When Are Nonconvex Problems Not Scary?

Ju Sun, Qing Qu, and John Wright
{js4038, qq2105, jw2966}@columbia.edu

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

October 20, 2015 Revised: January 22, 2022

Abstract

In this note, we focus on smooth nonconvex optimization problems that obey: (1) all local minimizers are also global; and (2) around any saddle point or local maximizer, the objective has a negative directional curvature. Concrete applications such as dictionary learning, generalized phase retrieval, and orthogonal tensor decomposition are known to induce such structures. We describe a second-order trust-region algorithm that provably converges to a global minimizer efficiently, without special initializations. Finally we highlight alternatives, and open problems in this direction.

1 Introduction

General nonconvex optimization problems (henceforth "NCVX problems" for brevity) are NP-hard, even the goal is computing only a local minimizer [MK87, Ber99]. In applied disciplines, however, NCVX problems abound, and heuristic algorithms such as gradient descent and alternating directions are often surprisingly effective. The ability of natural heuristics to find high-quality solutions for practical NCVX problems remains largely mysterious.

In this note, we study a family of NCVX problems that *can* be solved efficiently. This family cuts across central tasks in signal processing and machine learning, such as complete (sparse) dictionary learning [SQW15], generalized phase retrieval [SQW16], orthogonal tensor decomposition [GHJY15], and noisy phase synchronization and community detection [Bou16, BBV16].

Natural optimization formulations for these distinct tasks are nonconvex; surprisingly they exhibit a common characteristic structure. In each case, the goal is to estimate or recover an object from observed data. Under certain technical hypotheses, *every local minimizer of the objective function exactly recovers the object of interest*.

With this structure, the central issue is how to escape the saddle points and local maximizers. Fortunately, for these problems, *all saddle points and local maximizers are "typical" – the associated Hessian matrix has at least one negative eigenvalue*. Geometrically, this means around any saddle point or local maximizer, the objective function has a negative curvature in a certain direction. Particularly, we call saddles of this type *ridable saddles*;[1] the importance of this apparently extraneous restriction is illustrated in Figure 1. Intuitively, at saddle points or local maximizers, in the direction of negative curvature the objective function is also locally descending. One can use this to design algorithms

[1] They are also called "strict saddle" points in optimization literature, see, e.g., pp 38 of [RI10]; see also [GHJY15].

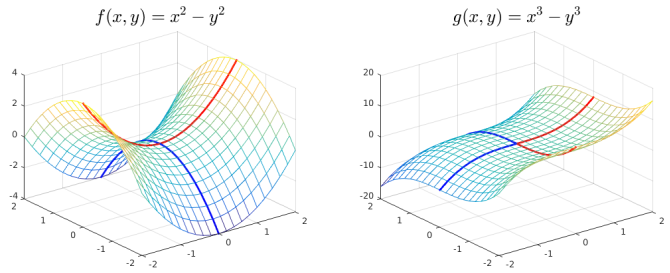

Figure 1: Not all saddle points are ridable! Shown in the plot are functions $f(x,y) = x^2 - y^2$ (left) and $g(x,y) = x^3 - y^3$ (right). For g, both first- and second-order derivatives vanish at $(0,0)$, producing a saddle that is induced by third-order derivatives. In both plots, red curves indicate local ascent directions and blue curves indicate local descent directions.

that escape from the saddle points and local maximizers concerned here. Indeed, consider a natural quadratic approximation to the objective f around a saddle point $\boldsymbol{x}$:

$$\widehat{f}(\boldsymbol{\delta}; \boldsymbol{x}) = f(\boldsymbol{x}) + \frac{1}{2}\boldsymbol{\delta}^* \nabla^2 f(\boldsymbol{x}) \boldsymbol{\delta}.$$

When $\boldsymbol{\delta}$ is chosen to align with one eigenvector associated with a negative eigenvalue $\lambda_{\mathrm{neg}}[\nabla^2 f(\boldsymbol{x})]$, it holds that

$$\widehat{f}(\boldsymbol{\delta}; \boldsymbol{x}) - f(\boldsymbol{x}) \leq -\frac{1}{2}|\lambda_{\mathrm{neg}}| \, \|\boldsymbol{\delta}\|^2 .$$

Thus, minimizing $\widehat{f}(\boldsymbol{\delta}; \boldsymbol{x})$ locally provides a direction $\boldsymbol{\delta}_\star$ that tends to decrease the objective f, provided local approximation of $\widehat{f}$ to f is reasonably accurate.[2] Based on this intuition, we derive an algorithmic framework that can exploit the second-order information to escape from saddle points and local maximizers and provably returns a global minimizer.

2 Nonconvex Optimization with Ridable Saddles

In this section, we present a more quantitative definition of the problem class we focus on and provide several concrete examples in this class.

We are interested in optimization problem of the form:

$$\text{minimize } f(\boldsymbol{x}), \quad \text{subject to} \quad \boldsymbol{x} \in \mathcal{M}. \tag{2.1}$$

Here we assume f is twice continuously differentiable, i.e., it has continuous first- and second-order derivatives, and $\mathcal{M}$ is a Riemannian manifold. Restricting f to $\mathcal{M}$ and (with abuse of notation) writing the restricted function as f also, one can effectively treat (2.1) as an unconstrained optimization on $\mathcal{M}$. We further use $\operatorname{grad} f(\boldsymbol{x})$ and $\operatorname{Hess} f(\boldsymbol{x})$ to denote the Riemannian gradient and Hessian of f at point $\boldsymbol{x}$ [3], which one can think of as Riemannian counterparts of Euclidean gradient and Hessian for functions, with the exception that $\operatorname{grad} f(\boldsymbol{x})[\cdot]$ and $\operatorname{Hess} f(\boldsymbol{x})[\cdot]$ only act on vectors in tangent space of $\mathcal{M}$ at $\boldsymbol{x}$, i.e., $T_{\boldsymbol{x}}\mathcal{M}$.

