol{Y})$ the vanilla formulation (I.1) (or equivalently (I.2)) is NP-hard to solve. Therefore, we need to resort to certain relaxations of the problem such that the new problem is substantial easier to optimize and its global solutions are still *close* to the expected target solutions. Similar to the idea of solving the sparse recovery problem [4, 42], one natural idea is to replace ℓ^0-norm with any sparsity promoting convex surrogate $\varphi(\cdot)$ (see Table I for an illustration, we will discuss the choices in Section V). However, the nonconvex constraint $\boldsymbol{q} \neq \boldsymbol{0}$ still makes the problem inherently difficult to optimize. Nonetheless, since we only hope to find the sparsest vector up a scaling, it is natural to consider replacing $\boldsymbol{q} \neq \boldsymbol{0}$ by certain unit norm constraints on $\boldsymbol{q}$.

Limitation of convex relaxations. In the context of the dictionary learning problem, Spielman et al. [46] considers ℓ^1-minimization constrained with $\lVert\boldsymbol{q}\rVert_\infty = 1$, introducing a convex relaxation of (I.1) with a sequence of linear programs:

$$\ell^1/\ell^\infty \text{ Relaxation:} \qquad \min_{\boldsymbol{q}} \ \varphi\left(\boldsymbol{Y}^\top\boldsymbol{q}\right), \quad \text{s.t.} \quad q(i) = 1, \quad 1 \leq i \leq n, \tag{III.1}$$

where here $\varphi(\cdot) = \lVert\cdot\rVert_1$ as shown in Table I. When $\varphi(\cdot) = \lVert\cdot\rVert_1$, they showed that the solutions of (III.1) are exactly the target sparse vectors up to a scaling when the subspace $\mathcal{S}$ is spanned by a set of random sparse basis vectors. However, this result provably breaks down merely when the sparsity level of each base vector is beyond[8] $\theta \in O\left(1/\sqrt{n}\right)$, while convex relaxation for standard sparse approximation problem can handle much higher sparsity levels $\theta \in \mathcal{O}(1)$ [42, 43]. For the problem (III.1), the same sparsity threshold is also observer for a simpler *planted sparse vector* (PSV) model, where there is a single sparse vector embedded in an otherwise random subspace $\mathcal{S}$ [36]. Moreover, for both models the most natural semidefinite programming (SDP) relaxation [34] also breaks down at exactly the same threshold[9]. Unfortunately, numerical simulations

[6] In other words, we assume that its circulant matrix $\boldsymbol{C}_{\boldsymbol{a}_0}$ is invertible.

[7] The Huber loss can also be written in the form $\varphi(\boldsymbol{z}) = \sum_k H_\mu(z_k)$ with $H_\mu(z) = \begin{cases} |z| & |z| \geq \mu \\ \frac{z^2}{2\mu} + \frac{\mu}{2} & |z| < \mu \end{cases}$.

[8] Here, θ denotes the probability of one entry being nonzero.

[9] This breakdown behavior is again in sharp contrast to the standard sparse approximation problem (with $\boldsymbol{b} \neq \boldsymbol{0}$), in which it is possible to handle very large fractions of nonzeros (say, $\theta = \Omega(1/\log n)$, or even $\theta = \Omega(1)$) using a very simple ℓ^1 relaxation [42, 43]

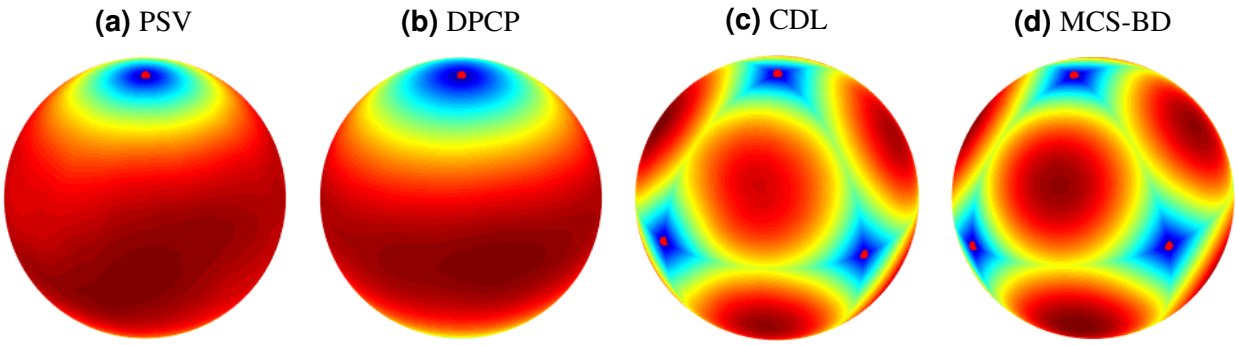


Fig. 2: Plots of optimization landscapes of (III.2) over the sphere for different problems in 3D. From the left to right: (a) planted sparse vector [34], (b) robust subspace recovery [51], (c) orthognal DL [47], and (d) sparse blind deconvolution [50, 81]. The colder color value means smaller function value, and vice versa. The red dots correspond to the target solutions.

confirm that these results are essentially sharp, so that one might naturally question: *is* $\theta \in O\left(1/\sqrt{n}\right)$ *simply the best we can do with efficient, guaranteed algorithms?*

Remarkably, this is not the case. Recently, Barak et al. introduced a new rounding technique for sum-of-squares (SoS) relaxations, showing that the sparsest vector can be recovered even when $\theta = \Omega(1)$ [93]. Unfortunately, the runtime of this approach is a high-degree polynomial in data dimension so that the result is mostly of theoretical interest. Therefore, the legitimate question still remains: *Is there a practical algorithm that provably recovers a sparse vector with* $\theta = \Omega(1)$ *portion of nonzeros from a generic subspace* $\mathcal{S}$*?*

Efficient solutions via nonconvex optimization. This question is addressed by recent advances on nonconvex optimization, where Qu et al. [34] first considered a *nonconvex* relaxation of (I.1),

$$\ell^1/\ell^2 \text{ Relaxation:} \qquad \min_{\boldsymbol{q}} \; f(\boldsymbol{q}) \;:=\; \varphi\left(\boldsymbol{Y}^\top \boldsymbol{q}\right), \qquad \text{s.t.} \quad \|\boldsymbol{q}\| \;=\; 1, \tag{III.2}$$

which relaxes $\boldsymbol{q} \neq \boldsymbol{0}$ by a *nonconvex* spherical constraint $\boldsymbol{q} \in \mathbb{S}^{n-1}$. Intuitively, the sphere $\mathbb{S}^{n-1}$ is a homogeneous manifold so that it could potentially handle higher sparsity levels. Indeed, for an idealized PSV model, Qu et al. showed that there is a simple, efficient nonconvex optimization method that provably recovers the sparsest vector even with $\theta = \Omega(1)$, breaking the $\theta \in O\left(1/\sqrt{n}\right)$ sparsity barrier [34]. Subsequently, Sun et al. showed that the same sparsity level can also be achieved with efficient methods for complete DL [47, 67]. Inspired by these results, optimizing variants of the nonconvex formulation (III.2) has led to new performance guarantees in robust subspace recovery [51, 52] and sparse blind deconvolution [50, 74, 81]. Nonetheless, as the problem formulation in (III.2) is nonconvex, it naturally raises the following question: *what are the underlying principles for efficiently solving these nonconvex problems to target solutions?*

IV. Geometry and Optimization Landscapes

In the following, we demystify the recent success of nonconvex approaches, by reviewing recent advances on geometric studies of the nonconvex optimization landscapes, towards providing a unified view for solving a broader class of nonconvex optimization problems. Correspondingly, in the next section (Section V) we show how to exploit these benign geometric properties plus extra ingredients to develop efficient nonconvex optimization methods, efficiently solving (III.2) to target (global) solutions.

A. Some Basic Facts

Basic notations. First, we introduce some basic notations for studying the global optimization properties. Let $\mathcal{Q}_\star \subset \mathbb{S}^{n-1}$ be the set of target solutions. To measure the distance between a vector $\boldsymbol{q} \in \mathbb{S}^{n-1}$ and the set $\mathcal{Q}_\star$, we introduce the following metric defined in the Eucildean space

$$\operatorname{dist}(\boldsymbol{q}, \mathcal{Q}_\star) \;:=\; \inf_{\boldsymbol{a} \in \mathcal{Q}_\star} \|\boldsymbol{q} - \boldsymbol{a}\| \,.$$

Accordingly, we define the set

$$\mathcal{B}(\boldsymbol{q}, \mathcal{Q}_\star) \;:=\; \left\{\boldsymbol{q} \in \mathbb{S}^{n-1} \;\middle|\; \operatorname{dist}(\boldsymbol{q}, \mathcal{Q}_\star) \le \varepsilon\right\}$$

that contains all the points on the sphere that are ε-close to $\mathcal{Q}_\star$.

Riemannian derivatives. Since we are solving a nonconvex optimization problem (III.2) that is constrained over a Riemannian manifold $\mathbb{S}^{n-1}$, to study the geometric properties of optimization landscape, we need formal definitions of the *slope* (gradient) and *curvature* (Hessian) of $f(\cdot)$ over the manifold. The sphere $\mathbb{S}^{n-1}$ is an embedded smooth manifold in $\mathbb{R}^n$; its *tangent space* at a point $\boldsymbol{q} \in \mathbb{S}^{n-1}$ can be identified with $\boldsymbol{q}^\perp$

$$\mathrm{T}_{\boldsymbol{q}}\mathbb{S}^{n-1} \;=\; \left\{\, \boldsymbol{v} \in \mathbb{R}^n \;\middle|\; \boldsymbol{q}^\top \boldsymbol{v} \;=\; 0 \right\}.$$

Thus, the projection onto the tangent space is given by $\mathcal{P}_{\boldsymbol{q}^\perp} = \boldsymbol{I} - \boldsymbol{q}\boldsymbol{q}^\top$. If f is smooth, the slope of $f(\cdot)$ over the sphere (formally, the Riemannian gradient) is defined in the tangent space $\mathrm{T}_{\boldsymbol{q}}\mathbb{S}^{n-1}$, which is simply the component of the standard (Euclidean) gradient $\nabla f(\boldsymbol{q})$ that is tangent to the sphere:

$$\operatorname{grad}[f](\boldsymbol{q}) \;=\; \mathcal{P}_{\boldsymbol{q}^\perp} \nabla f(\boldsymbol{q}).$$

For $f(\boldsymbol{q}) = \left\|\boldsymbol{Y}^\top \boldsymbol{q}\right\|_1$ that is nonsmooth with a subgradient[10] $\partial f(\boldsymbol{q}) = \boldsymbol{Y} \operatorname{sign}(\boldsymbol{Y}^\top \boldsymbol{q})$ where sign is an element-wise operator that takes the sign of the input if it is non-zero and sets 0 if the input is 0, we can similarly introduce the corresponding Riemannian subgradient[11]

$$\partial_{\mathrm{R}} f(\boldsymbol{q}) \;=\; \mathcal{P}_{\boldsymbol{q}^\perp} \partial f(\boldsymbol{q}).$$

On the other hand, if $f \in \mathcal{C}^2$, the curvature of $f(\cdot)$ over the sphere is slightly more involved. For any direction $\boldsymbol{\delta} \in \mathrm{T}_{\boldsymbol{q}}\mathbb{S}^{n-1}$, the second derivative of $f(\cdot)$ at point $\boldsymbol{q} \in \mathbb{S}^{n-1}$ along the geodesic curve[12] (great circle) is given by $\boldsymbol{\delta}^\top \operatorname{Hess}[f](\boldsymbol{q})\boldsymbol{\delta}$, where $\operatorname{Hess}[f](\boldsymbol{q})$ is the *Riemannian Hessian*

$$\operatorname{Hess}[f](\boldsymbol{q}) \;=\; \mathcal{P}_{\boldsymbol{q}^\perp}\Big(\underbrace{\nabla^2 f(\boldsymbol{q})}_{\text{curvature of } f(\cdot)} \;-\; \underbrace{\langle \boldsymbol{q}, \nabla f(\boldsymbol{q})\rangle \, \boldsymbol{I}}_{\text{curvature of the manifold}} \Big)\mathcal{P}_{\boldsymbol{q}^\perp}.$$

This expression contains two terms: (i) the first is the standard (Euclidean) hessian $\nabla^2 f$, which accounts for the curvature of the objective function f; (ii) the second term accounts for the curvature of the sphere itself. Thus, analogous to the case in Euclidean space, critical points can be characterized by $\operatorname{grad}[f](\boldsymbol{q}) = \boldsymbol{0}$ or $\boldsymbol{0} \in \partial_{\mathrm{R}} f(\boldsymbol{q})$; curvature can be studied through the eigenvalues of $\operatorname{Hess}[f](\boldsymbol{q})$.

