On the geometric analysis of a quartic–quadratic optimization problem under a spherical constraint

Haixiang Zhang\textsuperscript{1} · Andre Milzarek\textsuperscript{2,3,4} · Zaiwen Wen\textsuperscript{5,6} · Wotao Yin\textsuperscript{7}

Received: 9 August 2019 / Accepted: 10 August 2021 / Published online: 21 August 2021
© Springer-Verlag GmbH Germany, part of Springer Nature and Mathematical Optimization Society 2021

Abstract
This paper considers the problem of solving a special quartic–quadratic optimization problem with a single sphere constraint, namely, finding a global and local minimizer of $\frac{1}{2}z^*Az + \frac{\beta}{2} \sum_{k=1}^n |z_k|^4$ such that $\|z\|_2 = 1$. This problem spans multiple domains including quantum mechanics and chemistry sciences and we investigate the geometric properties of this optimization problem. Fourth-order optimality conditions are derived for characterizing local and global minima. When the matrix in the quadratic term is diagonal, the problem has no spurious local minima and global solutions can be represented explicitly and calculated in $O(n \log n)$ operations. When $A$ is a rank one matrix, the global minima of the problem are unique under certain phase shift schemes. The strict-saddle property, which can imply polynomial time convergence of second-order-type algorithms, is established when the coefficient $\beta$ of the quartic term is either at least $O(n^{3/2})$ or not larger than $O(1)$. Finally, the Kurdyka–Łojasiewicz exponent of quartic–quadratic problem is estimated and it is shown that the largest exponent is at least $1/4$ for a broad class of stationary points.

Keywords Constrained quartic–quadratic optimization · Geometric analysis · Strict-saddle property · Łojasiewicz inequality

Mathematics Subject Classification 15A45 · 47H60 · 58K30 · 58C40 · 90C26
1 Introduction

In this paper, we analyze the geometric properties of the following nonconvex quartic–quadratic problem under a single spherical constraint,

$$\min_{z \in \mathbb{C}^n} f(z) = \frac{1}{2} z^* A z + \frac{\beta}{2} \sum_{k \in [n]} |z_k|^4 \quad \text{s.t.} \quad \|z\|_2 = 1,$$

where $\beta > 0$ is a fixed interaction coefficient and $A \in \mathbb{C}^{n \times n}$ is a given Hermitian matrix. The so-called Bose–Einstein condensation (BEC) problem is an important application and examples that can be expressed using the optimization model (1.1) have attracted great interests in the atomic, molecule and optical physics community and in the condense matter community. In particular, utilizing a proper non-dimensionalization and discretization, the BEC problem can be written as a quartic–quadratic minimization problem of the form (1.1), where the matrix $A$ corresponds to the sum of the discretized Laplace operator and a diagonal matrix. If a non-rotating BEC problem is considered, then the variable $z$ can be restricted to the real space $\mathbb{R}^n$ and problem (1.1) becomes a real optimization problem. For a more detailed setup of the BEC problem and its specific mathematical formulation, we refer to [11,38,65].

Our interest in problem (1.1) and its geometric properties is primarily triggered by related numerical results and observations with Bose–Einstein condensates and Kohn–Sham density functional calculations, see, e.g., [32,41,78,79], and is motivated by recent landscape results for matrix completion [35,36,76], phase retrieval [24,72], phase synchronization [10,17,54], and quadratic programs with spherical constraints [31,53]. Understanding the geometric landscape of the nonconvex optimization problem (1.1) is a fundamental step towards understanding and explaining the global and local behavior of the problem and the performance of associated algorithms. Despite recent progress on the geometric properties of nonconvex minimization problems and due to the complex interaction of the quadratic and quartic terms, the landscape of (1.1) is still elusive. We further note that in [40], Hu et al. have shown that the minimization problem (1.1) can be interpreted as a special instance of the partition problem and thus, it is generally NP-hard to solve (1.1).

1.1 Related work and geometric concepts

Although nonconvex optimization problems are generally NP-hard [61], direct and traditional minimization approaches, such as basic gradient and trust region schemes, can still be applied to solve certain important classes of nonconvex problems—with astonishing success—and they remain the methods of choice for the practitioner [17]. A recent and steadily growing area of research concentrates on the identification of such classes of problems and tries to close the discrepancy between theoretical results and numerical performances, see, e.g., [26,43,71] for an overview. Herein geometric observations and techniques play a major role in understanding the landscape and the
global and local behavior of a nonconvex problem and of corresponding algorithms. Specifically, we are interested in the following geometric properties:

\((P_1)\) All local minimizers are also global solutions, i.e., there are no spurious local minimizers.

\((P_2)\) The objective function possesses negative curvature directions at all saddle points and local maximizers which allows to effectively escape those points.

Condition \((P_2)\) is the basis of the so-called strict-saddle property and was introduced in [34,35,72]. The strict-saddle and other related conditions can be used in the convergence analysis and in the design of algorithms to efficiently avoid saddle points. For instance, Sun et al. [75] established a polynomial-time convergence rate of a Riemannian trust region method that is tailored to solve phase retrieval problems which satisfy the strict-saddle property. Furthermore, in [48,49,63] it is shown that certain randomly initialized first-order methods can converge to local minimizers and escape saddle points almost surely if the strict-saddle property holds. In the following, we briefly review recent classes of nonconvex optimization problems for which the conditions \((P_1), (P_2)\), or other desirable geometric properties are satisfied.

The generalized phase retrieval (GPR) problem is a popular nonconvex problem which has seen remarkable progress these years, see, e.g., [42,69] for an overview. Classical methods that transform the GPR problem into a convex program include convex relaxation techniques [19,21,25] and Wirtinger flow algorithms with carefully-designed initialization [20]. Phase retrieval problems are typically formulated as a quartic and unconstrained least squares problem depending on \(m\) measurements \(y_k = | a_k^* z |, k = 1, \ldots, m\). Traditional GPR methods can recover the true signal \(z\) from the measurements as long as the sample size \(m\) satisfies \(m \gtrapprox n\) or \(m \gtrapprox n \log n\) where \(n\) is the dimension of the signal. A provable convergence rate for a randomly initialized trust region-type algorithm is given in [72] as long as \(m \gtrapprox n \log^3 n\) via showing that all the local minima are global and the strict-saddle property holds. When the signal and observations are real, the convergence rate of the vanilla gradient descent method is established by Chen et al. [24] under the assumption \(m \gtrapprox n \log^{13} n\). Another interesting class of amenable nonconvex problems are low-rank matrix factorization problems. Classical methods for matrix factorization are based on nuclear norm minimization [22,67] and are usually memory intensive or require long running times. In [44,45], Keshavan, Montanari, and Oh showed that the well-initialized gradient descent method can recover the ground truth of those problems. A strong convexity-type property is proved to hold around the optimal solution by Sun and Luo in [76] and the objective function is shown to be sharp and weakly convex in nonsmooth settings by Li et al. [51]. Further, the strict-saddle property for the low-rank matrix factorization problem is established in [35,36], as well as for other low rank problems such as robust PCA and matrix sensing. Other classes of nonconvex optimization problems with provable convergence or geometric properties comprise orthogonal tensor decomposition [34,37], complete dictionary learning [5,73,74], phase synchronization and community detection [10,17,54] and shallow neural networks [52]. There are also several numerical methods that work well in practice for solving the BEC problem (e.g., tools for numerical partial differential equations [3,29] or optimization methods [33,40,79]), but their geometric properties are not well understood.
So far the mentioned concepts allow to cover global structures and landscapes. Instead, local properties and the local behavior of (1.1) can be captured by the so-called Kurdyka–Łojasiewicz (KL) [47], or Łojasiewicz inequality [55]. The Łojasiewicz inequality is a useful tool to estimate the convergence rate of first-order iterative methods in the nonconvex setting [1,59,68]. Moreover, the convergence rate of first-order methods satisfying a certain line-search criterion and descent condition can be derived via the KL inequality [6,15,68], where the rate depends on the KL exponent \( \theta \). However, there is no general method to determine or estimate the KL exponent of specific optimization problems, though the existence of the KL exponent is guaranteed in many situations. For optimizing a real analytic function over a compact real analytic manifold (such as problem (1.1)), the existence of the KL exponent is established by Łojasiewicz in [55]. There are also several few works that derive explicit estimates of the KL exponent for certain structured problems, such as general polynomials [27,39,80], convex problems [50], non-convex quadratic optimization problems with simple convex constraints [30,50,56,57], and quadratic optimization problems with single spherical constraint [31,53]. Obviously, the above four cases do not cover our constrained quartic–quadratic optimization problem (1.1).

1.2 Contributions

In this work, we investigate different geometric concepts for the quadratic-quartic optimization problem (1.1) and give theoretical explanations why first- and second-order methods can perform well on it. In Sect. 2, we first derive several new second- and fourth-order optimality conditions for problem (1.1) that can be utilized to characterize local and global solutions. These conditions capture fundamental geometric properties of stationary points and local minima and form the basis of our geometric analysis. We then investigate problem (1.1) in the special case where \( A \) is a diagonal matrix. In this situation, we show that a complete characterization of the landscape can be obtained and that problem (1.1) does not possess any spurious local minima. Furthermore, global solutions can be computed explicitly using a closed-form expression that involves the projection onto an \( n \)-simplex which requires \( O(n \log n) \) operations. These results can be partially extended to the case where \( A \) is a rank-one matrix. In general, the complex interplay between the quartic and quadratic terms impedes the derivation of explicit expressions for stationary points and local minima and complicates the landscape analysis of \( f \) significantly. However, if either the quartic or the quadratic term dominates the objective function, we can establish the strict-saddle property (\( P_2 \)) and identify and calculate the location and number of local minima. Our methodology is based on a careful discussion of the quartic and quadratic terms for large and small interaction coefficients that is applicable for general deterministic and arbitrary choices of \( A \). We note that previous works and results only cover fourth-order unconstrained optimization problems (e.g., phase retrieval), quadratic constrained optimization problems (e.g., matrix completion and phase synchronization), or fourth-order constrained optimization problems without quadratic terms (e.g., fourth-order tensor decomposition). In particular, there is no interaction between quartic and quadratic terms and between their Riemannian derivatives. Different from most nonconvex problems discussed in
the literature, our problem does not have a natural probabilistic framework and thus, probabilistic techniques such as concentration inequalities can not be directly applied.