[2] For general saddles that seem to demand higher-order approximations, the computation may quickly become intractable. For example, third order saddle points seem to generally demand studying spectral property of three-way tensors, which entails NP-hard computational problems [HL13]; see [AG16] for a recent attempt in this line.

[3] Detailed introduction to these quantities can be found in [AMS09]. We prefer to keep this at an intuitive level not to obscure the main ideas.

Definition 2.1 ($(\alpha, \beta, \gamma, \delta)$-$\mathcal{X}$ functions) *A function $f : \mathcal{M} \mapsto \mathbb{R}$ is $(\alpha, \beta, \gamma, \delta)$-$\mathcal{X}$ (α, β, γ, $\delta > 0$) if:*

0) all local minimizers of f are also global minimizers;

and f is $(\alpha, \beta, \gamma, \delta)$-saddle,[4] *i.e., any point $\boldsymbol{x} \in \mathcal{M}$ obeys **at least one of the following**: ($T_{\boldsymbol{x}}\mathcal{M}$ is the tangent space of $\mathcal{M}$ at point $\boldsymbol{x}$)*

*1) **[Strong gradient]** $\|\mathrm{grad}\, f(\boldsymbol{x})\| \geq \beta$;*

*2) **[Negative curvature]** There exists $\boldsymbol{v} \in T_{\boldsymbol{x}}\mathcal{M}$ with $\|\boldsymbol{v}\| = 1$ such that $\langle \mathrm{Hess}\, f(\boldsymbol{x})[\boldsymbol{v}], \boldsymbol{v}\rangle \leq -\alpha$;*

*3) **[Strong convexity around minimizers]** There exists a local minimizer $\boldsymbol{x}_\star$ such that $\|\boldsymbol{x} - \boldsymbol{x}_\star\| \leq \delta$, and for all $\boldsymbol{y} \in \mathcal{M}$ that is in 2δ neighborhood of $\boldsymbol{x}_\star$, $\langle \mathrm{Hess}\, f(\boldsymbol{y})[\boldsymbol{v}], \boldsymbol{v}\rangle \geq \gamma$ for any $\boldsymbol{v} \in T_{\boldsymbol{y}}\mathcal{M}$ with $\|\boldsymbol{v}\| = 1$, i.e., the function f is γ-strongly convex in 2δ neighborhood of $\boldsymbol{x}_\star$.* [5]

In words, the function has no spurious local minimizers. Moreover, each point on the manifold $\mathcal{M}$ either has strong Riemannian gradient, or has Riemannian Hessian with at least one strictly negative eigenvalue, or lives in a small neighborhood of a local minimizer, such that the function is locally strongly convex. We remark in passing that requiring a function to be ridable may appear far too restrictive than it actually is. Indeed, one of the central results in Morse theory implies that a generic smooth function is ridable.

In this note, we deal exclusively with minimizing $\mathcal{X}$ functions. [6] These functions indeed appear in natural nonconvex formulations of important practical problems.

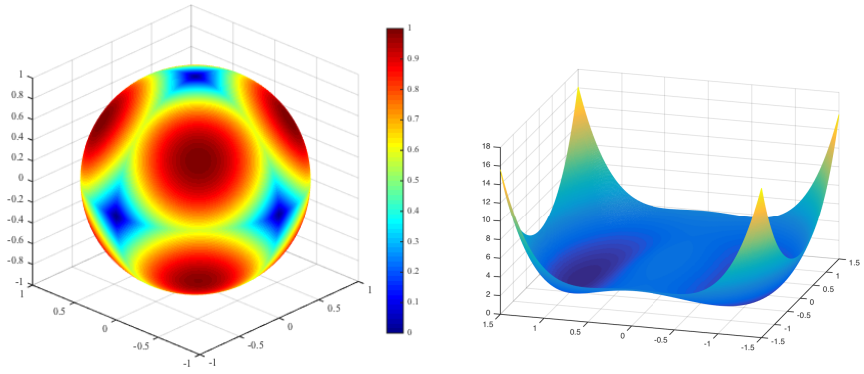

Figure 2: (Left) Function landscape of learning sparsifying complete basis via (2.3) in $\mathbb{R}^3$. (Right) Function landscape of generalized phase retrieval via (2.4), assuming the target signal $\boldsymbol{x}$ is real in $\mathbb{R}^2$. In each case, note the equivalent global minimizers and the ridable saddles.

- **The Eigenvector Problem.** For a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, the classic eigenvector problem is

$$\mathrm{maximize}_{\boldsymbol{x}\in\mathbb{R}^n}\ \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \quad \text{subject to} \quad \|\boldsymbol{x}\| = 1. \tag{2.2}$$

  Here the manifold is the sphere $\mathbb{S}^{n-1}$. It can be easily shown that (see, e.g., Section 4.6 of [AMS09]) the set of critical points to the problem is exactly the set of eigenvectors to $\boldsymbol{A}$.