B. Local Geometry: Basins of Attraction Around Target Solutions

At the nascent of theoretical understandings of global nonconvex optimization in early 2010s, in most cases people tend to believe nonconvex problems *only* have benign local geometric structures. It can be showed that there usually exist *local basins of attraction* around the target solutions, in the sense that the function either has local *strong convexity* or it satisfies certain *regularity condition* around the target solutions. Therefore, to have guaranteed global optimization, people developed data-driven initialization by using spectral methods to initialize into the local basin such that descent methods efficiently converge to the target solutions. The initialization plus local algorithmic analysis has led to global guarantees for several important problems in signal processing and machine learning, such as generalized phase retrieval [10, 96], low rank matrix recovery [28], tensor decomposition [97], and blind deconvolution with subspace model [98], and more.

[10] Its subdifferential which includes all the subgradients is given by $\boldsymbol{Y}\operatorname{Sign}(\boldsymbol{Y}^\top \boldsymbol{q})$, where Sign is an element-wise operator with $\operatorname{Sign}(a) = \operatorname{sign}(a)$ if $a \neq 0$ and $\operatorname{Sign}(a) = [-1, 1]$ if $a = 0$.

[11] We refer to [94] for a formal defintion of Riemannian subgradient. For a general nonsmooth function, the projection (onto the tangent space) of a subgradiment may not be a Riemannian subgradient. Fortunately, for problem that is regular (such as $f(\boldsymbol{q}) = \left\|\boldsymbol{Y}^\top \boldsymbol{q}\right\|_1$), according to [94], the Riemannian subgradient can be simply introduced by the projection of a subgradient.

[12] A geodesic curve is the shortest path connecting two points on the manifold, which can be parameterized by an exponential map $\exp_{\boldsymbol{q}}(t\delta)$. We refer readers to [95] for more technical details.

In the context of finding the sparsest vector in a subspace, for functions f that are $\mathcal{C}^\infty$ smooth, Sun et al. [47] and Li et al. [50] showed that the function is *locally* strongly convex tangent to $\boldsymbol{q}$, in the sense that

$$\mathrm{Hess}[f](\boldsymbol{q}) \;\succeq\; \alpha \cdot \boldsymbol{P}_{\boldsymbol{q}^\perp}, \quad \forall \boldsymbol{q} \in \mathcal{B}\left(\mathcal{Q}_\star, \epsilon_1\right). \tag{IV.1}$$

However, the regions that satisfy strong convexity are usually quite small (i.e., ϵ_1 is small), that they only cover a small measure of the sphere. For problems like complete DL [47] and sparse blind deconvolution [50], it is often very difficult to initialize into the region $\mathcal{B}\left(\mathcal{Q}_\star, \epsilon_1\right)$. Moreover, the strong convexity condition also needs the function to be at least $\mathcal{C}^2$ smooth, which is quite stringent.

A more general local condition is the so-called *regularity condition*, which often ensures local convergences of descent methods within a region of much larger radius. For instance, for a consideration of nonsmooth f, Bai et al. [69] and Zhu et al. [99] showed the following regularity condition[13]

$$\langle \boldsymbol{q} - \mathcal{P}_{\mathcal{Q}_\star}(\boldsymbol{q}), \partial_{\mathrm{R}} f(\boldsymbol{q}) \rangle \;\geq\; \alpha \,\mathrm{dist}(\boldsymbol{q}, \mathcal{Q}_\star), \qquad \forall\; \boldsymbol{q} \in \mathcal{B}(\mathcal{Q}_\star, \epsilon_2). \tag{IV.2}$$

As illustrated in Figure 3, the condition (IV.2) shows that the negative direction of the chosen Riemannian subgradient $\partial_{\mathrm{R}} f(\boldsymbol{q})$ is aligned with the direction $\boldsymbol{q} - \mathcal{P}_{\mathcal{Q}_\star}(\boldsymbol{q})$ pointing to the target solutions. In other words, this regularity condition will force the trajectory of (sub)gradient iterates getting closer to the target solutions when the step size is chosen appropriately, which we will discuss in more details in Section V. Moreover, (IV.2) also indicates a lower bound[14] for the Riemannian subgradient $\|\partial_R f(\boldsymbol{q})\| \geq \alpha$ for all $q \in \mathcal{B}(\mathcal{Q}_\star, \epsilon_2) \setminus \mathcal{Q}_\star$. Thus, if (IV.2) holds for all the Riemannian subgradients of any $q \in \mathcal{B}(\mathcal{Q}_\star, \epsilon_2)$—which is true for both orthogonal dictionary learning and DPCP problems—then one can conclude that there is no critical point other than the target solutions $\mathcal{Q}_\star$ in $\mathcal{B}(\mathcal{Q}_\star, \epsilon_2)$. This property further implies the possibility of finding a target solution by not only the Riemannian subgradient but also many other local iterative algorithms (which will be described in Section V) as long as they are initialized properly and converge to a critical point.

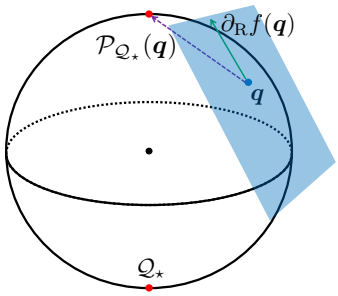


Fig. 3: Illustration of Equation (IV.2). Red nodes denote $\mathcal{Q}_\star$, with the top one closest to $\mathcal{Q}_\star$. Inequality (IV.2) requires the angle between $\mathcal{P}_{\mathcal{Q}^\star}(\boldsymbol{q}) - \boldsymbol{q}$ (purple arrow) and $-\partial_R f(\boldsymbol{q})$ (blue arrow) to be sufficiently small.

Minimal Example I: Robust Subspace Recovery. In the context of finding the sparsest vector in a subspace, we use the robust subspace recovery problem as an example to elaborate on the regularity condition (IV.2). As illustrated in Section II, given the dataset $\boldsymbol{Y}$ corrupted by outliers as $\boldsymbol{Y} = \begin{bmatrix}\boldsymbol{X} & \boldsymbol{O}\end{bmatrix} \boldsymbol{\Gamma}$, where $\boldsymbol{X} \in \mathbb{R}^{n \times p_1}$ are inliers generated from a subspace $\mathcal{S}_{\boldsymbol{X}}$, $\boldsymbol{O} \in \mathbb{R}^{n \times p_2}$ are outliers with no linear structure, and $\boldsymbol{\Gamma}$ is an unknown permutation matrix, the goal is to recover the underlying inlier subspace $\mathcal{S}_{\boldsymbol{X}}$. Noting that estimating $S_{\boldsymbol{X}}$ is equivalent to finding its orthogonal complement $\mathcal{S}_{\boldsymbol{X}}^\perp$, the DPCP approach [100] attempts to find one basis vector $\boldsymbol{q} \in \mathcal{S}_{\boldsymbol{X}}^\perp$ in each time. Once one basis vector is founded, we can then find another basis vector by removing the contribution from the previous one and repeat this process until finding all the basis vectors for $\mathcal{S}_{\boldsymbol{X}}^\perp$.

Recall that if $\boldsymbol{q}$ is in the orthogonal complement subspace $\mathcal{S}_{\boldsymbol{X}}^\perp$, then it is at least orthogonal to the n inliers $\boldsymbol{X}$. This motivates us to find such a basis $\boldsymbol{q} \in \mathcal{S}_{\boldsymbol{X}}^\perp$ by seeking a vector that is orthogonal to as many data points in $\boldsymbol{Y}$ as possible (i.e., the sparsest vector in $\boldsymbol{Y}^\perp$), resulting in (III.2) with φ being the ℓ_1-norm [52, 58, 100]. In this case, since the goal of DPCP is to compute a basis for $\mathcal{S}_{\boldsymbol{X}}^\perp$, the set of target solutions is $\mathcal{Q}_\star = \mathcal{S}_{\boldsymbol{X}}^\perp \cap \mathbb{S}^{n-1}$. With this choice of $\mathcal{Q}_\star$, it is proved in [99] that the DPCP problem (III.2) satisfies the regularity condition[15] (IV.2) with positive α and sufficiently large ϵ that depend on the number of

[13]The consideration of nonsmooth objective is only for the ease to resort to existing results [69, 99], and the simplicity of presenting the regularity condition. For smooth objectives,

[14]This lower bound can be obtained by applying the Cauchy-Schwartz inequality to the left hand side of (IV.2).

[15]This underlying regularity condition has been implicitly explored in [52, 100] in convergence analysis. Similar regularity condition has also been exploited in [101].

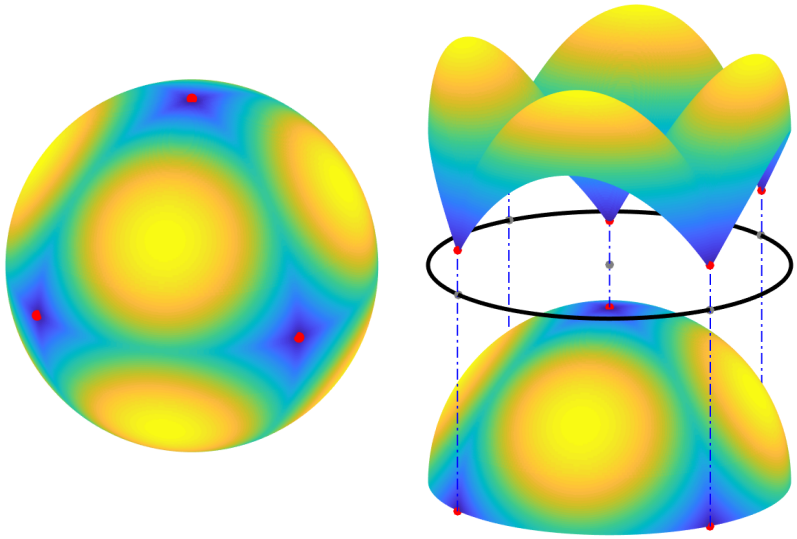


Fig. 4: **Orthogonal DL with permutation symmetry.** Left: $\varphi(\boldsymbol{Y}^\top \boldsymbol{q})$ as a function on the sphere $\mathbb{S}^2$. Local minimizers (red) are signed standard basis vectors $\mathcal{Q}_\star = \{\pm \boldsymbol{e}_i\}$. These are the maximally sparse vectors on $\mathbb{S}^2$. Right: graph of $\varphi(\boldsymbol{Y}^\top \boldsymbol{q})$ reparameterized into Euclidean space; notice the strong negative curvature at points that are not sparse.

inlier points p_1 and outlier points p_2 and how well distributed the inliers and outliers are. Moreover, starting from a spectral initialization (i.e., the eigenvector of $\boldsymbol{Y}\boldsymbol{Y}^\top$ corresponding to the smallest eigenvalue) that falls in $\mathcal{B}(\mathcal{Q}_\star, \epsilon)$, a basis vector for $\mathcal{S}_{\boldsymbol{X}}^{\perp}$ can be efficiently obtained by iterative algorithms which will be described in Section V.

Furthermore, this type of regularity condition also exist for other subspace models. For the orthogonal dictionary learning, Gilboa et al. [68] and Bai et al. [69] showed that random initialization falls into regions satisfies (IV.2) with constant probability, ensuring fast convergence of gradient methods. This result is later extended to the sparse blind deconvolution problems [74, 81], where similar results are established. Finally, it should be noted that all these convergence guarantees are based on this underlying geometric property, that we will discuss in more detail about exploiting these properties for algorithmic design in Section V.

C. Global Geometry: Negative Curvature Near Saddles

More surprisingly, more recent work [11, 28, 29, 47, 50] showed that in many cases nonconvex problems even have benign *global* geometric structures (see Figure 4 for an example), in the sense that

- There is *no* spurious local minimizer. (All) minimizers are (approximately) symmetric versions of the ground truth, and the optimization landscape around them exhibits local strong convexity or certain regularity properties that we discussed previously.
- There is *no* flat saddle points. All saddles are created by symmetric superposition of the target solutions, and they exhibit negative curvature[16] in symmetry breaking directions.

These two characteristics circumvent two computational obstacles (see Figure 1) for nonconvex optimization: existences of (i) *spurious* local minimizers and (ii) *high-order* critical points. This implies that starting from any initialization, any optimization method which is able to efficiently escape saddle points converges to the global solution up to symmetry ambiguity. This type of function is also called *strict saddle* or *ridable saddle* functions [29, 102].

[16]Here, for $\mathcal{C}^2$ smooth functions, the negative curvature direction means the negative eigenvector direction of the Hessian.