In addition, we estimate the largest KL exponent and establish a Riemannian Łojasiewicz-type inequality for problem (1.1). Again, the presence of the quartic term considerably complicates the theoretical analysis. In order to deal with the high-order terms appearing in the Taylor expansion, we first separate the nonzero and zero components of a stationary point in order to facilitate the discussion of the leading terms. The appearance of the quartic term requires the third-order and the fourth-order terms in the Taylor expansion to fully describe the local behavior, rather than merely the second-order terms. Due to the additional terms, the number of possible leading terms is significantly increased and we carefully analyze the relationship between those different terms. If the matrix $A$ is diagonal, we show that the Łojasiewicz inequality holds at every stationary point of (1.1) with exponent $\theta = \frac{1}{4}$. Moreover, this result can be extended to more general choices of $A$, if the problem is restricted to the real space and positive semi-definiteness of the stationary certification matrix is assumed. The proof is based on the diagonal case and on estimates of the local behavior of the objective function and the Riemannian gradient in different subspaces. The positive semi-definiteness assumption is utilized throughout the proof to handle the non-isolated case and can not be easily removed. Although this additional condition represents a stronger notion of global optimality, a wide range of global minima in the real case satisfy this condition. To the best of the authors’ knowledge, our work is the first to estimate and analyze these properties for quadratic-quartic optimization problems over a single sphere.

1.3 Organization and notations

This paper is organized as follows. In Sect. 2, we present second- and fourth-order optimality conditions and characterize global minimizer of problem (1.1). Next, in Sects. 3 and 4, we consider two special cases and investigate geometric properties of problem (1.1) when $A$ is either diagonal or has rank one. General landscape results for the real case are discussed in Sect. 5. Finally, in Sect. 6, we estimate the KL exponent of problem (1.1).

For $n \in \mathbb{N}$, we define $[n] := \{1, \ldots, n\}$ and for $z \in \mathbb{C}^n$, we set $\|z\| = \|z\|_2 = \sqrt{z^*z}$. Let $S^{n-1}$ and $CS^{n-1}$ denote the $n$-dimensional real and complex sphere, respectively. In the following sections, we will use the notation $M = S^{n-1}$ or $M = CS^{n-1}$ depending on whether we consider the real or the complex case. The tangent space of $CS^{n-1}$ at a point $z \in CS^{n-1}$ is given by $T_zM := \{v \in \mathbb{C}^n : \Re (v^*z) = 0\}$.

For $z \in \mathbb{C}^n$, $\text{diag}(z)$ is a diagonal matrix with diagonal entries $z_1, \ldots, z_n$ and we use $|z|^2$ to denote the component-wise absolute value, $|z|^2 = z \odot \bar{z}$, of $z$. We use $I$ to denote the $(n \times n)$ identity matrix. The Euclidean and corresponding Riemannian gradient of $f$ at $z$ on $M$ are denoted by $\nabla f(z)$ and $\text{grad } f(z)$. Similarly, $\nabla^2 f(z)$ and $\text{Hess } f(z)$ represent the Euclidean and Riemannian Hessian, respectively.

Throughout this paper and without loss of generality, we will assume that the matrix $A$ is positive definite. Furthermore, $A = P \Lambda P^*$ is an associated eigenvalue decomposition of the Hermitian matrix $A$ with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n > 0$, $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$, $P = (p_1, \ldots, p_n) \in \mathbb{C}^{n \times n}$, and $P^*P = I$. [Springer]
2 Wirtinger calculus and optimality conditions

Since the real-valued objective function \( f \) is nonanalytic in \( z \), we utilize the Wirtinger calculus \([46,70]\) to express the complex derivatives of \( f \). Specifically, the Wirtinger gradient and Hessian of \( f \) are defined as

\[
\nabla f(z) := \begin{bmatrix} \nabla_z f(z) \\ \nabla_{\bar{z}} f(z) \end{bmatrix}, \quad \nabla^2 f(z) := \begin{bmatrix} \frac{\partial}{\partial z} \left( \frac{\partial f}{\partial z} \right)^* & \frac{\partial}{\partial z} \left( \frac{\partial f}{\partial \bar{z}} \right)^* \\ \frac{\partial}{\partial \bar{z}} \left( \frac{\partial f}{\partial z} \right)^* & \frac{\partial}{\partial \bar{z}} \left( \frac{\partial f}{\partial \bar{z}} \right)^* \end{bmatrix},
\]

where \( \nabla_z f(z) := \left( \frac{\partial f}{\partial z} \right)^* \), \( \nabla_{\bar{z}} f(\bar{z}) := \left( \frac{\partial f}{\partial \bar{z}} \right)^* \) and following \([79]\), we obtain

\[
\nabla f(z) = \frac{1}{2} A z + \beta \text{diag}(|z|^2) z, \quad \nabla_{\bar{z}} f(\bar{z}) = \nabla f(z), \quad \nabla^2 f(z) = \begin{bmatrix} \frac{1}{2} A + 2\beta \text{diag}(|z|^2) & \beta \text{diag}(z_1^2, \ldots, z_n^2) \\ \beta \text{diag}(\bar{z}_1^2, \ldots, \bar{z}_n^2) & \frac{1}{2} \bar{A} + 2\beta \text{diag}(|z|^2) \end{bmatrix}.
\]

Furthermore, using the identification \( T_z M \equiv \{ v, \bar{v} \in \mathbb{C}^n : z^* v + z^T \bar{v} = 0 \} \), the Riemannian gradient and Hessian of \( f \) are given by

\[
\text{grad} f(z) = \nabla f(z) - \lambda \begin{bmatrix} z \\ \bar{z} \end{bmatrix} \text{ and } \text{Hess} f(z) = \nabla^2 f(z) - \lambda I_{2n}, \quad (2.1)
\]

where

\[
\lambda = z^* \nabla_z f(z) = \frac{1}{2} z^* A z + \beta \|z\|^4_4 \in \mathbb{R}, \quad (2.2)
\]

see, e.g., \([2, \text{Section 3.6 and 5.5}]\). Exploiting the symmetry in \( \text{grad} f \), the associated first-order optimality conditions for \((1.1)\) now take the form:

\[
\text{grad} f(z) = 0 \iff [A + 2\beta \text{diag}(|z|^2)] z = 2\lambda z. \quad (2.3)
\]

A point \( z \in \mathbb{C}^n \) satisfying the conditions \((2.3)\) will be called \emph{stationary point} of problem \((1.1)\). We define the quadratic form induced by the Hessian \( \text{Hess} f(z) \) along a direction \( v \in \mathbb{C}^n \) via

\[
H_f(z)[v] := \begin{bmatrix} v^* \\ v^T \end{bmatrix} \text{Hess} f(z) \begin{bmatrix} v \\ \bar{v} \end{bmatrix} = v^*[A + 4\beta \text{diag}(|z|^2)] v + 2\beta \sum_{k=1}^n \Re(v_k^2 \bar{z}_k^2) - 2\lambda \|v\|^2 = v^*[A + 2\beta \text{diag}(|z|^2) - 2\lambda I] v + 4\beta \sum_{k=1}^n \Re(v_k \bar{z}_k)^2. \quad (2.4)
\]
Let us note that this expression coincides with the standard curvature of $f$ at $z$ along the vector $v$ if $v \in \mathbb{C}S^{n-1}$. In the real case, the latter formulae reduce to grad $f(z) = [A + 2\beta\text{diag}(|z|^2)]z - 2\lambda z$, Hess $f(z) = A + 6\beta\text{diag}(|z|^2) - 2\lambda I_n$, and $H_f(z)[v] := v^T \text{Hess } f(z)v$.

### 2.1 Second-order optimality conditions

Using the Riemannian optimality conditions discussed in [81, Theorem 4.2 and 4.3], the second-order optimality conditions for (1.1) can be expressed as follows:

**Lemma 2.1** (Second-order necessary and sufficient conditions) Suppose that $z$ is a local solution of problem (1.1). Then, it holds that grad $f(z) = 0$ and we have $H_f(z)[v] \geq 0$ for all $v \in T_zM \cap \mathbb{C}S^{n-1}$. Conversely, if $z$ is a stationary point satisfying grad $f(z) = 0$ and $H_f(z)[v] > 0$ for all $v \in T_zM \cap \mathbb{C}S^{n-1}$, then $z$ is an isolated local minimum of problem (1.1).

Next, for some $z \in \mathbb{C}^n$ we define the equivalence class

$$[z] := \{y \in \mathbb{C}^n : |y_k| = |z_k|, \forall k \in [n]\}. \quad (2.5)$$

The following theorem gives a general sufficient condition for a stationary point to be a global minimum of problem (1.1).

**Theorem 2.1** Let $z \in \mathbb{C}^n$ be a stationary point of problem (1.1) and let $\lambda \in \mathbb{R}$ be given as in (2.2). Furthermore, suppose that the matrix

$$H := A + 2\beta\text{diag}(|z|^2) - 2\lambda I \succeq 0 \quad (2.6)$$

is positive semidefinite. Then $z$ is a global minimum and all global minima of problem (1.1) belong to the equivalence class $[z]$.

**Proof** Let $y \in \mathbb{C}^n$ be an arbitrary point with $\|y\| = 1$ and let us introduce the polar coordinates $z_i = r_i e^{i\theta_i}$, $y_i = t_i e^{i\phi_i}$ for $r_i, t_i \geq 0$, $\theta_i, \phi_i \in [0, 2\pi]$ and all $i \in [n]$. Using the stationarity condition grad $f(z) = 0$ and $\|z\| = \|y\| = 1$, it holds that

$$f(y) - f(z) = \frac{1}{2}y^* Ay - \frac{1}{2}z^*[2\lambda z - 2\beta\text{diag}(|z|^2)z] + \frac{\beta}{2} (\|y\|_4^4 - \|z\|_4^4)$$

$$= \frac{1}{2}y^* H y - \beta y^* \text{diag}(|z|^2)y + \frac{\beta}{2} (\|y\|_4^4 + \|z\|_4^4)$$

$$= \frac{1}{2}y^* H y + \frac{\beta}{2} \sum_{k=1}^{n} (t_k^2 - r_k^2)^2. \quad (2.7)$$

Consequently, the positive semidefiniteness of $H$ yields $f(y) - f(z) \geq 0$ for all $y \in \mathbb{C}^n$ with $\|y\| = 1$. Suppose now that $y$ is a global minimum with $y \notin [z]$. In this case the last sum in the above expression is strictly positive which, together with the positive semidefiniteness of $H$, yields a contradiction. \hfill $\Box$
If problem (1.1) has two different global minimizers \( y \) and \( z \) with \([y] \cap [z] = \emptyset\), Theorem 2.1 implies that \( H \) can not be positive semidefinite. Moreover, if condition (2.6) holds at a stationary point \( z \), it automatically has to hold at all global minimizers in \([z]\).