[4] See also strict-saddle function defined in [GHJY15]. When $\mathcal{M}$ is $\mathbb{R}^n$ or $\mathbb{C}^n$, the two definitions coincide. It is interesting to see if the two agree in general settings. Particularly, [GHJY15] deals only with sets defined by equalities of the form $c_i(\boldsymbol{x}) = 0$ with differentiable function c, which excludes many manifolds of interest, such as symmetric positive definite matrices of a fixed dimension. See this page: `http://www.manopt.org/tutorial.html#manifolds` for more examples. See also discussion in Introduction of this paper [ABG07] on relationship between manifold optimization and constrained optimization in the Euclidean space.

[5] Strong convexity is required for the sake of deriving concrete convergence rate near the minimizers. Less stringent conditions might already be sufficient, depending on the specific problems and target computational guarantees. Also, it is possible to modify the trust-region methods (described later) to take advantage of fine problem structure; see, e.g.,[SQW16].

[6] Though if the target is to compute any local minimizer of a ridable function, our method also applies.

Moreover, suppose $\lambda_1 > \lambda_2 \geq \ldots \lambda_{n-1} > \lambda_n$, with the corresponding eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$. Then, the only global maximizers are $\pm\boldsymbol{v}_1$, the only global minimizers are $\pm\boldsymbol{v}_n$, and all the intermediate eigenvectors and their negatives are ridable saddle points.[7] Quantitatively, one can show that the function is $(c(\lambda_{n-1}-\lambda_n), c(\lambda_{n-1}-\lambda_n)/\lambda_1, c(\lambda_{n-1}-\lambda_n), 2c(\lambda_{n-1}-\lambda_n)/\lambda_1)$-ridable over $\mathbb{S}^{n-1}$ for a certain absolute constant $c > 0$.

- **Complete Dictionary Recovery [SQW15].** Arising in signal processing and machine learning, dictionary learning tries to approximate a given data matrix $\boldsymbol{Y} \in \mathbb{R}^{n\times p}$ as the product of a dictionary $\boldsymbol{A}$ and a sparse coefficient matrix $\boldsymbol{X}$. In recovery setting, assuming $\boldsymbol{Y} = \boldsymbol{A}_0\boldsymbol{X}_0$ with $\boldsymbol{A}_0$ square and invertible, $\boldsymbol{Y}$ and $\boldsymbol{X}_0$ have the same row space. Under appropriate model on $\boldsymbol{X}_0$, it makes sense to recover one row of $\boldsymbol{X}_0$ each time by finding the sparsest direction[8] in $\mathrm{row}(\boldsymbol{Y})$ by solving the optimization:

$$\text{minimize}_{\boldsymbol{q}} \left\|\boldsymbol{q}^\top \boldsymbol{Y}\right\|_0 \quad \text{subject to} \quad \boldsymbol{q} \neq \boldsymbol{0},$$

which can be relaxed as

$$\text{minimize } f(\boldsymbol{q}) \doteq \frac{1}{p}\sum_{k=1}^{p} h(\boldsymbol{q}^\top \overline{\boldsymbol{y}}_k) \quad \text{subject to} \quad \|\boldsymbol{q}\|_2 = 1 \quad [\text{i.e., } \boldsymbol{q} \in \mathbb{S}^{n-1}]. \tag{2.3}$$

Here $h(\cdot)$ is a smooth approximation to the $|\cdot|$ function and $\overline{\boldsymbol{y}}_k$ the k-th column of $\overline{\boldsymbol{Y}}$, a proxy of $\boldsymbol{Y}$. The manifold $\mathcal{M}$ is $\mathbb{S}^{n-1}$ here. [SQW15] (Theorem 2.3 and Corollary 2.4) showed that when $h(\cdot) = \mu \log\cosh(\cdot/\mu)$ and p is reasonably large, those $\boldsymbol{q}$'s that help recover rows of $\boldsymbol{X}_0$ are *the only local minimizers* of f over $\mathbb{S}^{n-1}$.[9] Moreover, these exists a positive constant c such that f is $(c\theta, c\theta, c\theta/\mu, \sqrt{2}\mu/7)$-ridable over $\mathbb{S}^{n-1}$, where θ controls the sparsity level of $\boldsymbol{X}_0$.

- **(Generalized) Phase Retrieval [SQW16].** For complex signal $\boldsymbol{x} \in \mathbb{C}^n$, generalized phase retrieval (GPR) tries to recover $\boldsymbol{x}$ from the nonlinear measurements of the form $y_k = |\boldsymbol{a}_k^*\boldsymbol{x}|$, for $k = 1, \ldots, m$. This task has occupied the central place in imaging systems for scientific discovery [SEC+15]. Assuming i.i.d. Gaussian measurement noise, a natural formulation for GPR is

$$\text{minimize}_{\boldsymbol{z}\in\mathbb{C}^n} f(\boldsymbol{z}) \doteq \frac{1}{4m}\sum_{k=1}^{m}(y_k^2 - |\boldsymbol{a}_k^*\boldsymbol{z}|^2)^2. \tag{2.4}$$

The manifold $\mathcal{M}$ here is $\mathbb{C}^n$. It is obvious that for all $\boldsymbol{z}$, $f(\boldsymbol{z})$ has the same value as $f(\boldsymbol{z}\mathrm{e}^{\mathrm{i}\theta})$ for any $\theta \in [0, 2\pi)$. [SQW16] showed when $m \geq \Omega(n\log^3 n)$, $\{\boldsymbol{x}\mathrm{e}^{\mathrm{i}\theta}\}$ are the only local minimizers, and also global minimizers (as $f \geq 0$). Moreover, modulo the trivial equivalence discussed above, the function f is $(c, c/(n\log m), c, c/(n\log m))$-ridable for a certain absolute constant c, assuming $\|\boldsymbol{x}\| = 1$.