Minimal Example II: Orthogonal Dictionary Learning. In the context of finding the sparsest vector in a subspace $\mathcal{S} = \mathrm{row}(\boldsymbol{Y})$, let us first use orthgonal DL model $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$ in (II.2) as an example to elaborate more on this type of global geometric structures. Recall from Section II, given the generative model $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} \in \mathbb{R}^{n\times p}$ with orthogonal dictionary $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ and sparse coefficient $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, we aim to learn both $\boldsymbol{A}$ and $\boldsymbol{X}$ only given the data $\boldsymbol{Y}$. When the dictionary $\boldsymbol{A}$ is orthogonal, the observation is that $\mathrm{row}(\boldsymbol{Y}) = \mathrm{row}(\boldsymbol{X})$ and the row vectors of $\boldsymbol{X}$ are sparse. When $\boldsymbol{X}$ is random and Bernoulli-Gaussian, Spielman et al. [46] proved that the row vectors of $\boldsymbol{X}$ are the sparsest vectors in the subspace $\mathcal{S} = \mathrm{row}(\boldsymbol{Y})$ provided $p \geq \Omega(n \log n)$. Therefore, we can reduce the orthogonal DL problem to finding one sparse row vector of $\boldsymbol{X}$ by solving the problem (III.2). If one sparse row vector of $\boldsymbol{X}$ can be found, one may resort to deflation [47] or repeating random trials [46, 69] to recover $\boldsymbol{X}$ and $\boldsymbol{A}$ up to a *signed permutation* ambiguity.

The reason that we can only solve the problem up to a *signed permutation* $\mathrm{SP}(n)$ ambiguity, is because of the inherent symmetry structure, in the sense that signed permutation

$$\boldsymbol{Y} \quad = \quad \boldsymbol{A}\boldsymbol{X} \quad = \quad (\boldsymbol{A}\boldsymbol{\Gamma}) \cdot (\pm\boldsymbol{\Gamma}^\top \boldsymbol{X})$$

which creates *equivalent feasible solutions*, where $\boldsymbol{\Gamma} \in \mathrm{SP}(n)$ is any signed permutation matrix. To see how this symmetry plays out in shaping the benign global optimization landscape, let us consider a simple case[17] that the dictionary $\boldsymbol{A} = \boldsymbol{I}$, so that $\boldsymbol{Y} = \boldsymbol{X}$ and the target solution set $\mathcal{Q}_\star$ of our optimization variable $\boldsymbol{q}$ becomes the set of signed standard basis vectors $\mathcal{Q}_\star = \{\pm\boldsymbol{e}_1, \cdots, \pm\boldsymbol{e}_n\}$. Since the set $\mathcal{Q}_\star$ is also symmetric to signed permutations, it is obvious that the function values $f(\boldsymbol{\Gamma}\boldsymbol{q}) = f(\boldsymbol{q})$ for the problem (III.2). As observed from Figure 4, for all critical points over the sphere:

- All local minimizers are close to the signed standard basis vector $\boldsymbol{q}_\star = \pm\boldsymbol{e}_i$ with the positive Riemannian Hessian, in the sense that

$$\mathrm{Hess}[f](\boldsymbol{q}_\star) \;=\; \mathcal{P}_{\boldsymbol{q}_\star^\perp}\left(\boldsymbol{Y}\nabla^2\varphi(\boldsymbol{Y}^\top\boldsymbol{q}_\star)\boldsymbol{Y}^\top - \left\langle \boldsymbol{Y}\nabla\varphi(\boldsymbol{Y}^\top\boldsymbol{q}_\star), \boldsymbol{q}_\star\right\rangle \boldsymbol{I}\right)\mathcal{P}_{\boldsymbol{q}_\star^\perp} \succeq \alpha\cdot\mathcal{P}_{\boldsymbol{q}_\star^\perp}$$

  for some $\alpha > 0$, so that the function is strongly convex around the local minimizers.
- Saddle point $\boldsymbol{q}_s$ does exist, but they are balanced superpositions of target solutions

$$\boldsymbol{q}_s \;=\; \frac{1}{\sqrt{|\mathcal{I}|}}\sum_{i\in\mathcal{I}}\sigma_i\boldsymbol{e}_i,$$

  for every subset $\mathcal{I} \subseteq \{1, \cdots, m\}$ and sign scalar $\sigma_i \in \{\pm 1\}$. For each saddle point $\boldsymbol{q}_s$, the Riemannian Hessian manifests negative curvature

$$\boldsymbol{e}_i^\top\,\mathrm{Hess}[f](\boldsymbol{q}_\star)\boldsymbol{e}_i \;<\; 0$$

  along the direction pointing to any i-th standard basis with $i \in \mathcal{I}$.

Since there is no spurious local minimizer presenting for orthogonal DL, we can start from any point on the sphere and use any saddle point escaping method to find one sparse row vector from $\boldsymbol{X}$ via $\boldsymbol{Y}^\top\boldsymbol{q}_\star$.

So far, we only considered a relative simple case in DL, where the dictionary $\boldsymbol{A}$ is orthogonal. If the dictionary is complete (i.e., square and invertible), we can approximately reduce the complete DL to orthogonal DL via simple techniques such as preconditioning or whitening of the data $\boldsymbol{Y}$ [47, 103]. Aside from complete DL, recently similar benign global geometric properties have also been discovered for sparse blind deconvolution with multiple inputs [50] (see Figure 2). Similar to DL, this benign global landscape has also been induced by an intrinsic symmetry structure within the problem – the shift symmetry. Indeed, every local minimizer for the sparse blind deconvolution is corresponding to a circulant shift of the filter.

Table II summarizes representative references on the local and global geometric properties for finding the sparsest vector in a subspace in the context of DPCP, DL, and MCS-BD that are illustrated in Section II. Finally, we close this section by noting that the benign global geometric structure pertains to subspace models

[17]For orthogonal $\boldsymbol{A}$, without loss of generality, we can always assume that $\boldsymbol{A} = \boldsymbol{I}$. This is because a change of variable $\overline{\boldsymbol{q}} = \boldsymbol{A}^\top\boldsymbol{q}$ reduces the problem (III.2) to the case $\boldsymbol{A} = \boldsymbol{I}$, which only rotates the optimization landscape.

TABLE II: A selective summary of geometric analysis for problems (III.2). Here ? indicates that there is no existing result for this task.

Objective $\varphi(\cdot)$	Problem	Distance between minimum and the target solution	Local geometry	Global geometry
ℓ_1-norm	DL [69]	0	✓	?
	DPCP [51, 52]	0	✓	?
Huber loss	MCS-BD [81]	$O(\mu)$	✓	?
$\mathcal{C}^\infty$ smooth (e.g., Logcosh)	DL [47, 68]	$O(\mu)$	✓	✓
	MCS-BD [50]	$O(\mu)$	✓	✓

with certain symmetric structures, such as complete DL and sparse blind deconvolution. In both cases, the discrete symmetry such as permutation or shift only induce equivalent good solutions but no spurious local minimizers (see Figure 2). From this perspective, we conjecture that the DPCP problem could also obey benign global geometric property (by using a smooth objective) since in this case, the *continuous* symmetry such as the rotation may also only introduce equivalent good solutions but no spurious local minimizers.

V. Efficient Nonconvex Optimization Methods

The underlying benign geometric structures of the problem have strong implications for designing efficient, guaranteed optimization algorithms. In the following, we overview recent advances of optimization algorithms for solving problems (III.2) with different choices of convex surrogate φ, ranging from smooth to nonsmooth methods, and first-order to second-order approaches. We discuss the underlying principles and the advantages of each method.

A. Algorithms for smooth φ

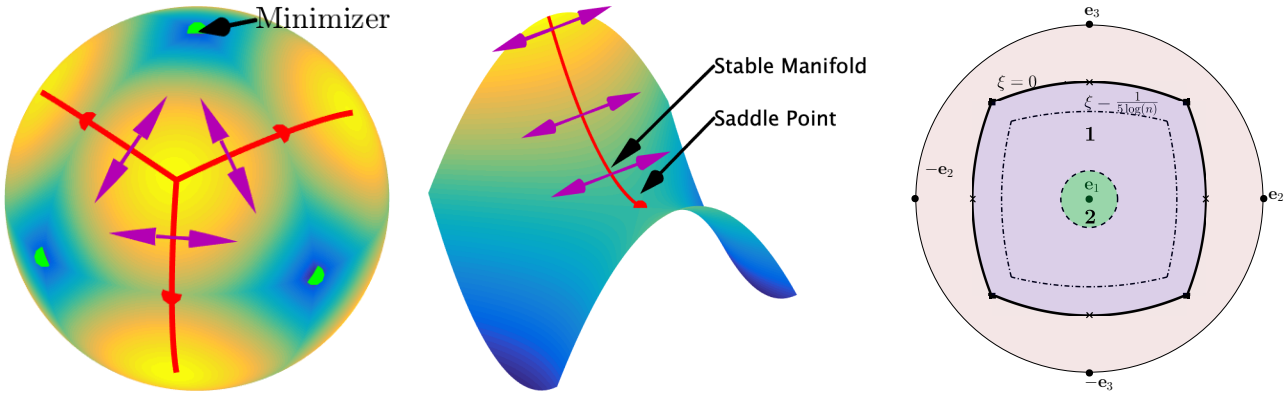
First, we consider the simplest setting where φ is smooth. Undeniably, the most natural way to enforce sparsity is using nonsmooth surrogates such as ℓ^1-penalty (e.g., $\varphi(\cdot) = \|\cdot\|_1$). However, nonsmooth loss often raises great challenges in optimization and algorithmic analysis, due to the non-Lipschitzness of its subgradient. As shown in Table I, there are many ways to replace ℓ^1-penalty with its smooth surrogate, such that we can adopt simple algorithms to tackle our problem. Nonetheless, the tradeoff is that smoothing will introduce approximation errors. To have exact recovery, we need extra rounding step as shown in [34, 67, 81].

a) First-order methods First, let us start with first-order iterative methods for solving the optimization problem (III.2) on the sphere. As we explained in the last section that the Riemannian gradient describes the notion of slope over the sphere, one simple algorithm is to iteratively perform two steps – move the iterate along the opposite direction of the Riemannian gradient and then project it back to the sphere – which is known as Riemannian gradient descent (RGD) [104] on the sphere. In particular, in the $(k+1)$-th step RGD computes

$$\boldsymbol{q}^{(k+1)} \;=\; \mathcal{P}_{\mathbb{S}^{n-1}}\left(\boldsymbol{q}^{(k)} - \eta_k \cdot \mathrm{grad}[f](\boldsymbol{q}^{(k)})\right), \tag{V.1}$$

where η_k represents the step size which can be chosen simply as a constant or selected by the line search method [104].

However, since RGD only uses the Riemannian gradient information, for general nonconvex problems it is only guaranteed to converge to a critical point [104, 105], i.e., it may get stuck at a saddle point. Fortunately, for specific nonconvex problems such as these considered in this paper, Jason et al. [106] proved that RGD escapes from saddle points with negative curvature and converges to a second order critical point almost surely when using random initialization and constant step size. This escaping saddle property was also recently proved in [107] for RGD with varying step sizes. Therefore, when all saddle points exhibis negative



(a) ODL: partition of the sphere (b) ODL: negative curvature (c) ODL: large basin of attraction

Fig. 5: **Illustration of partition of optimization landscape for orthogonal dictionary learning (ODL).**

curvature (or it is called *strict saddle property* [29]), RGD converges almost surely to a local minimum that satisfies second-order optimality condition. Furthermore, as we elaborated in Section IV, this type of local minimizer is close to a target solution for several important cases in finding the sparsest vector in a subspace.

Nonetheless, because the strict function is still too general, these results [106, 107] do not directly imply how fast RGD escapes the saddle points and converges to a local minimum. For general strict saddle functions, the only known result with global convergence rate is a perturbed version of RGD, which injects random noise into the descent iterates preventing stucking at saddle points. In particular, the results [108, 109] showed a sublinear convergence rate $O(1/\epsilon^2)$ for noisy RGD, where ϵ is the accuracy tolerance[18]. However, this type of results does not have direct implication in practice: (i) it is hard to control the level of noise to be injected, (ii) the convergence speed is hindered due to the random noise.