The definiteness condition in Theorem 2.1 can be equivalently rephrased as follows: the scalar \( \lambda = z^* \nabla_z f(z) \) associated with \( z \) is the minimum eigenvalue of the matrix \( A + 2\beta \text{diag}(|z|^2) \) and \( z \) is the corresponding eigenvector. Characterizations of this type are also known for (quadratic) trust-region subproblems and for general quadratic programs with quadratic constraints, see [60,77]. Furthermore, utilizing [18, Theorem 3.1], it can be shown that such an eigenvector \( z \) with the stated properties exists under the assumption \( 0 < \beta \leq \frac{(\lambda_n - 1 - \lambda_n)}{8} \). In this case, the condition (2.6) is necessary and sufficient for global optimality.

2.2 Fourth-order optimality conditions

In the following section, we derive several fourth-order optimality conditions based on a special and finer expansion of the objective function \( f \). In contrast to the sufficient conditions in Theorem 2.1, this allows us to fully characterize global optima. Let \( z \in CS^{n-1} \) be an arbitrary stationary point. For \( v \in T_z M \cap CS^{n-1} \) and \( \theta \in \mathbb{R} \), we consider the point \( y = \cos(\theta)z + \sin(\theta)v \in CS^{n-1} \). Using this decomposition in (2.7), \( H_z = 0 \), and \(|y_k|^2 = \cos^2(\theta)|z_k|^2 + 2 \sin(\theta) \cos(\theta) \Re(v_k \bar{z}_k) + \sin^2(\theta)|v_k|^2 \), we have

\[
\begin{align*}
f(y) - f(z) &= \frac{1}{2} y^* H y + \frac{\beta}{2} \sum_{k \in [n]} \left[ |y_k|^2 - |z_k|^2 \right]^2 \\
&= \frac{1}{2} \sin^2(\theta) \cdot v^* H v + 2 \beta \sin^2(\theta) \cos(\theta) \cdot \sum_{k \in [n]} \Re(\bar{z}_k v_k)^2 \\
&\quad + 2 \beta \sin^3(\theta) \cos(\theta) \cdot \sum_{k \in [n]} (|v_k|^2 - |z_k|^2) \Re(\bar{z}_k v_k) \\
&\quad + \frac{\beta}{2} \sin^4(\theta) \cdot \sum_{k \in [n]} (|v_k|^2 - |z_k|^2)^2 \\
&= \frac{1}{2} \sin^2(\theta) \cdot v^* H v + 2 \beta \sin^2(\theta) \cdot \sum_{k \in [n]} \Re(\bar{z}_k v_k)^2 \\
&\quad + 2 \sin^3(\theta) \cos(\theta) \cdot H_3(v) + \frac{1}{2} \sin^4(\theta) \cdot H_4(v) \\
&= \frac{1}{2} \sin^2(\theta) H_f(z)[v] + 2 \sin^3(\theta) \cos(\theta) \cdot H_3(v) + \frac{1}{2} \sin^4(\theta) \cdot H_4(v) \\
&= \frac{1}{2} \sin^2(\theta) \cos^2(\theta) \cdot H_f(z)[v] + 2 \sin^3(\theta) \cos(\theta) \cdot H_3(v) \\
&\quad + \frac{1}{2} \sin^4(\theta) \left[ H_f(z)[v] + H_4(v) \right],
\end{align*}
\]

where the terms \( H_3 \) and \( H_4 \) are defined via
On the geometric analysis of a quartic–quadratic...

\[ H_3(v) := \beta \cdot \sum_{k \in [n]} (|v_k|^2 - |z_k|^2) \Re(\bar{z}_k v_k), \]

\[ H_4(v) := \beta \cdot \sum_{k \in [n]} [(|v_k|^2 - |z_k|^2)^2 - 4\Re(\bar{z}_k v_k)^2]. \]

Setting \( \theta = \pi/2 \), this yields

\[ 2[f(v) - f(z)] = H_f(z)[v] + H_4(v) \quad \forall \, v \in T_z\mathcal{M} \cap \mathbb{C}^{n-1}. \quad (2.8) \]

Finally, dividing \( f(y) - f(z) \) by \( \sin^4(\theta) \) and using the relation (2.8), we obtain

\[ \frac{2[f(y) - f(z)]}{\sin^4(\theta)} = H_f(z)[v] \cdot \cot^2(\theta) + 4H_3(v) \cdot \cot(\theta) + 2[f(v) - f(z)] \]

\[ = G(v, \cot(\theta)), \quad (2.9) \]

for \( \theta \neq k\pi, k \in \mathbb{Z} \), where the function \( G : \mathbb{C}^n \times \mathbb{R} \to \mathbb{R} \) is defined as follows:

\[ G(v, t) := H_f(z)[v] \cdot t^2 + 4H_3(v) \cdot t + 2[f(v) - f(z)]. \]

We first propose necessary and sufficient optimality conditions for \( z \) being a local minimum.

**Theorem 2.2** (Characterization of local optimality) *Let \( z \) be a stationary point of problem (1.1). Then \( z \) is locally optimal if and only if there exists a constant \( M_z > 0 \) such that \( G(v, t) \geq 0 \) for all \( v \in T_z\mathcal{M} \cap \mathbb{C}^{n-1} \) and all \( t \) with \( |t| \geq M_z \).*

**Proof** By definition, \( z \) is a local minimum if and only if there exists a constant \( \delta_z > 0 \) such that \( f(y) \geq f(z) \) for all \( y \in B(z, \delta_z) \cap \mathbb{C}^{n-1} \). Due to \( ||y - z||^2 = 2 - 2\cos(\theta) \), the condition \( y \in B(z, \delta_z) \cap \mathbb{C}^{n-1} \) is equivalent to \( y = \cos(\theta)z + \sin(\theta)v \) with \( v \in T_z\mathcal{M} \cap \mathbb{C}^{n-1} \) and \( \cos(\theta) \geq 1 - \delta_z^2/2 \). Moreover, \( \cos(\theta) \geq 1 - \delta_z^2/2 \) implies

\[ |\sin(\theta)| \leq \delta_z \sqrt{1 - \delta_z^2/4}, \quad |\cot(\theta)| \geq \frac{1 - \delta_z^2/2}{\delta_z \sqrt{1 - \delta_z^2/4}} := M_z. \quad (2.10) \]

Consequently, it holds that \( y \in B(z, \delta_z) \cap \mathbb{C}^{n-1} \) if and only if \( v \in T_z\mathcal{M} \cap \mathbb{C}^{n-1} \) and \( |\cot(\theta)| \geq M_z \). Further, by definition of \( G \), the necessary and sufficient conditions for \( z \) being a local minima are equivalent to: there exists a constant \( M_z > 0 \) such that \( G(v, t) \geq 0 \) for all \( v \in T_z\mathcal{M} \cap \mathbb{C}^{n-1} \) and all \( t \) with \( |t| \geq M_z \).

Similarly, we can derive the fourth-order global optimality conditions.

**Theorem 2.3** (Characterization of global optimality) *Let \( z \) be a stationary point of problem (1.1). Then \( z \) is a global solution if and only if

\[ H_f(z)[v] \geq 0, \quad 2H_3(v)^2 \leq H_f(z)[v] \cdot [f(v) - f(z)], \quad \forall \, v \in T_z\mathcal{M} \cap \mathbb{C}^{n-1}, \]

\[ H_3(v) = 0, \quad H_4(v) \geq 0, \quad \forall \, v \in T_z\mathcal{M} \cap \mathbb{C}^{n-1} \text{ with } H_f(z)[v] = 0. \]
Proof A stationary point $\mathbf{z}$ is a global minimum if and only if $f(\mathbf{y}) \geq f(\mathbf{z})$ for all $\mathbf{y} \in \mathbb{C}S^{n-1}$, which is equivalent to $G(\mathbf{v}, t) \geq 0$ for all $t \in \mathbb{R}$ and all $\mathbf{v} \in T_{\mathcal{M}} \cap \mathbb{C}S^{n-1}$. However, nonnegativity of the (degenerated) quadratic function $t \mapsto G(\mathbf{v}, t) = H_f(\mathbf{z})[\mathbf{v}]^2 + 4H_3(\mathbf{v})t + 2[f(\mathbf{v}) - f(\mathbf{z})]$ on $\mathbb{R}$ is equivalent to the conditions stated in Theorem 2.3. \qed

Finally, we establish fourth-order necessary conditions for local optimality.

**Theorem 2.4** (Fourth-order necessary optimality conditions) Let $\mathbf{z}$ be a local minima of problem (1.1). Then it holds that

$$
\begin{aligned}
H_f(\mathbf{z})[\mathbf{v}] &\geq 0, \quad \forall \mathbf{v} \in T_{\mathcal{M}} \cap \mathbb{C}S^{n-1}, \\
H_3(\mathbf{v}) &= 0, \quad H_4(\mathbf{v}) \geq 0, \quad \forall \mathbf{v} \in T_{\mathcal{M}} \cap \mathbb{C}S^{n-1} \text{ with } H_f(\mathbf{z})[\mathbf{v}] = 0, \\
4H_3^2(\mathbf{v}, \mathbf{w}) &\leq H_f(\mathbf{z})[\mathbf{w}]H_4(\mathbf{v}), \\
\forall \mathbf{v}, \mathbf{w} &\in T_{\mathcal{M}} \cap \mathbb{C}S^{n-1} \text{ with } H_f(\mathbf{z})[\mathbf{v}] = 0, \quad H_f(\mathbf{z})[\mathbf{w}] > 0, \quad \mathbf{v}^*\mathbf{w} = 0,
\end{aligned}
$$

where $H_3(\mathbf{v}, \mathbf{w}) = \beta \sum_{k \in \mathcal{N}} [|v_k|^2 - |z_k|^2]|\Re(\bar{z}_k w_k)| + 2|\Re(\bar{v}_k w_k)|\Re(\bar{v}_k z_k)$.