- **Independent Component Analysis (ICA) and Orthogonal Tensor Decomposition [GHJY15].** Typical setting of ICA asks for a linear transformation $\boldsymbol{A}$ for a given data matrix $\boldsymbol{Y}$, such that

[7] One can state a more general version of the results, allowing multiplicity of maximum eigenvalues.

[8] The absolute scale is not recoverable.

[9] These local minimizers are all global when $p \to \infty$. For finite p that is large enough, these local minimizers assume very close values, and each of them produces a close approximation to a row of $\boldsymbol{X}_0$.

rows of $\boldsymbol{AY}$ achieve maximal statistical independence. Tensor decomposition generalizes spectral decomposition of matrices. Here we focus on *orthogonally decomposable* d-th order tensors $\mathcal{T}$ which can be represented as

$$\mathcal{T} = \sum_{i=1}^{n} \boldsymbol{a}_i^{\otimes d}, \quad \boldsymbol{a}_i^\top \boldsymbol{a}_j = \delta_{ij} \; \forall \; i, j, (\boldsymbol{a}_i \in \mathbb{R}^n \; \forall \; i)$$

where $\otimes$ generalizes the usual outer product of vectors. Tensor decomposition refers to finding (up to sign and permutation) the components $\boldsymbol{a}_i$'s given $\mathcal{T}$. With appropriate processing and up to small perturbation, ICA is showed to be equivalent to decomposition of a certain form of 4-th order orthogonally decomposable tensors [FJK96, AGMS12]. Specifically, [GHJY15] showed (Section C.1.) [10] the minimization problem

$$\text{minimize } f(\boldsymbol{u}) \doteq -\mathcal{T}(\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{u}, \boldsymbol{u}) = -\sum_{i=1}^{n} (\boldsymbol{a}_i^\top \boldsymbol{u})^4 \quad \text{subject to} \quad \|\boldsymbol{u}\|_2 = 1$$

has $\pm \boldsymbol{a}_i$'s as its only minimizers and the function f is $(7/n, 1/\mathrm{poly}(n), 3, 1/\mathrm{poly}(n))$-ridable over $\mathbb{S}^{n-1}$. Once one of the component is obtained, one can apply deflation to obtain the others. One alternative that tends to make the process more noise-stable is trying to recover all the components in one shot. To this end, [GHJY15] proposed to solve

$$\begin{aligned} &\text{minimize } g(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r) \doteq \sum_{i \neq j} \mathcal{T}(\boldsymbol{u}_i, \boldsymbol{u}_i, \boldsymbol{u}_j, \boldsymbol{u}_j) = \sum_{i \neq j} \sum_{k=1}^{n} (\boldsymbol{a}_k^\top \boldsymbol{u}_i)^2 (\boldsymbol{a}_k^\top \boldsymbol{u}_j)^2, \\ &\text{subject to } \|\boldsymbol{u}_i\| = 1 \; \forall i \in [n]. \end{aligned}$$

The object $\{\boldsymbol{U} \in \mathbb{R}^{n \times n} : \|\boldsymbol{u}_i\| = 1 \; \forall i\}$ is called the *oblique manifold*, which is a product space of multiple spheres. [GHJY15] showed all local minimizers of g are equivalent (i.e., signed permuted) copies of $[\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n]$. Moreover, g is $(1/\mathrm{poly}(n), 1/\mathrm{poly}(n), 1, 1/\mathrm{poly}(n))$-ridable.

- **Phase Synchronization and Community Detection [Bou16, BBV16].** Phase synchronization concerns recovery of unit-modulus complex scalars from their relative phases. More precisely, recovering an unknown vector $\boldsymbol{z} \in \mathbb{C}_1^n$ with

$$\mathbb{C}_1^n \doteq \{\boldsymbol{z} \in \mathbb{C}^n : |z_1| = \cdots = |z_n| = 1\},$$

from noisy measurements of the form $C_{ij} = z_i \overline{z_j} + \Delta_{ij}$. The problem is interesting when the noise is nonzero yet controlled, which demands robust solution schemes. Turning to the optimization approach, a natural formulation (if one assumes a Gaussian noise model) is

$$\text{minimize}_{\boldsymbol{x} \in \mathbb{C}_1^n} \; \|\boldsymbol{x}\boldsymbol{x}^* - \boldsymbol{C}\|_F^2,$$

where we have collected C_{ij} into a matrix $\boldsymbol{C}$. Assuming the noise is symmetric (i.e., $\Delta_{ij} = \Delta_{ji}$), the above formulation is equivalent to

$$\text{minimize}_{\boldsymbol{x} \in \mathbb{C}_1^n} \; -\boldsymbol{x}^* \boldsymbol{C} \boldsymbol{x}. \tag{2.5}$$

[10] [GHJY15] has not used the manifold language as we use here, but resorted to Lagrange multiplier and optimality of the Lagrangian function. For the two decomposition formulations we discussed here, one can verify that the gradient and Hessian they defined are exactly the Riemannian gradient and Hessian of the respective manifolds.