In practice, for many instances of finding the sparsest vector in a subspace, *vanilla* RGD with random initializations seemingly always converge to the target solution with linear rate (see Figure 7a and Figure 8a for an illustration). This is mainly due to the fact that particular problems often have extra structures other than strict saddle property. For example, in the orthgonal DL problem discussed in Section IV, as illustrated in Figure 5, the *stable manifold* (i.e., the set of points along the flow that are sent towards the saddle point) exhibits strong negative curvature, that the gradient increases geometrically moving away the the stable manifold (see Figure 5b). Therefore, we have a large basin of attraction for each target solutions (see Figure 5c), within which the function satisfies regularity conditions analgous to (IV.2) that we discussed in Section IV. Thus, it can be shown that a random initialization falls into one of these basins (the dotted region in Figure 5c) with constant probability. In particular, when $\varphi(\cdot)$ is a $\log\cosh$ function, Gilboa et al. [68] rigorously showed this is true for orthogonal DL and proved sublinear convergence of RGD method. Similar ideas have been adopted for solving sparse blind deconvolution with multiple inputs [81], with Huber loss and an improved analysis that guarantees linear convergence.

b) Second-order methods Another important class of methods that can naturally escape saddle points for problems with benign global geometry are the second-order methods. This type of methods usually forms quadratic approximations of the function in the tangent space, and search for the descent direction based on this approximation within a restricted radius. At a saddle point $\boldsymbol{q}_s \in \mathbb{S}^{n-1}$ with $\operatorname{grad}[f](\boldsymbol{q}_s) = \mathbf{0}$, this type of methods can directly exploit the Hessian information to find the descent direction that is aligned with the negative eigenvector of the Hessian. To see why this happens, consider the following quadratic

[18]Precisely, it produces a point $\boldsymbol{q}$ with gradient smaller than ϵ and Hessian within $\sqrt{\epsilon}$ of being positive semidefinite, i.e., $\|\operatorname{grad}[f](\boldsymbol{q})\| \le \epsilon, \operatorname{Hess}[f](\boldsymbol{q}) \succeq -\sqrt{\epsilon}\mathcal{P}_{\boldsymbol{q}^\perp}$.

approximation $\widehat{f}$ of the function f at a saddle point $\boldsymbol{q}_s \in \mathbb{S}^{n-1}$,

$$\widehat{f}\left(\boldsymbol{q}_s + \boldsymbol{d}\right) \;=\; f(\boldsymbol{q}_s) + \frac{1}{2}\boldsymbol{d}^\top \mathrm{Hess}[f](\boldsymbol{q}_s)\boldsymbol{d} + O(\|\boldsymbol{d}\|^3), \;\; \forall\, \boldsymbol{d} \in \mathrm{T}_{\boldsymbol{q}_s}\mathbb{S}^{n-1}.$$

Because the Riemannian Hessian has negative eigenvalues, if we can find a direction $\boldsymbol{d}$ aligned with the negative eigenvector of the Hessian, then we have $\boldsymbol{d}^\top \mathrm{Hess}[f](\boldsymbol{q}_s)\boldsymbol{d} < 0$. When $\|\boldsymbol{d}\|$ is small enough such that $\widehat{f}\left(\boldsymbol{q}_s + \boldsymbol{d}\right) \approx f\left(\boldsymbol{q}_s + \boldsymbol{d}\right)$, this further implies that $f\left(\boldsymbol{q}_s + \boldsymbol{d}\right) < f\left(\boldsymbol{q}_s\right)$. In other words, the saddle points can be efficiently escaped with second-order methods, directly exploiting the negative curvature information of the Riemannian Hessian.

In addition, when optimizing over the sphere, it should be noted that the Riemannian second-order methods can be viewed as a natural extension of classical second-order methods in the Euclidean space. The quadratic approximation is formed using Riemannian derivatives in the tangent space (which is also a linear space), while the only difference is that we need to perform an extra retraction step to retract the iterate from tangent space back to the sphere (see Figure 6). These Riemannian second-order methods include Riemannian Newton method [104], and Riemannian Quasi-Newton (RQN) method (e.g., the Riemannian trust-region method [104] and the Riemannian cubic-regularization method [110]), where we omit the algorithmic details and refer interested readers to these references [104, 110] for a closer look.

In comparison with first-order methods, the major advantage of second order methods is the convergence speed. As can be seen from Figure 7 and Figure 8, the second methods are usually 10 times faster than first-order methods in terms of iteration complexity. In theory, for example, the Riemannian trust-region method is proved to converge to a target solution at a local quadratic rate for complete DL [47]. Nonetheless, because a common requirement of second-order methods is to form and compute the Riemannian Hessian $\mathrm{Hess}\, f(\cdot)$, for each iteration the second-order methods are way much more expensive than first-order methods in terms of computation and memory cost. Therefore, it is often preferred to use second-order methods for small-scale problems, and use first-order methods for large-scale ones. In the future, it is intersting how to design fast algorithms with low computation cost per iteration (e.g., Riemannian versions of limited-memory BFGS algorithms [111, 112]).

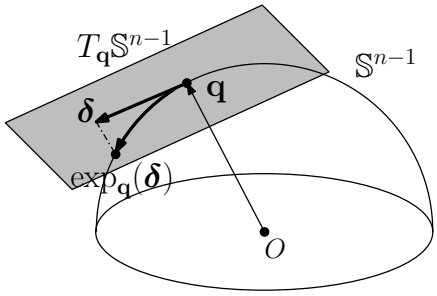


Fig. 6: Illustration of tangent space and retraction to the sphere.

B. Algorithms for non-smooth φ

As alluded in Section V-A, optimizing smooth surrogates often induces approximation errors of the solution, so that extra steps are required for finding exact solutions. In contrast, as demonstrated in Figure 7 and Figure 8, optimizing nonsmooth objectives directly produces exact solutions. In the following, we review recent advances on developing nonsmooth optimization methods for finding the sparsest vector in a subspace.

a) Riemannian SubGradient (RSG) methods. A natural modification of RGD for a non-smooth φ (i.e., the ℓ_1 loss) is a Riemannian SubGradient (RSG) method that replaces the Riemannian gradient by a Riemannian subgradient in (V.1). Although the iterate of RSG has a similar form as RGD in (V.1)

$$\boldsymbol{q}^{(k+1)} \;=\; \mathcal{P}_{\mathbb{S}^{n-1}}\left(\boldsymbol{q}^{(k)} - \eta_k \cdot \partial_R f(\boldsymbol{q}^{(k)})\right), \tag{V.2}$$

the convergence behavior of RSG is more complicated than RGD because of the nonsmoothness. Unlike RGD, the RSG with a constant step size may not converge to a critical point[19]. Moreover, in contrast to

[19]This is true even there is no sphere constraint [113]. As a simple example, consider minimizing $|x|$ by the subgradiment method $x_{k+1} = x_k - \eta_k \,\mathrm{sign}(x_k)$ and suppose that we take $x_0 = 0.01$ and $\eta_k = 0.02$ for all $k \geq 0$. Then, the iterates $\{x_k\}_{k\geq 0}$ will oscillate between the two points $x_+ = 0.01$ and $x_- = -0.01$ and never converge to the global minimum $x^\star = 0$. At best, one can only show that even for a convex function, the subgradient method with a constant step size will converge to a neighborhood of the set of global optima (with rate guarantees if the problem satisfies additional regularity conditions); see, e.g., [113–116]. To ensure the convergence of subgradient methods, a set of diminishing step sizes is generally needed [113, 117].

RGD, the convergence analysis of RSG for general nonsmooth Riemannian optimization problems are largely unexploited [118]. Very recently, Li et al. [119] provided the first convergence rate guarantees for RSG when utilized to optimize nonsmooth functions over Stiefel manifold (which includes the sphere as a special case), under reasonable regularities of the functions. Specifically, if the objective function is weakly convex[20] in the Euclidean space, then RSG with an arbitrary initialization and diminishing step size (e.g., $\eta_k = 1/\sqrt{k}$) converge to a critical point at a sublinear rate (e.g., $O(1/k^{1/4})$). A piecewise geometrically diminishing step size has also been used in [52, 101], resulting in a fast convergence speed. To avoiding tuning the step size, a modified backtracking line search technique has been used in [52], which works well in practice but without a formal convergence guarantee.

For problems (III.2) that satisfy the local regularity condition[21] (IV.2), it is possible to make RSG converge much faster to a target solution by exploiting this local geometric property. For example, if we choose a geometrically diminishing step size (i.e., $\mu_k = O(\beta^k)$ for appropriate $\beta \in (0,1)$) and an initialization within $\mathcal{B}(\mathcal{Q}_\star, \epsilon)$, then RSG converges to $\mathcal{Q}_\star$ with a linear rate [99], i.e., $\mathrm{dist}(\boldsymbol{q}^{(k)}, \mathcal{Q}_\star) \lesssim \beta^k$.

Finally, although RSG is guaranteed to converge to the target solutions under certain regularity condition such as (IV.2), it should be noted that the negative Riemannian subgradient $-\partial_R f(\boldsymbol{q})$ is not necessarily a descent direction. How to efficiently search for an appropriate descent direction to accelerate the convergence for nonsmooth objective is still an open and interesting question. Existing methods such as Riemannian gradient sampling algorithm [120] is often very expensive and lacks non-asymptotic convergence guarantees.

b) Manifold proximal point algorithm (ManPPA). The manifold proximal point algorithm (ManPPA) [121] adopts the idea of the classical proximal point method in the Euclidean space. It can be viewed as an effective approach to find a descent direction on the Moreau envolope of φ within the tangent space:

$$\begin{aligned} \boldsymbol{d}^{(k)} &= \operatorname*{argmin}_{\boldsymbol{d}\in\mathbb{R}^n} \varphi\left(\boldsymbol{Y}^\top(\boldsymbol{q}^{(k)}+\boldsymbol{d})\right) + \frac{1}{2t}\|\boldsymbol{d}\|^2 \quad \text{s.t. } \boldsymbol{d}^\top\boldsymbol{q}^{(k)} = 0, \\ \boldsymbol{q}^{(k+1)} &= \mathcal{P}_{\mathbb{S}^{n-1}}\left(\boldsymbol{q}^{(k)} + \alpha_k \boldsymbol{d}^{(k)}\right), \end{aligned} \tag{V.3}$$

where $t > 0$ and $\alpha_k > 0$ are the step sizes. The efficiency of ManPPA depends on whether we can efficiently solve the optimization subproblem for the descent direction in (V.3). Chen et al. [121] solves this convex subproblem by using an inexact augmented Lagrangian method together with a semi-smooth Newton method. In comparison with RSG, ManPPA converges much faster in terms of iteration complexity[22], but its overall computational complexity can still be higher because solving the subproblem in (V.3) is usually quite expensive even with efficient implementations.

c) Alternating linearization and projection (ALP) method. Another way to deal with nonsmooth ℓ^1 minimization problem with nonlinear connstraint is to linearize the nonlinear constraint, and solve a sequence of linear programs (LPs) until convergence. This is the so-called alternating linearization and projection (ALP) method [82, 100]. In particular, for our problem (III.2), we linearize the spherical constraint $\|\boldsymbol{q}\| = 1$ by using its first order Taylor approximation at each iterate $\boldsymbol{q}^{(k)}$, resulting in a linear constraint $\boldsymbol{q}^\top\boldsymbol{q}^{(k)} = 1$. Thus, we compute a sequence of iterates $\boldsymbol{q}^{(k)}$ via solving the following subproblem

$$\overline{\boldsymbol{q}}^{(k)} = \operatorname*{argmin}_{\boldsymbol{q}\in\mathbb{R}^n} \left\|\boldsymbol{Y}^\top\boldsymbol{q}\right\|_1 \text{ subject to } \boldsymbol{q}^\top\boldsymbol{q}^{(k)} = 1, \quad \text{and} \quad \boldsymbol{q}^{(k+1)} = \mathcal{P}_{\mathbb{S}^{n-1}}\left(\overline{\boldsymbol{q}}^{(k)}\right), \tag{V.4}$$

where the optimization subproblem is simply a LP. It turns out that ALP can be viewed as a special instance of ManPPA by choosing $t = \infty$ and $\alpha_k = 1$ in (V.3) and setting $\boldsymbol{q} = \boldsymbol{q}^{(k)} + \boldsymbol{d}$ in (V.4).

For general nonconvex problems, Späth and Watson [82] established the convergence of ALP to a critical point. For the DPCP problem, this proving technique is further utilized in [100, 122] to show the convergence

[20] We say f is weakly convex if there exists a τ such that $f(\cdot) + \frac{\tau}{2}\|\cdot\|^2$ is convex.

[21] [119] utilizes another property called sharpness, which together with the weak convexity also results a similar regularity condition (IV.2).

[22] A local quadratic convergence rate is established in [121] for problems obeying shaprness, which is satisfied for both DPCP and the orthogonal dictionary learning [119].