Proof Theorem 2.2 implies that there is $M_\mathbf{z} > 0$ such that $G(\mathbf{v}, t) \geq 0$ for all $\mathbf{v} \in T_{\mathcal{M}} \cap \mathbb{C}S^{n-1}$ and $|t| \geq M_\mathbf{z}$. Hence, for fixed $\mathbf{v}$, we have $G(\mathbf{v}, t) = H_f(\mathbf{z})[\mathbf{v}]^2 + 4H_3(\mathbf{v})t + 2[f(\mathbf{v}) - f(\mathbf{z})] \geq 0$ for $t \to \pm \infty$. Thus, it follows $H_f(\mathbf{z})[\mathbf{v}] \geq 0$ and if $H_f(\mathbf{z})[\mathbf{v}] = 0$, it must hold $H_3(\mathbf{v}) = 0$ and $2[f(\mathbf{v}) - f(\mathbf{z})] = H_f(\mathbf{z})[\mathbf{v}] + H_4(\mathbf{v}) = H_4(\mathbf{v}) \geq 0$ by equation (2.8). Now we prove the last condition. Suppose $\mathbf{v}$ and $\mathbf{w}$ are two vectors in $T_{\mathcal{M}} \cap \mathbb{C}S^{n-1}$ satisfying $H_f(\mathbf{z})[\mathbf{v}] = 0$, $H_f(\mathbf{z})[\mathbf{w}] > 0$ and $\mathbf{v}^*\mathbf{w} = 0$. Let us set $\mathbf{y} = \cos(\theta)\mathbf{v} + \sin(\theta)\mathbf{w} \in T_{\mathcal{M}} \cap \mathbb{C}S^{n-1}$, where $\theta \in [-\pi, \pi] \setminus \{0\}$. Using the definition of $H_f(\mathbf{z})[\mathbf{y}]$ in (2.4), we obtain

$$
H_f(\mathbf{z})[\mathbf{y}] = \cos^2(\theta) \cdot H_f(\mathbf{z})[\mathbf{v}] + 2\sin(\theta) \cos(\theta)\Re\left(\left[\begin{array}{c} \mathbf{v} \\ \mathbf{w} \end{array}\right]^* \text{Hess } \mathbf{f} \left[\begin{array}{c} \mathbf{z} \\ \mathbf{w} \end{array}\right]\right)
+ \sin^2(\theta) \cdot H_f(\mathbf{z})[\mathbf{w}]
$$

for all $\theta \in [-\pi, \pi]$. Consequently, due to $H_f(\mathbf{z})[\mathbf{v}] = 0$ and $H_f(\mathbf{z})[\mathbf{y}] \geq 0$ (for all $\theta \in [-\pi, \pi]$), we can infer

$$
H_f(\mathbf{z})[\mathbf{y}] = \sin^2(\theta) \cdot H_f(\mathbf{z})[\mathbf{w}]. \quad (2.11)
$$

We now consider the limiting process $\theta \to 0$ and discuss corresponding representations of $H_3(\mathbf{y})$ and $H_4(\mathbf{y})$. Applying $\cos^2(\theta) = 1 - \sin^2(\theta)$, it holds that

$$
\begin{aligned}
H_3(\mathbf{y}) &= \beta \sum_k [\cos^2(\theta)|v_k|^2 + 2\cos(\theta)\sin(\theta)\Re(\bar{v}_k w_k) \\
&\quad + \sin^2(\theta)|w_k|^2 - |z_k|^2] \cdot \Re(\bar{z}_k (\cos(\theta)v_k + \sin(\theta)w_k))
\end{aligned}
$$

\[\square\] Springer
\[ = \beta \sum_k [|v_k|^2 - |z_k|^2 + 2 \sin(\theta) \cos(\theta) \Re(\bar{v}_k w_k)] + \sin^2(\theta) (|w_k|^2 - |v_k|^2) \cdot \Re(\bar{z}_k (\cos(\theta) v_k + \sin(\theta) w_k)) \]
\[ = \beta \sum_k [\sin(\theta) (|v_k|^2 - |z_k|^2) \Re(\bar{z}_k w_k) + 2 \sin(\theta) \cos^2(\theta) \cdot \Re(\bar{v}_k w_k) \Re(\bar{z}_k v_k)] + \cos(\theta) \cdot H_3(v) + O(\sin^2(\theta)) \]
\[ = \sin(\theta) \cdot H_3(v, w) + O(\sin^2(\theta)) \]  
\[
H_4(y) = \beta \sum_k [\cos^2(\theta)|v_k|^2 + \sin(2\theta) \Re(\bar{v}_k w_k) + \sin^2(\theta)|w_k|^2 - |z_k|^2]^2 - 4\beta \sum_k \Re(\cos(\theta)\bar{z}_k v_k + \sin(\theta)\bar{z}_k w_k)^2 \]
\[ = \beta \sum_k [|v_k|^2 - |z_k|^2 + O(\sin(\theta))]^2 - 4\beta \sum_k [\Re(\bar{z}_k v_k)^2 + O(\sin(\theta))] \]
\[ = H_4(v) + O(\sin(\theta)). \]  

We first discuss the case \( H_4(v) = 0 \). Using (2.9), the discriminant of \( G(y, t) \)—as a quadratic function of \( t \)—is given by

\[ 16H_3^2(y) - 4H_f(z)[y](H_f(z)[y] + H_4(y)) = 16H_3^2(v, w) \sin^2(\theta) + O(\sin^3(\theta)). \]

If \( H_3(v, w) \neq 0 \), this term is positive for all sufficiently small \( \theta \) and hence, \( G(y, t) \) has two real roots. The larger absolute value of the roots is

\[ \frac{4|H_3(y)| + \sqrt{16H_3^2(y) - 4H_f(z)[y](H_f(z)[y] + H_4(y))}}{2H_f(z)[y]} = \Theta\left(\left|\sin(\theta)\right|^{-1}\right), \]

which implies that there does not exist a constant \( M_3 > 0 \) such that \( G(y, t) \geq 0 \) for all \( y \) and \( t \) with \( y \in \mathcal{T}_z \mathcal{M} \cap \mathbb{C}^{n-1} \) and \( |t| \geq M_3 \). Thus, we have \( H_3(v, w) = 0 \) in this case. Next, we consider the case \( H_4(v) > 0 \) and let us suppose \( 4H_3^2(v, w) > H_f(z)[y]H_4(v) = \Theta(\sin^2(\theta)) \). The discriminant of \( G(y, t) \) now satisfies

\[ 16H_3^2(y) - 4H_f(z)[y](H_f(z)[y] + H_4(y)) = 4\left[4H_3^2(v, w) - H_f(z)[w]H_4(v)\right] \sin^2(\theta) + O(\sin^3(\theta)) > 0, \]

for all sufficiently small \( \theta \neq 0 \). As in the last case, the absolute value of the larger root of \( t \mapsto G(y, t) \) converges to \( +\infty \) as \( \theta \to 0 \) which yields the same contradiction.
Consequently, we have $4H_3^2(v, w) \leq H_f(z)[w]H_4(v)$ by combining the two cases. Since these implications do not depend on $v$ and $w$, this concludes the proof.

The fourth-order optimality conditions in Theorems 2.3 and 2.4 resemble other known fourth-order conditions, see, e.g., [23, 28, 64], and might be hard to verify in practice. However, in the real case, the inequality $H_4(v) \geq 0$ is equivalent to checking $\|v\|^2 - 3|z|^2 \geq 2\sqrt{2}\|z\|^2_2$ since it holds that $H_4(v) = \beta\|v\|^2 - 3|z|^2 - 8\beta|z|^2_2$. In this situation, the framework presented in [4] can be used to verify the first two conditions in Theorem 2.4 in polynomial time.

**Theorem 2.5** (Fourth-order sufficient optimality conditions) Let $z \in \mathbb{C}S^{n-1}$ be a given point. Then, the fourth-order optimality conditions

\[
\begin{align*}
\text{z is a stationary point,} \\
H_f(z)[v] \geq 0 & \quad \forall v \in T_z\mathcal{M} \cap \mathbb{C}S^{n-1}, \\
H_3(v) = 0, & \quad H_4(v) > 0 \quad \forall v \in T_z\mathcal{M} \cap \mathbb{C}S^{n-1} \text{ with } H_f(z)[v] = 0,
\end{align*}
\]

are equivalent to the following fourth-order growth condition

\[
\exists \alpha, \delta_z > 0 \quad f(y) - f(z) \geq \alpha\|y - z\|^4 \quad \forall y \in B(z, \delta_z) \cap \mathbb{C}S^{n-1},
\]

(2.14)

where $H_3(v, w)$ is defined as in Theorem 2.4. In addition, if one of the latter conditions is satisfied, $z$ is a strict local minimum of (1.1)

**Proof** Let us first assume that the growth condition (2.14) is satisfied. In this case, $z$ is a local minimum and thus, the first-order optimality conditions (2.3) hold at $z$. As shown in the proof of Theorem 2.2, the condition $y = \cos(\theta)z + \sin(\theta)v$, $v \in T_z\mathcal{M} \cap \mathbb{C}S^{n-1}$ is equivalent to $y = \cos(\theta)z + \sin(\theta)v$, $v \in T_z\mathcal{M} \cap \mathbb{C}S^{n-1}$, and $\cos(\theta) \geq 1 - \frac{\delta_z^2}{2}$. Consequently, due to (2.9) and $\|y - z\|^2 = 2 - 2\cos(\theta)$, the growth condition (2.14) is equivalent to

\[
\sin(\theta)^4 \cdot G(v, \cot(\theta)) \geq 8\alpha(1 - \cos(\theta))^2 \quad \forall v \in T_z\mathcal{M} \cap \mathbb{C}S^{n-1}
\]

(2.15)

and for all $\theta$ with $\cos(\theta) \geq 1 - \frac{\delta_z^2}{2}$. Specifically, we obtain

\[
G(v, \cot(\theta)) \geq \frac{8\alpha}{(1 + \cos(\theta))^2} \geq 2\alpha \quad \forall v \in T_z\mathcal{M} \cap \mathbb{C}S^{n-1}, \quad \forall \theta \in \mathcal{I}_\delta,
\]

where $\mathcal{I}_\delta := \{\theta : \cos(\theta) \geq 1 - \frac{\delta_z^2}{2} \cap (-\pi, \pi)\} \{0\}$. Following the proof of Theorem 2.4 and by discussing the limits $\theta \downarrow 0$ and $\theta \uparrow 0$, we can infer $H_f(z)[v] \geq 0$ for all $v \in T_z\mathcal{M} \cap \mathbb{C}S^{n-1}$ and

\[
H_3(v) = 0, \quad H_4(v) > 0 \quad \forall v \in T_z\mathcal{M} \cap \mathbb{C}S^{n-1} \text{ with } H_f(z)[v] = 0.
\]

In order to verify the final statement in the fourth-order sufficient condition, we mimic the proof of Theorem 2.4. In particular, we can introduce the finer decomposition

\[
\mathbb{C}S^{n-1} = \mathbb{C}S^{\delta_z^2/2} \cup \mathbb{C}S^{\delta_z^2/2}
\]