Interestingly, for the phase synchronization model, i.e., $\boldsymbol{C} = \boldsymbol{z}\boldsymbol{z}^* + \boldsymbol{\Delta}$ with Hermitian noise matrix $\boldsymbol{\Delta}$, [Bou16] recently showed that (Theorem 4) when the noise $\boldsymbol{\Delta}$ is bounded in mild sense,

second-order necessary condition for optimality is also sufficient.

Particularly, this holds w.h.p. when the noise is i.i.d. complex Gaussians with small variance (Lemma 5). To understand the above statement, recall that second-order necessary condition asks for vanishing gradient and negative semidefinite Hessian at a point. The above statement asserts that such condition is sufficient to guarantee global optimality. In other words, at any critical points other than these verifying the condition have indefinite Hessians. Thus, [Bou16] has effectively showed that when $\boldsymbol{\Delta}$ is appropriately bounded,

the function $-\boldsymbol{x}^*\boldsymbol{C}\boldsymbol{x}$ over $\mathbb{C}_1^n$ is a "qualitative" $\mathcal{X}$ function. [11]

The real counterpart of phase synchronization is called *synchronization over* $\mathbb{Z}^2$, i.e., $\boldsymbol{z} \in \{1, -1\}^n$. In this case, an analogous formulation to (2.5) appears to be a hard combinatorial problem (think of `MAX-CUT`) in theory, and also not be friendly for numerical computation (the domain is discrete). Interestingly, [BBV16] showed certain *nonconvex* relaxation has a benign geometric structure. Specifically, applying the usual SDP lifting idea leads to

$$\text{minimize}_{\boldsymbol{X}\in\mathbb{R}^{n\times n}} \ -\langle \boldsymbol{X}, \boldsymbol{C}\rangle \quad X_{ii} = 1, \forall\, i, \quad \boldsymbol{X} \succeq \boldsymbol{0}, \ \ \text{rank}(\boldsymbol{X}) = 1.$$

Dropping the rank constraint results in a convex program (SDP), which is expensive to solve for large n. The Burer-Monteiro factorization approach [BM03, BM05] suggests substituting $\boldsymbol{X} = \boldsymbol{W}\boldsymbol{W}^\top$ for $\boldsymbol{W} \in \mathbb{R}^{n\times p}$ for $1 \le p \ll n$ such that the above relaxation is reformulated as

$$\text{minimize}_{\boldsymbol{W}\in\mathbb{R}^{n\times p}} \ -\text{tr}\left(\boldsymbol{W}^\top\boldsymbol{C}\boldsymbol{W}\right) \quad \|\boldsymbol{w}_i\| = 1 \ \forall\, i. \tag{2.6}$$

Classic results [Sha82, Bar95, Pat98] on this says (2.6) has the same optimal value as the SDP relaxation when p is large enough ($p \sim \Theta(\sqrt{n})$). Moreover, when p is set to be this scale, rank-deficient local optimizers are also global [BM05]. Surprisingly, [BBV16] showed (Theorem 4) that even $p = 2$, for the $\mathbb{Z}^2$ synchronization problem with small noise (i.e., small $\boldsymbol{\Delta}$), formulation (2.6) obeys

all points verifying the second-order necessary condition are global optimizes.

By analogous argument as for the complex case, this implies:

the function $-\text{tr}\left(\boldsymbol{W}^\top\boldsymbol{C}\boldsymbol{W}\right)$ over the oblique manifold $\left\{\boldsymbol{W} \in \mathbb{R}^{n\times 2} : \|\boldsymbol{w}_i\| = 1 \ \forall i \in [n]\right\}$ is a qualitative $\mathcal{X}$ function.

A similar result was derived in [BBV16] for the two-block community detection problem based on the stochastic block model (Theorem 6).[12]

[11] Strictly speaking, our definition of $\mathcal{X}$ functions requires the function to be locally *strongly* convex around the local/global minimizers, while the Hessian being positive semidefinite is weaker than that. No matter whether their result can be strengthened in this respect, we note that we impose the strong convexity assumption instead of just convexity is for the sake of deriving concrete convergence rates for optimization algorithms. One can relax the requirement when talking of the qualitative aspect of the structure. Similar comment applies to the ensuring discussion of the real version also.

[12] Both [BBV16] and [Mon16] also contain results that characterize local optimizers in terms of their correlation with the optimizer under less stringent/general conditions.

3 Second-order Trust-region Method and Proof of Convergence

The intuition that second-order information can help escape ridable saddles from the very start suggests a second-order method. We describe a *second-order trust-region algorithm on manifolds* [ABG07, AMS09] for this purpose.