TABLE III: Summary of Optimization Methods

Loss function φ	Methods	Order	Convergence	Pros	Cons
Smooth	RGD	1st	xx	xx	
	RQN	2nd	xx	xx	xx
Nonsmooth	RSG	1st	xx	xx	xx
	ManPP	?	xx	xx	xx
	IRLS	?	xx	xx	xx

to a target solution starting from a spectral initialization. The latter result is achieved is due to the underlying benign geometric structures of the problem that we discussed in Section IV-B. In practice, the ALP usually converges much faster than RSG in terms of iterations, but for each iteration it involves solving a LP (e.g., can be solved using Gurobi [123]) subproblem which is often time consuming.

Finally, we note that if $\boldsymbol{q}^{(0)}$ is very close to a global minimum, solving one LP in (V.4) exactly returns the target solution. This property has been exploit in [47, 81, 124] for rounding approximate solutions (often produced by optimizing smooth objectives) to the exact target points. Moreover, for the rounding step Qu et al. [81] proposed an efficient projected subgradient method that enjoys local linear convergence.

d) Iterative reweighted least squares (IRLS). While the ALP iteratively linearizes the nonconvex constraint, the iterative reweighted least squares (IRLS) [100, 125–127] attempts to smooth the nonsmooth objective by a weighted least squares. It should be noted that the IRLS is a classical method to solve ℓ^p-minimization problems ($p \neq 2$) such as compressive sensing [128–130]. The main idea behind IRLS is to alternatively solve a weighted least-squares problem (which often admits a closed-form solution) and update the weights. To illustrate the IRLS for solving (III.2) [100, 125–127], let us rewrite the nonsmooth ℓ^1-norm as $\|\boldsymbol{Y}^\top \boldsymbol{q}\|_1 = \sum_{i=1}^p \left|\boldsymbol{y}_i^\top \boldsymbol{q}\right| = \sum_{i=1}^p \frac{1}{|\boldsymbol{y}_i^\top \boldsymbol{q}|}(\boldsymbol{y}_i^\top \boldsymbol{q})^2$. This inpsires us to consider solving the following subproblem

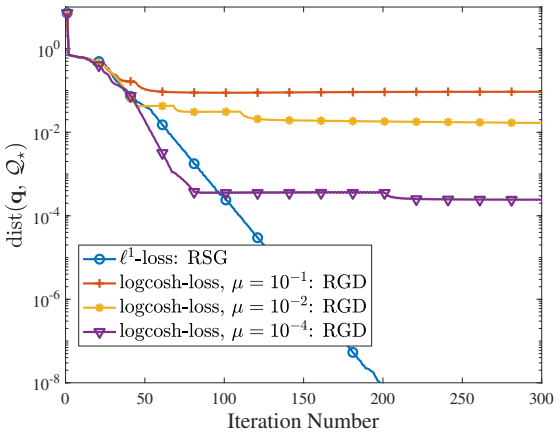
$$\boldsymbol{q}^{(k)} = \operatorname*{argmin}_{\boldsymbol{q}\in\mathbb{S}^{n-1}} \sum_{i=1}^p w_i^{(k-1)}(\boldsymbol{y}_i^\top \boldsymbol{q})^2, \quad \text{and} \quad w_i^{(k)} = \frac{1}{\max\{\delta, \left|\boldsymbol{y}_i^\top \boldsymbol{q}^{(k)}\right|\}} \ \forall i \in [p], \tag{V.5}$$

where δ is a small scalar to avoid numerical explosion. It is not difficult to show that the optimal solution of the subproblem (V.5) is given by the eigenvector corresponding to the smallest eigenvalue of $\sum_{i=1}^p w_i^{(k-1)} \boldsymbol{y}_i \boldsymbol{y}_i^\top$. The convergence behavior of IRLS is discussed in [126], where the global convergence to a critical point and a local convergence to an approximate target solution is established for solving DPCP. In comparison to RSG, IRLS converges much faster and it does not require tuning the step size. However, similar to ALP, the subproblem of IRLS is expensive as it requires performing an eigen-decomposition.

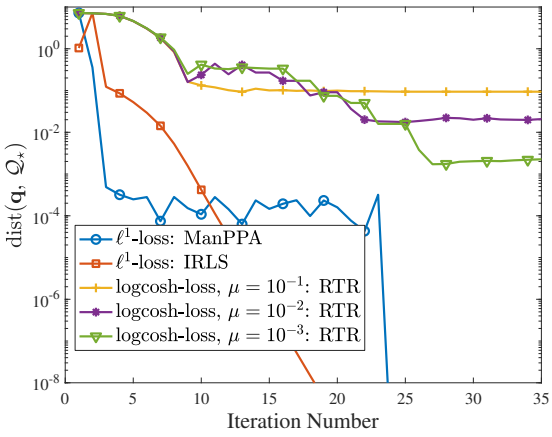
e) Other methods. Finally, we close this section by noting that there are many other methods developed for constrained nonsmooth problems that may also be used for solving (III.2). Typical examples include [131] SQP-GS (which combines sequential quadratic programming and gradient sampling techiniques), and a faster quasi-Newton type method which called GRANSO [132] which improves SQP-GS by employing the BFGS method. GRANSO has been used for solving orthogonal DL in [69] and converges very fast in practice, but there is no convergence guarantee yet.

C. Convergence evaluation

We will conduct simulations to compare different algorithms mentioned in the last two sections for orthogonal dictionary learning (ODL) and dual principle component pursuit (DPCP). For each application,



(a) DPCP: First order algorithms



(b) DPCP: Second order algorithms

Fig. 7: **Convergence performance for DPCP problem.**

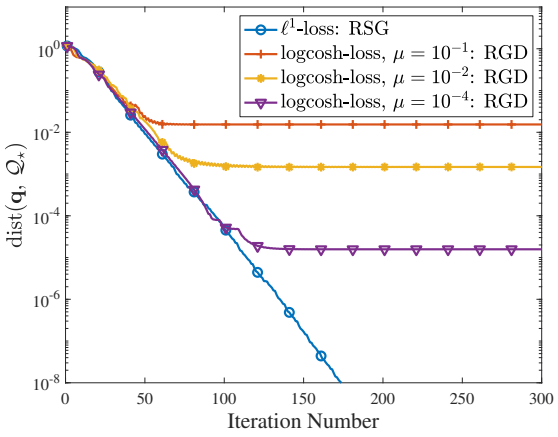
we examine two formulations: 1) the nonsmooth objective function using ℓ^1-loss, 2) the infinitely smooth objective function using logcosh-loss. The algorithms we execute for tackling ℓ^1-loss are the first order RSG, and higher order IRLS and ManPPA. On the other hand, we employ the first order algorithm RGD and higher order algorithm RTR for addressing the smooth logcosh loss.

For DPCP application, a subspace $\mathcal{S}_{\boldsymbol{X}}^{\perp}$ is randomly sampled with co-dimension $r = 40$ in ambient dimension $n = 100$. We then generate $m_1 = 1500$ inliers uniformly at random from the unit sphere in $\mathcal{S}_{\boldsymbol{X}}^{\perp}$ and $m_2 = 3500$ outliers uniformly at random from the unit sphere in $\mathbb{R}^n$. We initialize all the algorithm at the same point with its entries follow standard Gaussian distribution. **For ODL application**, we generate the synthetic data according to [69]. First, a random orthogonal dictionary $\boldsymbol{A} \in \mathbb{R}^n$ is generated with $n = 64$. Wet set the number of samples $m = 5120 \approx 10 \times n^{1.5}$. The sparse coefficient matrix $\boldsymbol{X} \in \mathbb{R}^{n \times m}$ is generated with each entry satisfying Bernoulli-Gaussian distribution with sparsity 0.25. Then the observation $\boldsymbol{Y}$ is generated as $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$. Similarly, we initialize all the algorithms at the same point whose entries follow standard Gaussian distribution. We illustrate the convergence speed of first order methods and higher order methods separately. The experimental results for DPCP and ODL are displayed in Figure 7 and Figure 8, respectively.

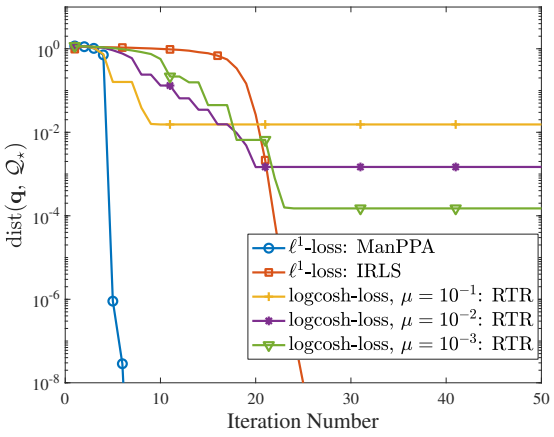
In terms of solution accuracy, one can observe from Figure 7a and Figure 8a that both RSG for ℓ^1-loss and RGD for logcosh-loss converge linearly. The difference lies in RSG for ℓ^1-loss is able to find the exact solution, while RGD for logcosh-loss can only admit an approximate solution and the accuracy directly depends on the smoothing parameter μ in the logcosh loss. In Figure 7b and Figure 8b, the ManPPA and IRLS for ℓ^1-loss converge to the accurate solution of the problem after a few iterations, whereas the RTR for logcosh-loss only finds an approximate one with the gap depending on smooth parameter μ. This observation corroborates with the summary in Table I. **In terms of convergence speed**, higher order methods converge faster than first order ones in terms of iteration number. However, first order methods often have a much cheaper computational load in each iteration. Thus, if the problem size is small scale, higher order methods are recommended, while first order methods will take less time to converge for large scale problems.

VI. Applications in learning low-complexity models from the data

High dimensional data often possess low dimensional structures such as sparsity. For a variety of applications in data science, one of the fundamental problems that we are facing today is how to learn those low-complexity structures/models only given the data. In the following, we present several engineering applications for which some of these challenging learning problems can be reduced to the task of finding the sparsest vector in a



(a) ODL: First order algorithms



(b) ODL: Second order algorithms

Fig. 8: **Convergence performance for ODL problem.**

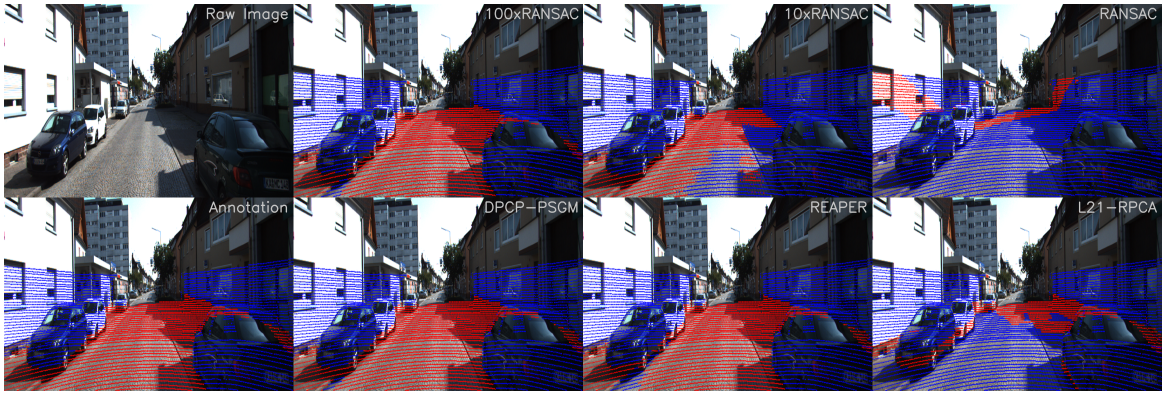


Fig. 9: **Road detection for autonomous driving [52].** Illustration of results on Frame 21 of KITTI-CITY-48 dataset [133], where DPCP [52] outperforms other methods.

subspace. Therefore, we can leverage on the nonconvex optimization approaches illustrated in this work, efficiently solving these problems with provable guarantees.

a) Machine Intelligence. How to endow machines with similar human intelligence has been a long term research interest for decades, which has broad applications in national security, autonomous driving, healthcare, etc. In many cases, it often requires learning low-complexity structures from the observations, in the meanwhile we need guaranteed methods to robustly deal with outliers. Many of these problems can be naturally reduced to finding the sparsest vector in a subspace. As an one example, the DPCP approach introduced in Section II has been successfully applied in the context of the three-view problem, which is of fundamental importance in many computer vision applications, such as 3D reconstruction from 2D images of the scene [51].

Another successful application of DPCP is on road plane detection from 3D point cloud data using the KITTI dataset [133], which is an important computer vision task in autonomous car driving systems. The dataset, recorded from a moving platform while driving in and around Karlsruhe, Germany, consists of image data together with corresponding 3D points collected by a rotating 3D laser scanner. As shown in Figure 9, one important problem is to determine the 3D points that lie off the road plane (outliers indicated by blue) and those on that plane (inliers indicated by red), follows which the road plane can then be easily estimated. Experimental results in Figure 9 show that DPCP outperforms other methods, in particular the RANSAC [134], which is one of the most popular methods for such computer vision applications.