 Springer
\( \mathbf{v} = \cos(\phi)\mathbf{u} + \sin(\phi)\mathbf{w} \) with \( \mathbf{u}, \mathbf{w} \in T_x \mathcal{M} \cap \mathbb{C}S^{n-1}, H_f(\mathbf{z})|_{\mathbf{u}} = 0, H_f(\mathbf{z})|_{\mathbf{w}} > 0, \) and \( \mathbf{u}^*\mathbf{w} = 0. \) Utilizing (2.11)–(2.13), the discriminant of the mapping \( t \mapsto G(\mathbf{v}, t) - 2\alpha \) is then given by

\[
16H_3^2(\mathbf{v}) - 4H_f(\mathbf{z})|_{\mathbf{v}}(H_f(\mathbf{z})|_{\mathbf{v}} + H_4(\mathbf{v}) - 2\alpha) \\
= 4\left[4H_3^2(\mathbf{u}, \mathbf{w}) - H_f(\mathbf{z})|_{\mathbf{w}}(H_4(\mathbf{u}) - 2\alpha)\right]\sin^2(\phi) + O(\sin^3(\phi)),
\]

for \( \phi \to 0. \) As in the proof of Theorem 2.4, if \( H_3^2(\mathbf{u}, \mathbf{v}) \geq H_f(\mathbf{z})|_{\mathbf{w}}H_4(\mathbf{u}), \) this implies that the larger root of \( t \mapsto G(\mathbf{v}, t) - 2\alpha \) converges to \( +\infty \) as \( \phi \to 0 \) and hence, the condition \( G(\mathbf{v}, \cot(\theta)) \geq 2\alpha \) can not hold for all \( \mathbf{v} \in T_x \mathcal{M} \cap \mathbb{C}S^{n-1} \) and \( \theta \in \mathcal{T}_{\delta}. \)

Conversely, let us assume that the fourth-order optimality conditions are satisfied. We define the sets \( \mathcal{N} := \{ \mathbf{v} \in \mathbb{C}^n : H_f(\mathbf{z})|_{\mathbf{v}} = 0 \}, \ \mathcal{U} := [T_x \mathcal{M} \cap \mathbb{C}S^{n-1}] \cap \mathcal{N}, \) and \( \mathcal{W} := [T_x \mathcal{M} \cap \mathbb{C}S^{n-1}] \cap \mathcal{N}^\perp. \) Since both \( \mathcal{U} \) and \( \mathcal{W} \) are compact and the functions \( H_f(\mathbf{z})|_{\cdot}, H_3(\cdot, \cdot), \) and \( H_4(\cdot) \) are continuous, we have

\[
\min_{\mathbf{u} \in \mathcal{U}} H_4(\mathbf{u}) := \epsilon_1 > 0, \quad \min_{(\mathbf{u}, \mathbf{w}) \in \mathcal{U} \times \mathcal{V}} H_f(\mathbf{z})|_{\mathbf{w}}H_4(\mathbf{u}) - 4H_3^2(\mathbf{u}, \mathbf{w}) := \epsilon_2 > 0,
\]

and \( \min_{\mathbf{w} \in \mathcal{W}} H_f(\mathbf{z})|_{\mathbf{w}} = \epsilon_3 > 0. \) Let \( \mathbf{v} \in T_x \mathcal{M} \cap \mathbb{C}S^{n-1} \) be arbitrary and let \( \xi > 0 \) be a small constant that will be determined later. Setting \( \epsilon := \frac{1}{2} \min\{\epsilon_1, \epsilon_2\}, \) we now consider three different cases.

Case 1. \( H_f(\mathbf{z})|_{\mathbf{v}} = 0. \) In this case, it follows \( G(\mathbf{v}, t) = H_4(\mathbf{v}) > \epsilon \) for all \( t \in \mathbb{R}. \)

Case 2. \( H_f(\mathbf{z})|_{\mathbf{v}} > \xi. \) The roots of the quadratic function \( t \mapsto G(\mathbf{v}, t) - \epsilon \) are bounded by

\[
\frac{4|H_3(\mathbf{v})| + \sqrt{\left|16H_3^2(\mathbf{v}) - 4H_f(\mathbf{z})|_{\mathbf{v}}(H_f(\mathbf{z})|_{\mathbf{v}} + H_4(\mathbf{v}) - \epsilon)\right|}}{2H_f(\mathbf{z})|_{\mathbf{v}}} < \frac{2 + \sqrt{7}}{\xi}M,
\]

where \( M := \max_{\mathbf{v} \in T_x \mathcal{M} \cap \mathbb{C}S^{n-1}} \max \left\{ |H_f(\mathbf{z})|_{\mathbf{v}}, |H_3(\mathbf{v})|, |H_4(\mathbf{v})|, \epsilon \right\}. \) The continuity of the functions \( H_f(\mathbf{z})|_{\cdot}, |H_3(\cdot)|, \) and \( |H_4(\cdot)| \) implies \( M < +\infty. \) Hence, we have \( G(\mathbf{v}, t) > \epsilon \) for all \( t \in \mathbb{R} \) such that \( |t| \geq (2 + \sqrt{7})M\xi^{-1}. \)

Case 3. \( H_f(\mathbf{z})|_{\mathbf{v}} \in (0, \xi]. \) As before, we can further decompose \( \mathbf{v} \) via \( \mathbf{v} = \cos(\phi)\mathbf{u} + \sin(\phi)\mathbf{w}, \ \mathbf{u} \in \mathcal{U}, \ \mathbf{w} \in \mathcal{W}, \ \mathbf{u}^*\mathbf{w} = 0, \) and \( \phi \in [-\pi, \pi]. \) (For instance, the vectors \( \mathbf{u} \) and \( \mathbf{w} \) can be set as normalized orthogonal projections of \( \mathbf{v} \) onto the subspaces \( T_x \mathcal{M} \cap \mathcal{N} \) and \( T_x \mathcal{M} \cap \mathcal{N}^\perp, \) respectively). Using (2.11) and the underlying assumption of case 3, we have

\[
\sin^2(\phi) = \left(H_f(\mathbf{z})|_{\mathbf{w}}\right)^{-1} \cdot H_f(\mathbf{z})|_{\mathbf{v}} \leq \epsilon_3^{-1}\xi.
\]

Furthermore, by (2.11)–(2.13), the discriminant of \( t \mapsto G(\mathbf{v}, t) - M^{-1}\epsilon \) satisfies
where $C_z > 0$ is a constant only depending on $z$. Notice that such a uniform constant can be found by using the compactness of $\mathbb{CS}^{n-1}$ and a continuity argument in (2.12) and (2.13). Choosing $0 < \xi < (4\epsilon/C_z)^2\epsilon_3$ and due to (2.16), we obtain $|\sin(\phi)| < 4\epsilon/C_z$, which implies that the discriminant of $t \mapsto G(v, t) - M^{-1}\epsilon$ is negative. This yields $G(v, t) \geq M^{-1}\epsilon$ for all $t \in \mathbb{R}$.

Overall, setting $M_2 := (2 + \sqrt{7})M_2\xi^{-1}$ and $\alpha := \min\{\epsilon/8, (8M)^{-1}\epsilon\}$, it follows $G(v, t) \geq 8\alpha$ for all $v \in T_{zA}M \cap \mathbb{CS}^{n-1}$ and $t \in \mathbb{R}$ such that $|t| \geq M_2$. Due to (2.10), $\delta_z \in (0, 1]$ can be chosen such that we have $|\cot(\theta)| \geq M_2$ for all $\theta$ with $\cos(\theta) \geq 1 - \delta_z^2/2$. For all of those $\theta$ and all $v \in T_{zA}M \cap \mathbb{CS}^{n-1}$, we then can infer

$$G(v, \cot(\theta)) \geq 8\alpha \geq \frac{8\alpha}{(1 + \cos(\theta))^2} = 8\alpha \cdot \frac{(1 - \cos(\theta))^2}{\sin^2(\theta)},$$

where the second inequality follows from $\cos(\theta) \geq 0$. By (2.15), this shows that $z$ is a strict local minimum. \hfill $\square$

### 3 Geometric analysis of the diagonal case

In this section, we investigate the geometric properties of problem (1.1) under the assumption that $A$ is a diagonal matrix, i.e., $A = \text{diag}(a) = \text{diag}(a_1, a_2, \ldots, a_n)$.

By setting $u_k = |z_k|^2$, we can reformulate problem (1.1) as a convex problem

$$\min_{u \in \mathbb{R}^n} \frac{1}{2} a^T u + \frac{\beta}{2} \|u\|^2 \quad \text{s.t.} \quad u \in \Delta_n,$$

where $\Delta_n = \{u \in \mathbb{R}^n : u_k \in [0, 1], k \in [n], \sum_{k \in [n]} u_k = 1\}$ is the $n$-simplex. We will use this connection later to show that there are no spurious local minima in the diagonal case and that the global solutions can be characterized via the unique solution of the strongly convex problem (3.1).

We first derive an explicit representation of critical points of problem (1.1).

**Lemma 3.1** (Characterizing stationary points) Suppose that $A$ is diagonal and let $z \in \mathbb{CS}^{n-1}$ be given. Let us set $\mathcal{I} := \{k \in [n] : z_k \neq 0\}$ and

$$u_k := 0, \quad \forall \ k \in \mathcal{C}, \quad u_k := \frac{1}{|\mathcal{I}|} + \frac{1}{2\beta} \left[ \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} a_i - a_k \right], \quad \forall \ k \in \mathcal{I}.$$

Then, $z$ is a stationary point if and only if there exist $\theta_k \in [0, 2\pi), k \in [n]$, such that $u_k \in (0, 1]$ for all $k \in \mathcal{I}$ and $z_k = \sqrt{u_k}e^{i\theta_k}$ for all $k$. \hfill $\triangle$
Proof In the diagonal case, introducing the polar form $z = (r_1e^{i\theta_1}, \ldots, r_ne^{i\theta_n})^T$, the first-order optimality conditions reduce to
\[
(a_k + 2\beta r_k^2 - 2\lambda)r_k e^{i\theta_k} = 0 \quad \forall \; k \in [n].
\]
(3.2)
where $\lambda$ is the associated multiplier defined in (2.2). Specifically, for all $k \in I$, we have $a_k + 2\beta r_k^2 = 2\lambda$ and summing these equations, we obtain
\[
\lambda = \frac{2\beta + \sum_{k \in I} a_k}{2|I|}, \quad r_k^2 = \frac{2\lambda - a_k}{2\beta},
\]
and $2\lambda - a_k \in (0, 2\beta]$ for all $k \in I$. The claimed result in Lemma 3.1 now follows immediately by setting $u_k = r_k^2$, $k \in [n]$. \qed

Next, we discuss the local minimizer of problem (1.1). By combining Theorems 2.1 and 3.1, we see that there are no spurious local minimizer in the diagonal case, i.e., all local solutions are automatically global solutions of problem (1.1).