For the generic problem (2.1), we start from any feasible $\boldsymbol{x}^{(0)} \in \mathcal{M}$, and form a sequence of iterates $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \cdots \in \mathcal{M}$ as follows. For the current iterate $\boldsymbol{x}^{(k)}$, we consider the quadratic approximation

$$\widehat{f}(\boldsymbol{\delta}; \boldsymbol{x}^{(k)}) \doteq f(\boldsymbol{x}^{(k)}) + \left\langle \boldsymbol{\delta}, \operatorname{grad} f(\boldsymbol{x}^{(k)}) \right\rangle + \frac{1}{2} \left\langle \operatorname{Hess} f(\boldsymbol{x}^{(k)})[\boldsymbol{\delta}], \boldsymbol{\delta} \right\rangle \tag{3.1}$$

which is defined for all $\boldsymbol{\delta} \in T_{\boldsymbol{x}^{(k)}}\mathcal{M}$. The next iterate is determined by minimizing the quadratic approximation within a small radius Δ (i.e., the trust region) of $\boldsymbol{x}^{(k)}$, i.e.,

$$\boldsymbol{\delta}^{(k+1)} \doteq \operatorname*{arg\,min}_{\boldsymbol{\delta} \in T_{\boldsymbol{x}^{(k)}}\mathcal{M}, \|\boldsymbol{\delta}\|_2 \leq \Delta} \widehat{f}\left(\boldsymbol{\delta}; \boldsymbol{x}^{(k)}\right), \tag{3.2}$$

which is called the Riemannian trust-region subproblem. The vector $\boldsymbol{x}^{(k)} + \boldsymbol{\delta}^{(k+1)}$ is generally not a point on $\mathcal{M}$. One then performs a *retraction* step $R_{\boldsymbol{x}^{(k)}}$ that pulls the vector back to the manifold, resulting in the update formula

$$\boldsymbol{x}^{(k+1)} = R_{\boldsymbol{x}^{(k)}}(\boldsymbol{x}^{(k)} + \boldsymbol{\delta}^{(k+1)}).$$

Most manifolds of practical interest are embedded submanifolds of $\mathbb{R}^{m \times n}$ and the tangent space is a subspace of $\mathbb{R}^{m \times n}$. For an $\boldsymbol{x}^{(k)} \in \mathcal{M}$ and an orthonormal basis $\boldsymbol{U}$ for $T_{\boldsymbol{x}^{(k)}}\mathcal{M}$, one can solve (3.2) by solving the recast Euclidean trust-region subproblem

$$\boldsymbol{\xi}^{(k+1)} \doteq \operatorname*{arg\,min}_{\|\boldsymbol{\xi}\| \leq \Delta} \widehat{f}(\boldsymbol{U}\boldsymbol{\xi}; \boldsymbol{x}^{(k)}), \tag{3.3}$$

for which efficient numerical algorithms exist [MS83, CGT00, FW04, HK14]. Design choice of the retraction is often problem-specific, ranging from the classical exponential map to the Euclidean projection that works for many matrix manifolds [AM12].

To show the trust-region algorithm converges to a global minimizer, we assume Δ is small enough such that approximation error of (3.1) to f is "negligible" locally. Each step around a negative-curvature or strong-gradient point decreases the objective by a certain amount. Indeed, it is clear there is always one descent direction in such cases. Thus, the trust-region step will approximately follow one descent direction and decrease the function value. When the iterate sequence moves into a strongly convex region around a global minimizer, a step is either constrained such that it also deceases the objective by an amount, or unconstrained, which is a good indicator that the target minimizer is within a radius Δ. In the latter case, the algorithm behaves like the classical Newton method and quadratic sequence convergence can be shown.

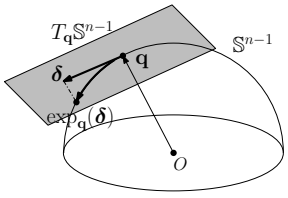

Figure 3: Illustrations of the tangent space $T_{\boldsymbol{q}}\mathbb{S}^{n-1}$ and exponential map $\exp_{\boldsymbol{q}}(\boldsymbol{\delta})$ defined on the sphere $\mathbb{S}^{n-1}$.

Quantitative convergence proof demands knowledge of the ridability parameters, smoothness parameters of the objective, and elements of Riemannian geometry. We refer the reader to [SQW15, SQW16] for practical examples of convergence analyses.

4 Discussion

Recently, there is a surge of interest in understanding nonconvex heuristics for practical problems [KMO10, JNS13, Har14, HW14, NNS+14, JN14, SL14, ZL15, TBSR15, CW15, NJS13, CLS15, CC15, WWS15, JO14, AGJ14b, AGJ14a, AJSN15, YCS13, SA14, LWB13, QSW14, LWB13, AAJ+13, AAN13, AGM13, AGMM15, ABGM14, JJKN15]. Majority of the work start from clever initializations, and then proceed with analysis of local convergence. In comparison, it is clear that for $\mathcal{X}$ functions, second-order trust-region algorithms with any initialization guarantee to retrieve one target minimizer. Identifying $\mathcal{X}$ functions has involved intensive technical work [SQW15, SQW16, GHJY15]. It is interesting to see if streamlined toolkits can be developed, say via operational rules or unified potential functions. This would facilitate study of other practical problems, such as the deep networks of which saddle points are believed to be prevalent and constitute significant computational bottleneck [PDGB14, DPG+14, CHM+14]. To match heuristics computationally, more practical algorithms other than the second-order trust-region methods are needed. Practical trust-region solvers with saddle-escaping capability may be possible for structured problems [BMAS14, SQW16]. Moreover, simulations with several practical problems suggest gradient-style algorithms with random initializations succeed. [GHJY15, LSJR16] are recent endeavors towards this direction.