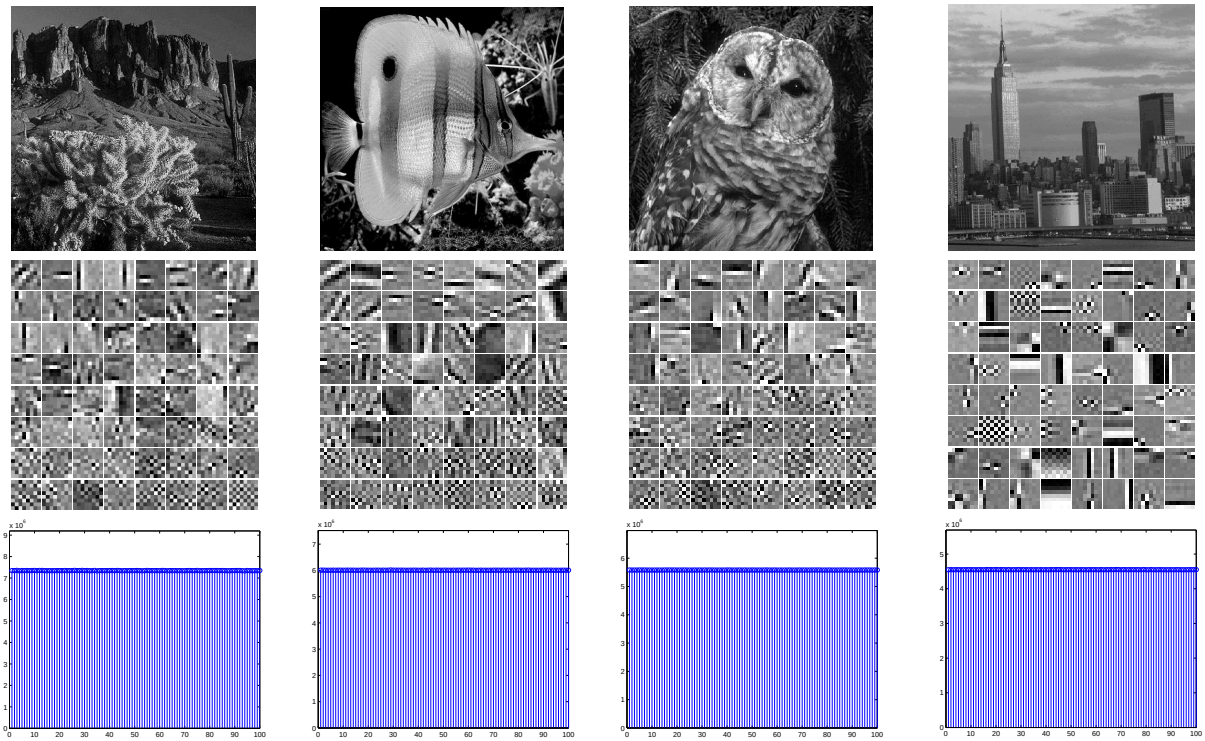
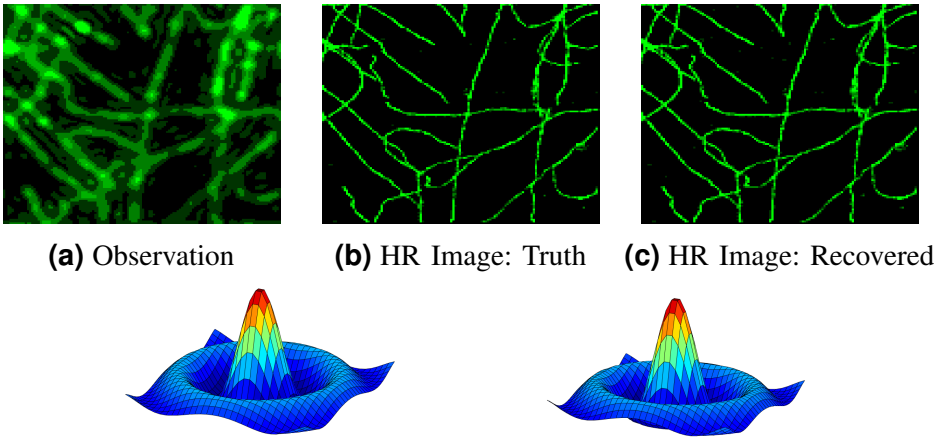


Fig. 10: **Learning representations of natural images [67]. Top:** natural images. **Middle:** 64 dictionary elements of size 8×8 learned via solving (III.2) and deflation. **Bottom:** the plots show the values of $\left\|\boldsymbol{A}_\star^\top \boldsymbol{Y}\right\|_1$ across 100 independent repetitions, where $\boldsymbol{A}_\star$ is the obtained solution for each trial.

b) Representation Learning. High dimensional data often contains quite a lot redundant information, and they often possess low-dimensional structures/representations. The performance of modern machine learning and data analytical methods heavily depends on appropriate low-complexity data representations (or features) which capture hidden information underlying the data. While we used to manually craft representations in the past, it has been demonstrated that learned representations from the data show much superior performance [39]. Therefore, (unsupervised) learning of latent representations of high-dimensional data becomes a fundamental problem in signal processing, machine learning, theoretical neuroscience and many other [135]. As alluded in Section II, one of the most important unsupervised representation learning problems is learning sparsely-used dictionaries [136], which aims to learn a compact dictionary such that every data point can be represented by only a few atoms from the dictionary.

However, despite of recent algorithmic and empirical success [63, 64], most of the methods based on alternating minimizations are lacking theoretical justifications for when and why these algorithms work for dictionary learning. As shown in Section II, when the dictionary is complete, it can be reduced to finding the sparsest vector in a subspace. Moreover, Sun et al. [47, 67] showed that this problem can be solved to the target solutions with efficient algorithms and optimal sparsity level. Figure 10 shows the learned compact representations from natural images using this approach, which is optimized by a second order Riemannian trust region algorithm followed by deflation [67]. As we observe, the method does not only enjoy global performance guarantees (see the bottom of Figure 10) but also efficiently learn meaningful representation from the data (see the middle of Figure 10).

c) Scientific Imaging. In many imaging science applications, we often face the problem of recovering a low-complexity signal from the observation taken from an unknown physical system. For instance, in fluorescent optical microscopy imaging, super-resolution microscopy is a new computation based imaging



(a) Observation **(b)** HR Image: Truth **(c)** HR Image: Recovered

(d) PSF: Ground truth **(e)** PSF: Recovered

Fig. 11: **Solving sparse blind deconvolution for solving super-resolution microscopy imaging [81].** Results on a standard stochastic optical reconstruction microscopy [139] dataset. **Top:** from left to right, observed blurred image, ground truth and recovered HR images **Bottom:** from left to right, ground truth and recovered PSFs.

technique which breaks the resolution limits of conventional optical fluorescence microscopy [137–139]. The basic principle is using photoswitchable florescent probes to create multiple sparse frames of individual molecules to temporally separates the spatially overlapping low resolution image. To improve the resolution limit, we need to computationally recover a sequence of sparse high resolution (HR) images from their convolution with a point spread function (i.e., low resolution images). However, in many scenarios (especially in 3D imaging), as it is often difficult to directly estimate the PSF due to defocus and unknown aberrations [140], it is more desired to jointly estimate both PSF and high resolution image by solving a sparse blind deconvolution problem with multiple inputs.

As discussed in Section II, this sparse blind deconvolution problem can be reduced to finding the sparsest vector in a subspace, which can be efficiently solved by the algorithms in Section V. As a demonstration on effectiveness, we test this approach[23] on a realistic simulated dataset obtained from SMLM challenge website[24] using 1000 video frames. The fluorescence wavelength is 690 nanometer (nm) and the imaging frequency is $f = 25Hz$. Each frame is of size 128×128 with 100 nm pixel resolution, and we solve the single-molecule localization problem on the same grid[25]. As observed in Figure 11, by reducing and solving the finding the sparsest vector in a subspace problem using simple algorithms, it can near perfectly recover both the underlying PSF and HR images, producing accurate recovery results.

VII. Conclusion and Future Directions

This work is part of a recent surge of research efforts on deriving provable and practical nonconvex algorithms to central problems in modern signal processing and machine learning, In this paper, we reviewed several important aspects of recent advances on nonconvex optimization method for solving the problem of finding the sparsest vector in a subspace, ranging from problem formulation, geometric analysis of optimization landscapes, to efficient algorithms and applications. In the following, we discuss several open problems to be addressed along this line of research in the near future.

[23]Here, we consider the Huber-loss for φ, and solve the problem via RGD.

[24]Available at http://bigwww.epfl.ch/smlm/datasets/index.html?p=tubulin-conjal647.

[25]Here, we are estimating the HR images on the same grid as the original image. To obtain even higher resolution than the result we obtain here, people are usually estimating the HR images on a finer grid.

a) Towards more disciplined nonconvex optimization theory. Despite of recent theoretical and algorithmic advances, our understandings of nonconvex optimization is still *far* from satisfactory – the current analysis is delicate, case-by-case, and pertains to problems with elementary symmetry (e.g., permutation or shift symmetry) and simple manifold (e.g., sphere). Analogous to the study of convex functions [9], there is a pressing need for simpler analytic tools, to identify and generalize benign properties for new nonconvex problems appearing in signal processing and machine learning.

b) Learning low-complexity structures over more complicated manifold. In this work, we formulate the problems such as robust subspace recovery and dictionary learning as finding a sparse vector in a subspace, which is constrained over the sphere. However, more natural and robust formulations for these problems involves optimization over more complicated manifolds, such as Stiefel manifold. More technical tools need to be developed towards a better understanding of optimization over these complicated manifolds, despite recent endeavors [99, 103, 119, 141].

c) Applications. In this paper, we reviewed a variety of optimization algorithm for finding a sparse vector in a subspace, with global and strong theoretical guarantees. Moreover, these algorithms are practical for handling large dataset as we demonstrated on several applications in unsupervised learning and imaging sciences. However, we believe the potential of seeking sparse/structured element in a subspace is still unexplored, despite the cases we mentioned at the start. We hope the motivating application discussed in this review could inspire more application ideas of these results.

ACKNOWLEDGEMENT

QQ thanks the generous support of the Microsoft graduate research fellowship and Moore-Sloan fellowship. ZZ and RV are partly supported by NSF Grant 1704458. XL would like to acknowledge the support by Grant CUHK14210617 from the Hong Kong Research Grants Council.

AUTHOR BIOGRAPHY

Qing Qu (qq213@nyu.edu) is a Moore-Sloan research fellow at Center for Data Science, New York University. He received his Ph.D from Columbia University in Electrical Engineering in Oct. 2018. He received his B.Eng. from Tsinghua University in Jul. 2011, and a M.Sc.from the Johns Hopkins University in Dec. 2012, both in Electrical and Computer Engineering. He interned at U.S. Army Research Laboratory in 2012 and Microsoft Research in 2016, respectively. His research interest lies at the intersection of signal/image processing, machine learning, numerical optimization, with focus on developing efficient nonconvex methods and global optimality guarantees for solving engineering problems in signal processing, computational imaging, and machine learning. He is the recipient of Best Student Paper Award at SPARS'15 (with Ju Sun, John Wright), and the recipient of 2016-18 Microsoft Research Fellowship in North America.

Zhihui Zhu (zzhu29@jhu.edu) received the B.Eng. degree in communication engineering from the Zhejiang University of Technology, Hangzhou, China, in 2012, and the Ph.D. degree in electrical engineering from the Colorado School of Mines, Golden, CO, USA, in 2017. He is currently a Postdoctoral Fellow with the Mathematical Institute for Data Science, Johns Hopkins University, Baltimore, MD, USA. His research interests include exploiting inherent structures and applying optimization methods with guaranteed performance for signal processing, machine learning, and data analysis.

Xiao Li (xli@ee.cuhk.edu.hk) received B.Eng. degree in communication engineering with the Zhejiang University of Technology in 2016. He is currently pursuing his Ph.D. degree at The Chinese University of Hong Kong. His current research interests lie in utilizing nonconvex formulation and provable algorithms to problems arising from signal processing, machine learning and data science in order for fast and efficient computation.

Manolis C. Tsakiris (mtsakiris@shanghaitech.edu.cn) is an electrical engineering and computer science graduate of the National Technical University of Athens, Greece. He holds an MS degree in signal processing

from Imperial College London, UK, and a PhD degree in theoretical machine learning from Johns Hopkins University, USA, under the supervision of Prof. René Vidal. Since August 2017 he is an assistant professor at the School of Information Science and Technology (SIST) at ShanghaiTech University. He pursues research on fundamental aspects of subspace learning and related problems in commutative algebra.