Theorem 3.1 (Characterization of local minimizer) Let $A$ be a diagonal matrix. A point $z \in \mathbb{C}^n$ is a local minimizer of problem (1.1) if and only if
\[
\text{grad } f(z) = 0 \quad \text{and} \quad H = A + 2\beta \text{diag}(|z|^2) - 2\lambda I \succeq 0,
\]
(3.3)
where $\lambda \in \mathbb{R}$ is the associated Lagrange multiplier. In addition, every local solution $z$ can be represented explicitly and has to satisfy
\[
z_k = \sqrt{u_k} e^{i\theta_k}, \quad \theta_k \in [0, 2\pi), \quad u = \mathcal{P}_{\Delta_n}(-a/2\beta), \quad \forall \; k \in [n],
\]
where $\mathcal{P}_{\Delta_n}$ denotes the Euclidean projection onto the n-simplex $\Delta_n$.

Proof According to Theorem 2.1, a point satisfying the conditions (3.3) is a global minimum of problem (1.1) and hence, it is also a local minimum. Let $z \in \mathbb{C}^{n-1}$ now be an arbitrary local minimum. Then, the first- and second-order necessary optimality conditions hold at $z$, i.e., we have grad $f(z) = 0$ and
\[
H_f(z)[v] = v^* H v + 4\beta \cdot \sum_{k=1}^n r_k^2 t_k^2 \cos^2(\theta_k - \phi_k) \geq 0
\]
(3.4)
for all $v \in T_z M$, where $(r_1e^{i\theta_1}, \ldots, r_ne^{i\theta_n})^T$ and $(t_1e^{i\phi_1}, \ldots, t_ne^{i\phi_n})^T$ are the corresponding polar coordinates of $z$ and $v$, respectively.

As shown in Lemma 3.1 and using the stationarity condition grad $f(z) = 0$, it follows $H_{kk} = a_k + 2\beta |z_k|^2 - 2\lambda = 0$ for all $k \in I = \{k : z_k \neq 0\}$. Next, for $k \in I^C$, we define $v := e_k$, where $e_k$ denotes the k-th unit vector. This choice of $v$ obviously fulfills $\Re(v^*z) = 0$ and thus, the optimality condition (3.4) implies $H_{kk} = H_f(z)[v] \geq 0$. Since $H$ is diagonal, this yields $H \succeq 0$. \qed

Springer
In order to verify the explicit characterization of local minimizers, we notice that \( \mathbf{u} = P_{\Delta_n}(-\mathbf{a}/2\beta) \) is the unique solution of the strongly convex problem (3.1). Moreover, using the identity \( u_k \equiv |z_k|^2, k \in [n] \), every global solution of (1.1) corresponds to a global minimizer of the problem (3.1) and vice versa. Since problem (3.1) does not possess spurious local minimizers, this finishes the proof of Theorem 3.1. \( \square \)

The latter theorem shows that we can identify and explicitly compute the unique equivalence class \([z]\) of global minimizer by a projection onto the \(n\)-simplex. This can be realized numerically in \(O(n \log n)\) operations, see, e.g., [66].

Inspired by the analysis of phase synchronization problems in [9], we now study the behavior of global minimizer when the diagonal matrix \(A\) is perturbed by a random noise matrix \(W\).

**Theorem 3.2** Let \(A\) be a diagonal matrix and let \(W \in \mathbb{C}^{n \times n}\) be a Hermitian matrix. Suppose that \(z_0\) is a global minimizer of (1.1) and that the point \(y \in \mathbb{C}^{n-1}\) satisfies \(f_\sigma(y) \leq \min_{z \in [z_0]} f_\sigma(z)\), where \(f_\sigma(z) := f(z) + \frac{\sigma}{2} z^* W z\) and \(\sigma > 0\) is a given noise parameter. Then, it holds that

\[
\min_{z \in [z_0]} \|y - z\|_4 \leq \sqrt{2\sigma \beta^{-1} \|W\|_2 n^{1/4}}.
\]

**Proof** As usual, we introduce the polar coordinates \(z_0 = (r_1 e^{i\theta_1}, ..., r_n e^{i\phi_n})^T\) and \(y = (t_1 e^{i\phi_1}, ..., t_n e^{i\phi_n})^T\). Due to Theorem 3.1, we can assume \(\theta_k = \phi_k\) and we have \(a_k + 2\beta r_k^2 - 2\lambda \geq 0\) for all \(k \in [n]\), where \(a_k = A_{kk}\) and \(\lambda\) is the associated multiplier of \(z_0\). Moreover, applying (3.2), we can infer \(a_k + 2\beta r_k^2 - 2\lambda = 0\) for all \(k\) with \(r_k > 0\). Thus, using \(f_\sigma(y) \leq f_\sigma(z_0)\) and \(\|z_0\| = \|y\| = 1\), this implies

\[
\sigma (z_0^* W z_0 - y^* W y) \\
\geq (y^* A y - z_0^* A z_0) + \beta \sum_{k \in [n]} (t_k^4 - r_k^4) \\
= \sum_{k \in [n]} (a_k + \beta (t_k^2 + r_k^2))(t_k^2 - r_k^2) \\
= \sum_{k \in [n], r_k > 0} (a_k + \beta (t_k^2 + r_k^2))(t_k^2 - r_k^2) + \sum_{k \in [n], r_k = 0} (a_k + \beta t_k^2) t_k^2 \\
\geq \sum_{k \in [n], r_k > 0} (2\lambda + \beta (t_k^2 - r_k^2))(t_k^2 - r_k^2) + \sum_{k \in [n], r_k = 0} (2\lambda + \beta t_k^2) t_k^2 \\
= \sum_{k \in [n]} (2\lambda + \beta (t_k^2 - r_k^2))(t_k^2 - r_k^2) \\
= \beta \sum_{k \in [n]} (t_k^2 - r_k^2)^2 \geq \beta \sum_{k \in [n]} (t_k - r_k)^4 = \beta \|y - z_0\|_4^4.
\]
Note that the last inequality follows from \((tk + rk)^2 \geq (tk - rk)^2\). Furthermore, by Hölder’s inequality and by \(\|x\|_{4/3} \leq n^{1/4}\|x\|\), we have

\[
\begin{align*}
  z_0^* Wz_0 - y^* W y &= \Re((z_0 - y)^* W(z_0 + y)) \\
  &\leq \|z_0 - y\|_4\|W(z_0 + y)\|_{4/3} \\
  &\leq n^{1/4}\|z_0 - y\|_4\|W(z_0 + y)\| \\
  &\leq 2n^{1/4}\|W\|_2\|z_0 - y\|_4.
\end{align*}
\]

Combining the last inequalities, we can conclude the proof. \(\Box\)

**Remark 1** If \(W \in \mathbb{C}^{n \times n}\) is a Hermitian random matrix with i.i.d. off-diagonal entries following a standard complex normal distribution and with zero diagonal entries, then Bandeira et al. [9], have shown that the bound \(\|W\|_2 \leq 3\sqrt{n}\) holds with probability at least \(1 - \frac{2}{n} - \frac{5}{4} - e^{-n/2}\). Combining this observation with Theorem 3.2, we can obtain

\[
\min_{z \in [z_0]} \|y - z\|_4 \leq 3\sqrt[3]{6}\sigma_\beta^{-1} \cdot n^{1/4}
\]

with probability at least \(1 - 2n^{-5/4} - e^{-n/2}\).

### 4 Geometric analysis of the rank-one case

In this section, we investigate the case when \(A\) is rank-one and positive semidefinite, i.e., we can write \(A = aa^*\) for some \(a \in \mathbb{C}^n\) and the quartic–quadratic problem (1.1) reduces to

\[
\min_{z \in \mathbb{C}^n} f(z) = \frac{1}{2}|a^* z|^2 + \frac{\beta}{2}\|z\|_4^4 \quad \text{s.t. } \|z\|_2 = 1. \tag{4.1}
\]

The associated first- and second-order necessary optimality conditions are given by

\[
a^* z \cdot a + 2\beta\text{diag}(|z|^2)z = 2\lambda z, \quad 2\lambda = |a^* z|^2 + 2\beta\|z\|_4^4 \tag{4.2}
\]

and \(H_f(z)[v] = v^*[aa^* + 2\beta\text{diag}(|z|^2) - 2\lambda I]\) for all \(v \in \mathbb{C}^n\). We now present a first structural and preparatory property of local and global minima.

**Lemma 4.1** Suppose that \(z\) is a local minimizer of (4.1). Then, for all \(k \in [n]\) with \(a_k = 0\) it holds that \(|z_k|^2 = \frac{\lambda}{\beta}\).

**Proof** If \(a_k = 0\), the first-order optimality conditions (4.2) imply \(\beta|z_k|^2 z_k = \lambda z_k\). Let us assume \(z_k = 0\) and let us choose \(v \in \mathbb{C}^n\) with \(v_k = 1\) and \(v_j = 0\) for all \(j \neq k\). Due to \(\lambda \geq \beta\|z\|_4^4 \geq \frac{\beta}{n} > 0\), we obtain

\[
H_f(z)[v] = -2\lambda < 0,
\]

which contradicts the second-order necessary optimality conditions. Hence, we have \(|z_k|^2 = \frac{\lambda}{\beta}\). \(\square\)
In the following sections, we discuss two different classes of local minima, which are characterized by their orthogonality to the vector $a$.

### 4.1 Orthogonal local minima

We first analyze the case where the local minimizer $z$ satisfies $a^*z = 0$.

**Theorem 4.1** Suppose that $z$ is a local minimizer satisfying $a^*z = 0$. Then, $z$ has at most one zero component and all of its nonzero components must have the same modulus.

**Proof** By the first-order optimality conditions, it follows $z_k = 0$ or $|z_k|^2 = \frac{1}{n}$ for all $k$. Hence, all nonzero components of $z$ have the same modulus.

Without loss of generality we now assume that $|z_k|^2 = \frac{1}{n}$ for $1 \leq k \leq n$ and $z_k = 0$ for $\tau + 1 \leq k \leq n$. Due to Lemma 4.1, we have $a_{n-1}, a_n \neq 0$ if $\tau \leq n - 2$. Let us set

$$v_k = 0, \quad k \in [n-2], \quad v_{n-1} = \frac{\tilde{a}_n}{\sqrt{|a_{n-1}|^2 + |a_n|^2}}, \quad v_n = \frac{-\tilde{a}_{n-1}}{\sqrt{|a_{n-1}|^2 + |a_n|^2}}.$$

Then, it holds that $H_f(z|v) = -2\lambda < 0$, which is a contradiction. Thus, we have $\tau = n - 1$ or $\tau = n$ (which means that all components $z_k$ are nonzero).