References

[AAJ+13] Alekh Agarwal, Animashree Anandkumar, Prateek Jain, Praneeth Netrapalli, and Rashish Tandon, *Learning sparsely used overcomplete dictionaries via alternating minimization*, arXiv preprint arXiv:1310.7991 (2013).

[AAN13] Alekh Agarwal, Animashree Anandkumar, and Praneeth Netrapalli, *Exact recovery of sparsely used overcomplete dictionaries*, arXiv preprint arXiv:1309.1952 (2013).

[ABG07] Pierre-Antoine. Absil, Christopher G. Baker, and Kyle A. Gallivan, *Trust-region methods on Riemannian manifolds*, Foundations of Computational Mathematics **7** (2007), no. 3, 303–330.

[ABGM14] Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma, *More algorithms for provable dictionary learning*, arXiv preprint arXiv:1401.0579 (2014).

[AG16] Anima Anandkumar and Rong Ge, *Efficient approaches for escaping higher order saddle points in non-convex optimization*, arXiv preprint arXiv:1602.05908 (2016).

[AGJ14a] Animashree Anandkumar, Rong Ge, and Majid Janzamin, *Analyzing tensor power method dynamics: Applications to learning overcomplete latent variable models*, arXiv preprint arXiv:1411.1488 (2014).

[AGJ14b] _____, *Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates*, arXiv preprint arXiv:1402.5180 (2014).

[AGM13] Sanjeev Arora, Rong Ge, and Ankur Moitra, *New algorithms for learning incoherent and overcomplete dictionaries*, arXiv preprint arXiv:1308.6273 (2013).

[AGMM15] Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra, *Simple, efficient, and neural algorithms for sparse coding*, arXiv preprint arXiv:1503.00778 (2015).

[AGMS12] Sanjeev Arora, Rong Ge, Ankur Moitra, and Sushant Sachdeva, *Provable ICA with unknown gaussian noise, with implications for gaussian mixtures and autoencoders*, Advances in Neural Information Processing Systems, 2012, pp. 2375–2383.

[AJSN15] Animashree Anandkumar, Prateek Jain, Yang Shi, and Uma Naresh Niranjan, *Tensor vs matrix methods: Robust tensor decomposition under block sparse perturbations*, arXiv preprint arXiv:1510.04747 (2015).

[AM12] P.-A. Absil and Jérôme Malick, *Projection-like retractions on matrix manifolds*, SIAM Journal on Optimization **22** (2012), no. 1, 135–158.

[AMS09] Pierre-Antoine. Absil, Robert Mahoney, and Rodolphe Sepulchre, *Optimization algorithms on matrix manifolds*, Princeton University Press, 2009.

[Bar95] Alexander I. Barvinok, *Problems of distance geometry and convex properties of quadratic maps*, Discrete & Computational Geometry **13** (1995), no. 2, 189–202.

[BBV16] Afonso S Bandeira, Nicolas Boumal, and Vladislav Voroninski, *On the low-rank approach for semidefinite programs arising in synchronization and community detection*, arXiv preprint arXiv:1602.04426 (2016).

[Ber99] Dimitri P. Bertsekas, *Nonlinear programming*, Athena scientific, 1999.

[BM03] Samuel Burer and Renato D.C. Monteiro, *A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization*, Mathematical Programming **95** (2003), no. 2, 329–357.

[BM05] Samuel Burer and Renato D. C. Monteiro, *Local minima and convergence in low-rank semidefinite programming*, Mathematical Programming **103** (2005), no. 3, 427–444.

[BMAS14] Nicolas Boumal, Bamdev Mishra, P.-A. Absil, and Rodolphe Sepulchre, *Manopt, a Matlab toolbox for optimization on manifolds*, Journal of Machine Learning Research **15** (2014), 1455–1459.

[Bou16] Nicolas Boumal, *Nonconvex phase synchronization*, arXiv preprint arXiv:1601.06114 (2016).

[CC15] Yuxin Chen and Emmanuel J. Candès, *Solving random quadratic systems of equations is nearly as easy as solving linear systems*, arXiv preprint arXiv:1505.05114 (2015).

[CGT00] Andrew R. Conn, Nicholas I. M. Gould, and Philippe L. Toint, *Trust-region methods*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2000.

[CHM+14] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun, *The loss surface of multilayer networks*, arXiv preprint arXiv:1412.0233 (2014).

[CLS15] Emmanuel J. Candès, Xiaodong Li, and Mahdi Soltanolkotabi, *Phase retrieval via wirtinger flow: Theory and algorithms*, Information Theory, IEEE Transactions on **61** (2015), no. 4, 1985–2007.

[CW15] Yudong Chen and Martin J. Wainwright, *Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees*, arXiv preprint arXiv:1509.03025 (2015).

[DPG+14] Yann N. Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio, *Identifying and attacking the saddle point problem in high-dimensional non-convex optimization*, Advances in Neural Information Processing Systems, 2014, pp. 2933–2941.

[FJK96] Alan Frieze, Mark Jerrum, and Ravi Kannan, *Learning linear transformations*, focs, IEEE, 1996, p. 359.

[FW04] Charles Fortin and Henry Wolkowicz, *The trust region subproblem and semidefinite programming*, Optimization methods and software **19** (2004), no. 1, 41–67.

[GHJY15] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan, *Escaping from saddle points—online stochastic gradient for tensor decomposition*, Proceedings of The 28th Conference on Learning Theory, 2015, pp. 797–842.