John N. Wright (jw2966@columbia.edu) received his B.S. degree in computer engineering, his M.S. degree in electrical engineering, and his Ph.D. degree in electrical engineering from the University of Illinois at Urbana–Champaign (UIUC) in 2004, 2007, and 2009, respectively. From 2009 to 2011, he was with Microsoft Research Asia. He is currently an associate professor in the Electrical Engineering Department at Columbia University, New York. His has received a number of awards, including the 2012 Conference on Learning Theory Best Paper Award (with Dan Spielman and Huan Wang), the 2009 Lemelson-Illinois Prize for Innovation for his work on face recognition, and the 2009 UIUC Martin Award for Excellence in Graduate Research. His research interests include high-dimensional data analysis. He is a Member of the IEEE.

René Vidal (rvidal@jhu.edu) received the B.S. degree in electrical engineering (highest honors) from the Pontificia Universidad Catolica de Chile, Santiago, Chile, in 1997 and the M.S. and Ph.D. degrees in electrical engineering and computer sciences from the University of California, Berkeley, CA, USA, in 2000 and 2003, respectively. He was a Research Fellow at the National ICT Australia in the fall of 2003 and has been a faculty member in the Department of Biomedical Engineering and the Center for Imaging Science of The Johns Hopkins University since 2004. He is co-author of the book "Generalized Principal Component Analysis" (2016), co-editor of the book "Dynamical Vision," and co-author of more than 200 articles in machine learning, computer vision, biomedical image analysis, hybrid systems, robotics, and signal processing. He is or has been Associate Editor of Medical Image Analysis, the IEEE Transactions on Pattern Analysis and Machine Intelligence, the SIAM Journal on Imaging Sciences, Computer Vision and Image Understanding, and the Journal of Mathematical Imaging and Vision, and a Guest Editor of the International Journal on Computer Vision and Signal Processing Magazine. He is a member of the ACM and SIAM.

REFERENCES

[1] P. Jain, P. Kar, *et al.*, "Non-convex optimization for machine learning," *Foundations and Trends® in Machine Learning*, vol. 10, no. 3-4, pp. 142–336, 2017.

[2] J. Sun, "Provable nonconvex methods/algorithms." https://sunju.org/research/nonconvex/.

[3] K. G. Murty and S. N. Kabadi, "Some np-complete problems in quadratic and nonlinear programming," *Mathematical programming*, vol. 39, no. 2, pp. 117–129, 1987.

[4] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling [a sensing/sampling paradigm that goes against the common knowledge in data acquisition]," *IEEE signal processing magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[5] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, p. 717, 2009.

[6] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *Journal of the ACM (JACM)*, vol. 58, no. 3, p. 11, 2011.

[7] E. J. Candes, Y. C. Eldar, T. Strohmer, and V. Voroninski, "Phase retrieval via matrix completion," *SIAM review*, vol. 57, no. 2, pp. 225–251, 2015.

[8] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, vol. 87. Springer Science & Business Media, 2013.

[9] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[10] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," in *Advances in Neural Information Processing Systems*, pp. 2796–2804, 2013.

[11] J. Sun, Q. Qu, and J. Wright, "A geometric analysis of phase retreival," *arXiv preprint arXiv:1602.06664*, 2016.

[12] E. J. Candès, X. Li, and M. Soltanolkotabi, "Phase retrieval via wirtinger flow: Theory and algorithms," *arXiv preprint arXiv:1407.1065*, 2014.

[13] Y. Chen and E. J. Candes, "Solving random quadratic systems of equations is nearly as easy as solving linear systems," *arXiv preprint arXiv:1505.05114*, 2015.

[14] N. Boumal, "Nonconvex phase synchronization," *SIAM Journal on Optimization*, vol. 26, no. 4, pp. 2355–2377, 2016.
[15] H. Liu, M.-C. Yue, and A. Man-Cho So, "On the estimation performance and convergence rate of the generalized power method for phase synchronization," *SIAM Journal on Optimization*, vol. 27, no. 4, pp. 2426–2446, 2017.
[16] K. Lee, N. Tian, and J. Romberg, "Fast and guaranteed blind multichannel deconvolution under a bilinear system model," *IEEE Transactions on Information Theory*, vol. 64, no. 7, pp. 4792–4818, 2018.
[17] X. Li, S. Ling, T. Strohmer, and K. Wei, "Rapid, robust, and reliable blind deconvolution via nonconvex optimization," *Applied and computational harmonic analysis*, 2018.
[18] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon, "Learning sparsely used overcomplete dictionaries via alternating minimization," *arXiv preprint arXiv:1310.7991*, 2013.
[19] J. Sun, Q. Qu, and J. Wright, "Complete dictionary recovery over the sphere," *arXiv preprint arXiv:1504.06785*, 2015.
[20] S. Burer and R. D. Monteiro, "A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization," *Mathematical Programming*, vol. 95, no. 2, pp. 329–357, 2003.
[21] S. Burer and R. D. Monteiro, "Local minima and convergence in low-rank semidefinite programming," *Mathematical Programming*, vol. 103, no. 3, pp. 427–444, 2005.
[22] P. Jain, P. Netrapalli, and S. Sanghavi, "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.
[23] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht, "Low-rank solutions of linear matrix equations via procrustes flow," in *International Conference on Machine Learning*, pp. 964–973, 2016.
[24] R. Sun and Z.-Q. Luo, "Guaranteed matrix completion via non-convex factorization," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 6535–6579, 2016.
[25] S. Bhojanapalli, B. Neyshabur, and N. Srebro, "Global optimality of local search for low rank matrix recovery," in *Advances in Neural Information Processing Systems*, pp. 3873–3881, 2016.
[26] R. Ge, J. D. Lee, and T. Ma, "Matrix completion has no spurious local minimum," in *Advances in Neural Information Processing Systems*, pp. 2973–2981, 2016.
[27] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin, "Global optimality in low-rank matrix optimization," *IEEE Transactions on Signal Processing*, vol. 66, no. 13, pp. 3614–3628, 2018.
[28] Y. Chi, Y. M. Lu, and Y. Chen, "Nonconvex optimization meets low-rank matrix factorization: An overview," *arXiv preprint arXiv:1809.09573*, 2018.
[29] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points—online stochastic gradient for tensor decomposition," in *Proceedings of The 28th Conference on Learning Theory*, pp. 797–842, 2015.
[30] D. C. Sorensen, "Newton?s method with a model trust region modification," *SIAM Journal on Numerical Analysis*, vol. 19, no. 2, pp. 409–426, 1982.
[31] Y. Nesterov and B. T. Polyak, "Cubic regularization of newton method and its global performance," *Mathematical Programming*, vol. 108, no. 1, pp. 177–205, 2006.
[32] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, "Gradient descent only converges to minimizers," in *Conference on learning theory*, pp. 1246–1257, 2016.
[33] Q. Li, Z. Zhu, and G. Tang, "Alternating minimizations converge to second-order optimal solutions," in *International Conference on Machine Learning*, pp. 3935–3943, 2019.
[34] Q. Qu, J. Sun, and J. Wright, "Finding a sparse vector in a subspace: Linear sparsity using alternating directions," *IEEE Transactions on Information Theory*, vol. 62, no. 10, pp. 5855–5880, 2016.
[35] B. Barak, J. A. Kelner, and D. Steurer, "Rounding sum-of-squares relaxations," in *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pp. 31–40, ACM, 2014.
[36] L. Demanet and P. Hand, "Scaling law for recovering the sparsest element in a subspace," *Information and Inference: A Journal of the IMA*, vol. 3, no. 4, pp. 295–309, 2014.
[37] S. B. Hopkins, T. Schramm, J. Shi, and D. Steurer, "Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors," in *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pp. 178–191, ACM, 2016.
[38] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.
[39] M. Elad, *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer Science & Business Media, 2010.
[40] Y. C. Eldar and G. Kutyniok, *Compressed sensing: theory and applications*. Cambridge university press, 2012.
[41] S. Foucart and H. Rauhut, "An invitation to compressive sensing," in *A mathematical introduction to compressive*

sensing, pp. 1–39, Springer, 2013.
[42] E. J. Candès and T. Tao, "Decoding by linear programming," *Information Theory, IEEE Transactions on*, vol. 51, no. 12, pp. 4203–4215, 2005.
[43] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ^1-norm solution is also the sparsest solution," *Communications on pure and applied mathematics*, vol. 59, no. 6, pp. 797–829, 2006.
[44] S. T. McCormick, "A combinatorial approach to some sparse matrix problems.," tech. rep., DTIC Document, 1983.
[45] T. F. Coleman and A. Pothen, "The null space problem i. complexity," *SIAM Journal on Algebraic Discrete Methods*, vol. 7, no. 4, pp. 527–537, 1986.
[46] D. A. Spielman, H. Wang, and J. Wright, "Exact recovery of sparsely-used dictionaries," in *Conference on Learning Theory*, 2012.
[47] J. Sun, Q. Qu, and J. Wright, "Complete dictionary recovery over the sphere i: Overview and the geometric picture," *IEEE Transactions on Information Theory*, vol. 63, no. 2, pp. 853–884, 2016.
[48] L. Wang and Y. Chi, "Blind deconvolution from multiple sparse inputs," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1384–1388, 2016.
[49] M. Rahmani and G. Atia, "Innovation pursuit: A new approach to the subspace clustering problem," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2874–2882, JMLR. org, 2017.
[50] Y. Li and Y. Bresler, "Global geometry of multichannel sparse blind deconvolution on the sphere," in *Advances in Neural Information Processing Systems*, pp. 1132–1143, 2018.
[51] M. C. Tsakiris and R. Vidal, "Dual principal component pursuit," *Journal of Machine Learning Research*, vol. 19, pp. 1–49, 2018.
[52] Z. Zhu, Y. Wang, D. Robinson, D. Naiman, R. Vidal, and M. Tsakiris, "Dual principal component pursuit: Improved analysis and efficient algorithms," in *Advances in Neural Information Processing Systems*, pp. 2171–2181, 2018.
[53] I. Jolliffe, *Principal component analysis*. Springer, 2011.
[54] G. Lerman, T. Zhang, *et al.*, "Robust recovery of multiple subspaces by geometric ℓ_p minimization," *The Annals of Statistics*, vol. 39, no. 5, pp. 2686–2715, 2011.
[55] H. Xu, C. Caramanis, and S. Sanghavi, "Robust pca via outlier pursuit," *IEEE Transactions on Information Theory*, vol. 5, no. 58, pp. 3047–3064, 2012.
[56] G. Lerman and T. Maunu, "An overview of robust subspace recovery," *Proceedings of the IEEE*, vol. 106, no. 8, pp. 1380–1410, 2018.
[57] R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis," 2016.
[58] T. Ding, Z. Zhu, T. Ding, Y. Yang, D. Robinson, M. Tsakiris, and R. Vidal, "Noisy dual principal component pursuit," in *International Conference on Machine Learning*, vol. 97, pp. 1617–1625, 09–15 Jun 2019.
[59] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, p. 607, 1996.
[60] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.
[61] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1045–1057, 2010.
[62] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image processing*, vol. 15, no. 12, pp. 3736–3745, 2006.
[63] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1031–1044, 2010.
[64] J. Mairal, F. Bach, and J. Ponce, "Sparse modeling for image and vision processing," *arXiv preprint arXiv:1411.3230*, 2014.
[65] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM review*, vol. 51, no. 1, pp. 34–81, 2009.
[66] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 4, pp. 791–804, 2011.
[67] J. Sun, Q. Qu, and J. Wright, "Complete dictionary recovery over the sphere ii: Recovery by riemannian trust-region method," *IEEE Transactions on Information Theory*, vol. 63, no. 2, pp. 885–914, 2016.
[68] D. Gilboa, S. Buchanan, and J. Wright, "Efficient dictionary learning with gradient descent," in *International Conference on Machine Learning*, pp. 2252–2259, 2019.
[69] Y. Bai, Q. Jiang, and J. Sun, "Subgradient descent learns orthogonal dictionaries," *arXiv preprint arXiv:1810.10702*, 2018.