Next, we derive conditions under which the existence of such local minima can be ensured. Before we present the formal statement and proof of the main theorem, we discuss a result that is used later in Theorem 4.2.

**Lemma 4.2** If $\|a\|_\infty \leq \frac{1}{2} \|a\|_1$, there exist phases $\{\theta_k\}_{k \in [n]}$, $\theta_k \in [0, 2\pi]$, such that $\sum_{k \in [n]} e^{i\theta_k}a_k = 0$.

**Proof** Let us assume $a_k \neq 0$ for all $k$. If $n = 1$, the condition $\|a\|_\infty \leq \frac{1}{2} \|a\|_1$ is never satisfied and hence, the statement in Lemma 4.2 holds automatically. In the case $n = 2$, we have $2 \max\{|a_1|, |a_2|\} \leq |a_1| + |a_2|$. This implies $|a_1| = |a_2|$ and thus, we can choose $\theta_1$ and $\theta_2$ such that $e^{i\theta_1}a_1 = |a_1|$ and $e^{i\theta_2}a_2 = -|a_1|$.

Otherwise assume $n \geq 3$ and $|a_1| = \|a\|_\infty$. Let $b_1 = \sum_{k \geq 3} |a_k|$, $b_2 = |a_2|$ and $b_3 = |a_1|$. It holds that $b_1 + b_2 = \|a\|_1 - |a_1| \geq b_3$, which means that the numbers $b_1, b_2, b_3$ can be interpreted as sides of a (degenerated) triangle. Let $ABC$ be such a triangle embedded into the complex space, where $A, B, C \in \mathbb{C}$ denote the nodes of $ABC$ with $|B - C| = b_1, |C - A| = b_2, |A - B| = b_3$. Consequently, there exist $\phi_1, \phi_2, \phi_3 \in [0, 2\pi)$ such that $B - C = e^{i\phi_1}b_1, C - A = e^{i\phi_2}b_2, A - B = e^{i\phi_3}b_3$.

But then we have $\sum_{k=1}^3 e^{i\phi_k}b_k = 0$, which completes the proof of the lemma.

**Theorem 4.2** (Existence of orthogonal local minima) There exists a local minimizer $z$ of (4.1) such that $a^*z = 0$ if and only if $\|a\|_\infty \leq \frac{1}{2} \|a\|_1$, or $a$ has only one nonzero component and we have $\|a\|^2 \geq 2\beta/(n-1)$. Further, if $a$ satisfies such conditions, all local minima with $a^*z = 0$ are the only global minima of (4.1).
Let \( z = (r_1 e^{i \theta_1}, \ldots, r_n e^{i \theta_n})^T \) be a local minimizer of problem (4.1) such that \( a^*z = 0 \). By Theorem 4.1, we only need to consider the cases when the local minimizer has no zero component or exactly one zero component.

**Case 1.** If \( z \) does not have any zero component, then it follows \( |z_k|^2 = \frac{1}{n} \) for all \( k \) and we have

\[
a^*z = \frac{1}{\sqrt{n}} \sum_{k \in [n]} e^{i \theta_k} a_k = 0,
\]

which implies \( |a_j| = |\sum_{k \neq j} e^{i \theta_k} a_k| \leq \sum_{k \neq j} |a_k| = \|a\|_1 - |a_j| \) for all \( j \in [n] \).

Choosing \( a_j \) to be the element with maximal modulus, we get \( \|a\|_\infty \leq \frac{1}{2} \|a\|_1 \).

**Case 2.** Let us suppose \( z_n = 0 \). Then, due to Lemma 4.1, we obtain \( a_n \neq 0 \). Let us assume that there exists another component \( a_k \neq 0 \) for some \( k \in [n-1] \). Setting

\[
v_j = iz_j, \quad j \neq k, \quad v_k = (1 - |a_n|)iz_k, \quad v_n = \frac{a_n}{|a_n|} \bar{a}_k \cdot iz_k,
\]

and normalizing \( v \), we have \( a^*v = i \cdot a^*z = 0 \) and \( \Re(z^*v) = 0 \) and the quadratic form \( H_f(z)[v] \) satisfies

\[
H_f(z)[v] = 2\beta \sum_{k \in [n-1]} |z_k|^2 |v_k|^2 - 2\lambda \sum_{k \in [n]} |v_k|^2 = -2\lambda |v_n|^2 < 0,
\]

which contradicts the second-order optimality conditions. Note that we used Lemma 4.1 and \( |z_k|^2 = \lambda/\beta \) for \( k \in [n-1] \) in the second equality. Consequently, we can infer \( a_k = 0 \) for all \( k \in [n-1] \), which implies that \( a \) has only one nonzero component. By Theorem 4.1, we then have \( |z_k|^2 = \frac{1}{n-1} \) for all \( k \in [n-1] \). Choosing \( v = (0, 0, \ldots, 0, 1)^T \in T_z \mathcal{M} \) in the second-order necessary optimality conditions, we obtain

\[
H_f(z)[v] = |a_nv_n|^2 + \frac{2\beta}{n-1} (1 - |v_n|^2) - \frac{2\beta}{n-1} = |a_n|^2 - \frac{2\beta}{n-1} \geq 0,
\]

and thus, it follows \( \|a\|^2 = |a_n|^2 \geq \frac{2\beta}{n-1} \).

We continue with the proof of the second direction. In particular, suppose that \( a \) satisfies the conditions stated in Theorem 4.2. We again discuss two cases.

**Case 1.** By Lemma 4.2, if \( \|a\|_\infty \leq \frac{1}{2} \|a\|_1 \), we can choose phases \( \{\theta_k\}_{k \in [n]} \) such that \( \sum_{k \in [n]} e^{i \theta_k} a_k = 0 \). Let us set \( z_k = e^{i \theta_k} / \sqrt{n} \) for all \( k \in [n] \). Then, we have \( a^*z = 0 \) and \( f(z) = \beta/(2n) \), which is the lower bound of the objective function \( f \). Thus, in this case \( z \) is a global minimizer of (4.1). Moreover, the objective function attains its optimal value if and only if \( a^*z = 0 \) and \( \|z\|_2^2 = 1/n \) and hence all global minima are orthogonal minima. Due to Theorem 4.1, we have \( |z_k|^2 = 1/n \) for all orthogonal local minima (those minima can not have zero components since this would imply that \( a \) has only one nonzero component \( a_n \)), which yields \( \|z\|_2^2 = 1/n \). This shows that all orthogonal local minima are the only global solutions.
Case 2. If \( \mathbf{a} \) has only one nonzero component \( a_n \) with \( \| \mathbf{a} \|_2^2 \geq \frac{2\beta}{n-1} \), the matrix \( \mathbf{A} \) is diagonal and we can apply the results derived in Sect. 3. Specifically, by Theorems 3.1 and 4.1, every local minimizer \( \mathbf{z} \) is already a global minimizer of problem (4.1) and it has to satisfy

\[
|z_k| = \frac{1}{\sqrt{n-1}}, \quad \forall \ k \in [n-1], \quad \text{and} \quad z_n = 0.
\]  

(4.3)

Theorem 3.1 further implies that the set of local minima of (4.1) is fully characterized by (4.3). Hence, in this situation, all local minima fulfill the orthogonality condition \( \mathbf{a}^* \mathbf{z} = 0 \) and are the only global minima of (4.1).

This finishes the proof of Theorem 4.2. \( \square \)

4.2 Non-orthogonal local minima

We discuss the case when there is no local minimizer \( \mathbf{z} \) such that \( \mathbf{a}^* \mathbf{z} = 0 \), or, equivalently, \( \mathbf{a} \) does not satisfy the conditions stated in Theorem 4.2. Let us introduce the polar coordinates \( z_k = r_k e^{i\theta_k}, a_k = t_k e^{i\phi_k} \), and \( \mathbf{a}^* \mathbf{z} = s e^{i\zeta} \) for all \( k \in [n] \) and let \( \mathbf{z} \) be a stationary point of (4.1) with \( \mathbf{a}^* \mathbf{z} \neq 0 \). By the optimality condition (4.2), we have

\[
e^{i(\zeta+\phi_k-\theta_k)} = \frac{(2\lambda - 2\beta r_k^2 r_k}{s t_k} \in \mathbb{R} \quad \forall \ k \text{ with } a_k \neq 0,
\]

which implies \( \theta_k - \phi_k - \zeta \in \{\ell \pi : \ell \in \mathbb{Z}\} \) for all those \( k \). Since a global shift of the form \( (z_1, \ldots, z_n) \rightarrow (e^{i\xi_1} z_1, \ldots, e^{i\xi_n} z_n) \) will not change the objective function value and the first-order optimality conditions, if we choose \( \xi_k = \xi \in \mathbb{R} \) for all \( k \) with \( a_k \neq 0 \) and \( \xi_k \in \mathbb{R} \) arbitrarily for all \( k \) with \( a_k = 0 \), we can shift \( \mathbf{z} \) such that \( \theta_k + \xi - \phi_k \in \{\ell \pi : \ell \in \mathbb{Z}\} \) for all \( k \) with \( a_k \neq 0 \). The shifted stationary point then satisfies \( [e^{i\xi_k} z_k] \tilde{a}_k \in \mathbb{R} \). Furthermore, in the case \( a_k = 0 \), we can adjust \( z_k \) to be a real number. Consequently, for every stationary point of problem (4.1), we can find a corresponding stationary point which has the same objective function value and satisfies the following ‘consistency’ property.

Definition 4.1 A stationary point \( \mathbf{z} \) of problem (4.1) is called consistent, if it satisfies

\[
\tilde{a}_k z_k \in \mathbb{R}, \quad \forall \ k \text{ with } a_k \neq 0 \quad \text{and} \quad z_k \in \mathbb{R}, \quad \forall \ k \text{ with } a_k = 0.
\]

Note that the corresponding consistent stationary point of a local minimizer of problem (4.1) does not need to be a local minimizer. On the other hand, a globally shifted consistent stationary point \( (e^{i\xi_1} z_1, \ldots, e^{i\xi_n} z_n) \) of a global minimizer \( \mathbf{z} \) remains a global minimizer since the objective function value does not change. In this subsection, we focus on structural properties of consistent stationary points of (4.1).