[Har14] Moritz Hardt, *Understanding alternating minimization for matrix completion*, Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on, IEEE, 2014, pp. 651–660.

[HK14] Elad Hazan and Tomer Koren, *A linear-time algorithm for trust region problems*, arXiv preprint arXiv:1401.6757 (2014).

[HL13] Christopher J. Hillar and Lek-Heng Lim, *Most tensor problems are NP-hard*, Journal of the ACM (JACM) **60** (2013), no. 6, 45.

[HW14] Moritz Hardt and Mary Wootters, *Fast matrix completion without the condition number*, Proceedings of The 27th Conference on Learning Theory, 2014, pp. 638–678.

[JJKN15] Prateek Jain, Chi Jin, Sham M. Kakade, and Praneeth Netrapalli, *Computing matrix squareroot via non convex local search*, arXiv preprint arXiv:1507.05854 (2015).

[JN14] Prateek Jain and Praneeth Netrapalli, *Fast exact matrix completion with finite samples*, arXiv preprint arXiv:1411.1087 (2014).

[JNS13] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi, *Low-rank matrix completion using alternating minimization*, Proceedings of the forty-fifth annual ACM symposium on Theory of Computing, ACM, 2013, pp. 665–674.

[JO14] Prateek Jain and Sewoong Oh, *Provable tensor factorization with missing data*, Advances in Neural Information Processing Systems, 2014, pp. 1431–1439.

[KMO10] Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh, *Matrix completion from a few entries*, Information Theory, IEEE Transactions on **56** (2010), no. 6, 2980–2998.

[LSJR16] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht, *Gradient descent converges to minimizers*, arXiv preprint arXiv:1602.04915 (2016).

[LWB13] Kiryung Lee, Yihong Wu, and Yoram Bresler, *Near optimal compressed sensing of sparse rank-one matrices via sparse power factorization*, arXiv preprint arXiv:1312.0525 (2013).

[MK87] Katta G. Murty and Santosh N. Kabadi, *Some NP-complete problems in quadratic and nonlinear programming*, Mathematical programming **39** (1987), no. 2, 117–129.

[Mon16] Andrea Montanari, *A grothendieck-type inequality for local maxima*, arXiv preprint arXiv:1603.04064 (2016).

[MS83] Jorge J. Moré and Danny C. Sorensen, *Computing a trust region step*, SIAM Journal on Scientific and Statistical Computing **4** (1983), no. 3, 553–572.

[NJS13] Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi, *Phase retrieval using alternating minimization*, Advances in Neural Information Processing Systems, 2013, pp. 2796–2804.

[NNS+14] Praneeth Netrapalli, Uma Naresh. Niranjan, Sujay Sanghavi, Animashree Anandkumar, and Prateek Jain, *Non-convex robust PCA*, Advances in Neural Information Processing Systems, 2014, pp. 1107–1115.

[Pat98] Gábor Pataki, *On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues*, Mathematics of operations research **23** (1998), no. 2, 339–358.

[PDGB14] Razvan Pascanu, Yann N Dauphin, Surya Ganguli, and Yoshua Bengio, *On the saddle point problem for non-convex optimization*, arXiv preprint arXiv:1405.4604 (2014).

[QSW14] Qing Qu, Ju Sun, and John Wright, *Finding a sparse vector in a subspace: Linear sparsity using alternating directions*, Advances in Neural Information Processing Systems, 2014, pp. 3401–3409.

[RI10] Simeon Reich and Aleksandr Davidovich Ioffe, *Nonlinear analysis and optimization: Optimization*, vol. 2, American Mathematical Soc., 2010.

[SA14] Hanie Sedghi and Animashree Anandkumar, *Provable tensor methods for learning mixtures of classifiers*, arXiv preprint arXiv:1412.3046 (2014).

[SEC+15] Yoav Shechtman, Yonina C. Eldar, Oren Cohen, Henry N. Chapman, Jianwei Miao, and Mordechai Segev, *Phase retrieval with application to optical imaging: A contemporary overview*, Signal Processing Magazine, IEEE **32** (2015), no. 3, 87–109.

[Sha82] Alexander Shapiro, *Rank-reducibility of a symmetric matrix and sampling theory of minimum trace factor analysis*, Psychometrika **47** (1982), no. 2, 187–199.

[SL14] Ruoyu Sun and Zhi-Quan Luo, *Guaranteed matrix completion via non-convex factorization*, arXiv preprint arXiv:1411.8003 (2014).

[SQW15] Ju Sun, Qing Qu, and John Wright, *Complete dictionary recovery over the sphere*, arXiv preprint arXiv:1504.06785 (2015).

[SQW16] ———, *A geometric analysis of phase retreival*, arXiv preprint arXiv:1602.06664 (2016).

[TBSR15] Stephen Tu, Ross Boczar, Mahdi Soltanolkotabi, and Benjamin Recht, *Low-rank solutions of linear matrix equations via procrustes flow*, arXiv preprint arXiv:1507.03566 (2015).

[WWS15] Chris D. White, Rachel Ward, and Sujay Sanghavi, *The local convexity of solving quadratic equations*, arXiv preprint arXiv:1506.07868 (2015).

[YCS13] Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi, *Alternating minimization for mixed linear regression*, arXiv preprint arXiv:1310.3745 (2013).

[ZL15] Qinqing Zheng and John Lafferty, *A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements*, arXiv preprint arXiv:1506.06081 (2015).