[70] Y. Zhang, Y. Lau, H.-w. Kuo, S. Cheung, A. Pasupathy, and J. Wright, "On the global geometry of sphere-constrained sparse blind deconvolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4894–4902, 2017.
[71] H.-W. Kuo, Y. Lau, Y. Zhang, and J. Wright, "Geometry and symmetry in short-and-sparse deconvolution," *arXiv preprint arXiv:1901.00256*, 2019.
[72] P. Campisi and K. Egiazarian, *Blind image deconvolution: theory and applications*. CRC press, 2016.
[73] Y. Li, *Bilinear inverse problems with sparsity: optimal identifiability conditions and efficient recovery*. PhD thesis, University of Illinois at Urbana-Champaign, 2018.
[74] L. Shi and Y. Chi, "Manifold gradient descent solves multi-channel sparse blind deconvolution provably and efficiently," *arXiv preprint arXiv:1911.11167*, 2019.
[75] E. Moulines, P. Duhamel, J.-F. Cardoso, and S. Mayrargue, "Subspace methods for the blind identification of multichannel fir filters," *IEEE Transactions on signal processing*, vol. 43, no. 2, pp. 516–525, 1995.
[76] D. Kundur and D. Hatzinakos, "Blind image deconvolution," *IEEE signal processing magazine*, vol. 13, no. 3, pp. 43–64, 1996.
[77] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman, "Understanding and evaluating blind deconvolution algorithms," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1964–1971, IEEE, 2009.
[78] J. T. Vogelstein, A. M. Packer, T. A. Machado, T. Sippy, B. Babadi, R. Yuste, and L. Paninski, "Fast nonnegative deconvolution for spike train inference from population calcium imaging," *Journal of neurophysiology*, vol. 104, no. 6, pp. 3691–3704, 2010.
[79] E. A. Pnevmatikakis, D. Soudry, Y. Gao, T. A. Machado, J. Merel, D. Pfau, T. Reardon, Y. Mu, C. Lacefield, W. Yang, *et al.*, "Simultaneous denoising, deconvolution, and demixing of calcium imaging data," *Neuron*, vol. 89, no. 2, pp. 285–299, 2016.
[80] N. Kazemi and M. D. Sacchi, "Sparse multichannel blind deconvolution," *Geophysics*, vol. 79, no. 5, pp. V143–V152, 2014.
[81] Q. Qu, X. Li, and Z. Zhu, "A nonconvex approach for exact and efficient multichannel sparse blind deconvolution," *In submission*.
[82] H. Späth and G. Watson, "On orthogonal linear ℓ_1 approximation," *Numerische Mathematik*, vol. 51, no. 5, pp. 531–543, 1987.
[83] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of computational and graphical statistics*, vol. 15, no. 2, pp. 265–286, 2006.
[84] A. d'Aspremont, L. El Ghaoui, M. Jordan, and G. Lanckriet, "A direct formulation of sparse PCA using semidefinite programming," *SIAM Review*, vol. 49, no. 3, 2007.
[85] J. R. Gilbert and M. T. Heath, "Computing a sparse basis for the null space," *SIAM Journal on Algebraic Discrete Methods*, vol. 8, no. 3, pp. 446–459, 1987.
[86] L.-A. Gottlieb and T. Neylon, "Matrix sparsification and the sparse null space problem," in *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pp. 205–218, Springer, 2010.
[87] Y.-B. Zhao and M. Fukushima, "Rank-one solutions for homogeneous linear matrix equations over the positive semidefinite cone," *Applied Mathematics and Computation*, vol. 219, no. 10, pp. 5569–5583, 2013.
[88] Y. Dai, H. Li, and M. He, "A simple prior-free method for non-rigid structure-from-motion factorization," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2018–2025, IEEE, 2012.
[89] G. Beylkin and L. Monzón, "On approximation of functions by exponential sums," *Applied and Computational Harmonic Analysis*, vol. 19, no. 1, pp. 17–48, 2005.
[90] M. Zibulevsky and B. A. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural computation*, vol. 13, no. 4, pp. 863–882, 2001.
[91] A. Anandkumar, D. Hsu, M. Janzamin, and S. M. Kakade, "When are overcomplete topic models identifiable? uniqueness of tensor tucker decompositions with structured sparsity," in *Advances in Neural Information Processing Systems*, pp. 1986–1994, 2013.
[92] J. Ho, Y. Xie, and B. Vemuri, "On a nonlinear generalization of sparse coding and dictionary learning," in *Proceedings of The 30th International Conference on Machine Learning*, pp. 1480–1488, 2013.
[93] B. Barak, J. A. Kelner, and D. Steurer, "Dictionary learning and tensor decomposition via the sum-of-squares method," in *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pp. 143–151, ACM, 2015.
[94] W. H. Yang, L.-H. Zhang, and R. Song, "Optimality conditions for the nonlinear programming problems on riemannian manifolds," *Pacific Journal of Optimization*, vol. 10, no. 2, pp. 415–434, 2014.
[95] P.-A. Absil, R. Mahoney, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*. Princeton University

Press, 2009.
[96] E. J. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval via wirtinger flow: Theory and algorithms," *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 1985–2007, 2015.
[97] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2773–2832, 2014.
[98] X. Li, S. Ling, T. Strohmer, and K. Wei, "Rapid, robust, and reliable blind deconvolution via nonconvex optimization," *Applied and computational harmonic analysis*, vol. 47, no. 3, pp. 893–934, 2019.
[99] Z. Zhu, T. Ding, D. Robinson, M. Tsakiris, and R. Vidal, "A linearly convergent method for non-smooth non-convex optimization on the grassmannian with applications to robust subspace and dictionary learning," in *Advances in Neural Information Processing Systems*, pp. 9437–9447, 2019.
[100] M. C. Tsakiris and R. Vidal, "Dual principal component pursuit," *Journal of Machine Learning Research*, vol. 19, pp. 1–49, 2018.
[101] T. Maunu, T. Zhang, and G. Lerman, "A well-tempered landscape for non-convex robust subspace recovery.," *Journal of Machine Learning Research*, vol. 20, no. 37, pp. 1–59, 2019.
[102] J. Sun, Q. Qu, and J. Wright, "When are nonconvex problems not scary?," *arXiv preprint arXiv:1510.06096*, 2015.
[103] Y. Zhai, Z. Yang, Z. Liao, J. Wright, and Y. Ma, "Complete dictionary learning via $ell4$-norm maximization over the orthogonal group," *arXiv preprint arXiv:1906.02435*, 2019.
[104] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.
[105] N. Boumal, P.-A. Absil, and C. Cartis, "Global rates of convergence for nonconvex optimization on manifolds," *IMA Journal of Numerical Analysis*, vol. 39, no. 1, pp. 1–33, 2018.
[106] J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, and B. Recht, "First-order methods almost always avoid strict saddle points," *Mathematical Programming*, pp. 1–27, 2019.
[107] I. Panageas, G. Piliouras, and X. Wang, "First-order methods almost always avoid saddle points: the case of vanishing step-sizes," *arXiv preprint arXiv:1906.07772*, 2019.
[108] C. Criscitiello and N. Boumal, "Efficiently escaping saddle points on manifolds," *arXiv preprint arXiv:1906.04321*, 2019.
[109] Y. Sun, N. Flammarion, and M. Fazel, "Escaping from saddle points on riemannian manifolds," *arXiv preprint arXiv:1906.07355*, 2019.
[110] J. Zhang and S. Zhang, "A cubic regularized newton's method over riemannian manifolds," *arXiv preprint arXiv:1805.05565*, 2018.
[111] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization," *Mathematical programming*, vol. 45, no. 1-3, pp. 503–528, 1989.
[112] X. Yuan, W. Huang, P.-A. Absil, and K. A. Gallivan, "A riemannian limited-memory bfgs algorithm for computing the matrix geometric mean," *Procedia Computer Science*, vol. 80, pp. 2147–2157, 2016.
[113] N. Z. Shor, *Minimization Methods for Non-Differentiable Functions*, vol. 3 of *Springer Series in Computational Mathematics*. Berlin Heidelberg: Springer-Verlag, 1985.
[114] A. Nedić and D. Bertsekas, "Convergence Rate of Incremental Subgradient Algorithms," in *Stochastic Optimization: Algorithms and Applications* (S. Uryasev and P. M. Pardalos, eds.), vol. 54 of *Applied Optimization*, Dordrecht: Springer Science+Business Media, 2001.
[115] D. P. Bertsekas, "Incremental Gradient, Subgradient, and Proximal Methods for Convex Optimization," in *Optimization for Machine Learning* (S. Sra, S. Nowozin, and S. J. Wright, eds.), Neural Information Processing Series, pp. 85–119, Cambridge, Massachusetts: MIT Press, 2012.
[116] D. Davis, D. Drusvyatskiy, K. J. MacPhee, and C. Paquette, "Subgradient Methods for Sharp Weakly Convex Functions," *Journal of Optimization Theory and Applications*, vol. 179, no. 3, pp. 962–982, 2018.
[117] J.-L. Goffin, "On Convergence Rates of Subgradient Optimization Methods," *Mathematical programming*, vol. 13, no. 1, pp. 329–347, 1977.
[118] P.-A. Absil and S. Hosseini, "A collection of nonsmooth riemannian optimization problems," in *Nonsmooth Optimization and Its Applications*, pp. 1–15, Springer, 2019.
[119] X. Li, S. Chen, Z. Deng, Q. Qu, Z. Zhu, and A. M. C. So, "Nonsmooth optimization over stiefel manifold: Riemannian subgradient methods," *arXiv preprint arXiv:1911.05047*, 2019.
[120] S. Hosseini and A. Uschmajew, "A riemannian gradient sampling algorithm for nonsmooth optimization on manifolds," *SIAM Journal on Optimization*, vol. 27, no. 1, pp. 173–189, 2017.
[121] S. Chen, Z. Deng, S. Ma, and A. M.-C. So, "Manifold proximal point algorithms for dual principal component

pursuit and orthogonal dictionary learning," in *Asilomar Conference on Signals, Systems, and Computers*, 2019.
[122] Z. Zhu, Y. Wang, D. P. Robinson, D. Q. Naiman, R. Vidal, and M. C. Tsakiris, "Dual principal component pursuit: Probability analysis and efficient algorithms," *arXiv preprint arXiv:1812.09924*, 2018.
[123] G. Optimization, "Inc.,"gurobi optimizer reference manual," 2015," 2014.
[124] Q. Qu, J. Sun, and J. Wright, "Finding a sparse vector in a subspace: Linear sparsity using alternating directions," in *Advances in Neural Information Processing Systems*, pp. 3401–3409, 2014.
[125] G. Lerman, M. B. McCoy, J. A. Tropp, and T. Zhang, "Robust computation of linear models by convex relaxation," *Foundations of Computational Mathematics*, vol. 15, no. 2, pp. 363–410, 2015.
[126] G. Lerman and T. Maunu, "Fast, robust and non-convex subspace recovery," *Information and Inference: A Journal of the IMA*, vol. 7, no. 2, pp. 277–336, 2017.
[127] T. Zhang and G. Lerman, "A novel m-estimator for robust pca," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 749–808, 2014.
[128] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *Journal of Fourier analysis and applications*, vol. 14, no. 5-6, pp. 877–905, 2008.
[129] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3869–3872, IEEE, 2008.
[130] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk, "Iteratively reweighted least squares minimization for sparse recovery," *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 63, no. 1, pp. 1–38, 2010.
[131] F. E. Curtis and M. L. Overton, "A sequential quadratic programming algorithm for nonconvex, nonsmooth constrained optimization," *SIAM Journal on Optimization*, vol. 22, no. 2, pp. 474–500, 2012.
[132] F. E. Curtis, T. Mitchell, and M. L. Overton, "A bfgs-sqp method for nonsmooth, nonconvex, constrained optimization and its evaluation using relative minimization profiles," *Optimization Methods and Software*, vol. 32, no. 1, pp. 148–181, 2017.
[133] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.
[134] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.
[135] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.
[136] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?," *Vision research*, vol. 37, no. 23, pp. 3311–3325, 1997.
[137] E. Betzig, G. H. Patterson, R. Sougrat, O. W. Lindwasser, S. Olenych, J. S. Bonifacino, M. W. Davidson, J. Lippincott-Schwartz, and H. F. Hess, "Imaging intracellular fluorescent proteins at nanometer resolution," *Science*, vol. 313, no. 5793, pp. 1642–1645, 2006.
[138] S. T. Hess, T. P. Girirajan, and M. D. Mason, "Ultra-high resolution imaging by fluorescence photoactivation localization microscopy," *Biophysical journal*, vol. 91, no. 11, pp. 4258–4272, 2006.
[139] M. J. Rust, M. Bates, and X. Zhuang, "Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (storm)," *Nature Methods*, vol. 3, no. 10, p. 793, 2006.
[140] P. Sarder and A. Nehorai, "Deconvolution methods for 3-d fluorescence microscopy images," *IEEE Signal Processing Magazine*, vol. 23, no. 3, pp. 32–45, 2006.
[141] J. Hu, X. Liu, Z. Wen, and Y. Yuan, "A brief introduction to manifold optimization," 2019.