Remark 2 Suppose \( \mathbf{a} \) does not satisfy the conditions in Theorem 4.2, then there exists a local minimizer \( \mathbf{z} \) such that \( \mathbf{a}^* \mathbf{z} \neq 0 \). If we have \( z_k = 0 \) for some \( k \in [n] \), then the first-order optimality conditions imply \( a_k = 0 \) which contradicts Lemma 4.1. Hence, for any local minima \( \mathbf{z} \) with \( \mathbf{a}^* \mathbf{z} \neq 0 \), we have \( z_k \neq 0 \) for all \( k \in [n] \).
In the following result, we show that consistent local minima must belong to the same equivalence class defined in (2.5).

**Theorem 4.3** Suppose that \( z, y \in \mathbb{C}S^{n-1} \) are two consistent local minima of problem (4.1) with \( a^*z \neq 0 \) and \( a^*y \neq 0 \). Then, we have \( y \in [z] \).

**Proof** The consistency of the stationary points \( y \) and \( z \) implies \( y^*z \in \mathbb{R} \) and thus, it holds that \( i^*y \in T_z M \cap \mathbb{C}S^{n-1} \) and \( i^*z \in T_y M \cap \mathbb{C}S^{n-1} \). By the second-order necessary optimality conditions, we have

\[
H_f(z)[i^*y] = |a^*y|^2 + 2\beta \sum_{k \in [n]} |z^*yk|^2 - 2\lambda z \geq 0,
\]

\[
H_f(y)[i^*z] = |a^*z|^2 + 2\beta \sum_{k \in [n]} |z^*yk|^2 - 2\lambda y \geq 0,
\]

where \( 2\lambda z = |a^*z|^2 + 2\beta \|z\|_4^4 \) and \( 2\lambda y = |a^*y|^2 + 2\beta \|y\|_4^4 \). Summing those two inequalities yields

\[
0 \leq -2\beta \left[ \|z\|_4^4 - 2 \sum_{k \in [n]} |z^*yk|^2 + \|y\|_4^4 \right] = -2\beta \sum_{k \in [n]} (|z_k|^2 - |y_k|^2)^2.
\]

Hence, we have \( |z_k| = |y_k| \) for all \( k \in [n] \). \( \square \)

**Remark 3** Similarly, if \( z, y \in \mathbb{C}S^{n-1} \) are two local minima of problem (4.1) such that \( z_k \) and \( y_k \) have the same phases for all \( k \in [n] \), then \( y \in [z] \).

We now prove that there are no spurious consistent local minima.

**Theorem 4.4** If \( a \) does not satisfy the conditions in Theorem 4.2, then all the consistent local minima of problem (4.1) are global minima.

**Proof** Suppose \( z, y \in \mathbb{C}S^{n-1} \) are two consistent local minima. Using the inequalities in (4.4) and \( |z_k| = |y_k| \), we obtain \( |a^*z|^2 = |a^*y|^2 \). Hence, all consistent local minima have the same objective function value. Since \( a \) does not satisfy the conditions in Theorem 4.2, there exists a consistent global minimizer satisfying \( a^*z \neq 0 \). This shows that all consistent local minima are global minima. \( \square \)

**Remark 4** Combining the last results, we see that the moduli \( |z_k| \) of every global minimum \( z \) of problem (4.1) coincide and are uniquely determined for all \( k \). If these moduli are known and fixed, then problem (4.1) reduces to a phase synchronization problem with rank-one observation matrix \( aa^* \).

### 5 Analyzing the geometric landscape: the real case

We now investigate a variant of the so-called strict-saddle property introduced by Ge, Jin, and Zheng in [35, Definition 2]. More specifically, as in [72, Theorem 2.2], we
strengthen the first condition in [35, Definition 2] to uniform positive definiteness of the Riemannian Hessian.

**Definition 5.1** Let $\xi, \epsilon, \zeta > 0$ be given constants. A function $f$ is called $(\xi, \epsilon, \zeta)$-strict-saddle if for all $z \in \mathcal{M}$ one of the following conditions holds:

1. (Strong convexity). For all $v \in T_z \mathcal{M} \cap S^{n-1}$ we have $H_f(z)[v] \geq \xi$.
2. (Large gradient). It holds that $\|\text{grad} f(z)\| \geq \epsilon$.
3. (Negative curvature). There exists $v \in T_z \mathcal{M} \cap S^{n-1}$ with $H_f(z)[v] \leq -\zeta$.

The strict-saddle property can be utilized to establish polynomial time convergence rates of second-order optimization algorithms, such as the Riemannian trust region method [72], and almost sure convergence to local minimizer of Riemannian gradient descent methods with random initialization, see, e.g., [48,49].

In the following sections, we analyze the geometric landscape of problem (1.1) and show that the strict-saddle property is satisfied in the real case when the interaction coefficient $\beta$ is sufficiently small or large. In general, due to the intricate relation between the quadratic and quartic terms, we can not expect the conditions in Definition 5.1 to hold for all choices of $\beta$ and $A$.

For instance, let us consider the example $A := \alpha 11^T$, $\alpha \geq 0$, and $z := 1/\sqrt{n} \in S^{n-1}$. Then, the associated multiplier $\lambda$ is given by $2\lambda = \alpha n + 2\beta/n$ and it can be shown that $z$ is a stationary point of (1.1) for all $\alpha$. Furthermore, for all $v \in T_z \mathcal{M}$, it follows

$$H_f(z)[v] = v^T [A + 6\beta \text{diag}(|z|^2) - 2\lambda I]v = [4\beta n^{-1} - \alpha n]\|v\|^2.$$ 

Hence, in the case $\alpha = 4\beta/n^2$, the strict-saddle property can not hold. Let us further note that the strong convexity condition in Definition 5.1 is never satisfied at stationary points in the complex case. In particular, if $z \in \mathbb{C}S^{n-1}$ is a critical point, then we have $iz \in T_z \mathcal{M} \cap \mathbb{C}S^{n-1}$ and $H_f(z)[iz] = 0$ which contradicts condition 1.

In Fig. 1, we illustrate different landscapes of the mapping $f$ when the problem is real and three-dimensional and the parameter $\beta$ changes. We consider the setup

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \quad \beta \in \{0.25, 0.75, 1.25, 3.25\}, \quad (5.1)$$

and the eigenvalues of $A$ are given by 2, 1, and 0, respectively. Moreover, we use spherical coordinates $(\phi, \theta) \mapsto (\cos(\phi), \sin(\phi) \cos(\theta), \sin(\phi) \sin(\theta))^T$, $\phi \in [0, 2\pi]$, $\theta \in [0, \pi]$, to plot the objective function $f$ on the sphere.

Figure 1 demonstrates that the landscape of the objective function varies a lot when the interaction coefficient $\beta$ changes. Specifically, it shows that the number of stationary points and local minima increases from 6 to 26 and from 2 to 8 as $\beta$ increases from 0.25 to 3.25. In Sects. 5.1 and 5.2, we investigate basic geometric features and the strict-saddle property for large and small choices of $\beta$ while keeping the matrix $A$ fixed. In particular, our results will allow us to characterize and describe the geometric landscape of $f$ in the sub-figures (a) and (d) of Fig. 1. In the following, we assume $n \geq 2$ since the landscape of $f$ is trivial in the case $n = 1$. 
Fig. 1 Plot of the landscape of the objective function for fixed $A$ and different values of $\beta$. The red point marker depicts the location of saddle points. Local and global minima are indicated by non-filled and filled diamond markers. The location of local and global maxima is marked by non-filled and filled squares.

### 5.1 Large interaction coefficient

In this section, we prove that the objective function possesses the strict-saddle property if the coefficient $\beta$ is chosen sufficiently large and satisfies $\beta \approx C\rho n^{3/2}$ for some constant $C > 0$. Here, $\rho := \lambda_1 - \lambda_n$ denotes the difference between the largest and the smallest eigenvalue of $A$. We analyze the geometric properties and behavior of $f$ on the following sets:

1. (Strong convexity) $\mathcal{R}_1 := \{z \in \mathbb{S}^{n-1} : \|z\|^2 - 1/n \leq 1/2n\}$,
2. (Large gradient) $\mathcal{R}_2 := \{z \in \mathbb{S}^{n-1} : \|z\|^2 - 1/n \geq 1/2n, \min_{k \in [n]} z_k^2 \geq (n - 1)/4n^2\}$,
3. (Negative curvature) $\mathcal{R}_3 := \{z \in \mathbb{S}^{n-1} : \min_{k \in [n]} z_k^2 \leq (n - 1)/4n^2\}$.

Obviously, these three regions cover the sphere $\mathbb{S}^{n-1}$. An illustration of the regions $\mathcal{R}_1 - \mathcal{R}_3$ is given later in Fig. 2.

We first present a preparatory result that is used to estimate the spectrum of the Riemannian Hessian.

**Lemma 5.1** The following estimate holds for all $z \in \mathbb{S}^{n-1}$:

$$z_0^2 \leq \min_{v \in T_M \cap \mathbb{S}^{n-1}} \sum_{k \in [n]} v_k^2 z_k^2 \leq \frac{n}{n - 1} (1 - z_0^2)z_0^2, \quad (5.2)$$

where $z_0^2 := \min_{k \in [n]} z_k^2$. 

Springer
Proof If there exists \( k \in [n] \) with \( z_k = 0 \), then the optimal value of the latter optimization problem is 0 and it is attained for \( \mathbf{v} = e_k \). Next, let us assume \( z_k^2 > 0 \) and let \( i \in [n] \) be given with \( z_i^2 = z_0^2 \). We define \( \mathbf{v} \in T_2 \mathcal{M} \cap \mathbb{S}^{n-1} \) via

\[
v_i = -\frac{(n-1)c}{z_i}, \quad v_k = \frac{c}{z_k}, \quad c = \left[ \frac{(n-1)^2}{z_i^2} + \sum_{k \neq i} \frac{1}{z_k^2} \right]^{-1/2}, \quad k \neq i.
\]

Using the reverse Hölder inequality, we have \( \sum_{k \neq i} z_k^{-2} \geq (n-1)^2 \left( \sum_{k \neq i} \frac{1}{z_k^2} \right)^{-1} \) and thus, we obtain

\[
\sum_{k \in [n]} z_k^2 v_k^2 = n(n-1)c^2 \leq \frac{n(n-1)}{z_i^2} + \frac{(n-1)^2}{1-z_0^2} = \frac{n}{n-1} \left( 1 - \frac{z_0^2}{z_i^2} \right).
\]

This establishes the upper bound in (5.2). The lower bound follows from \( \|\mathbf{v}\| = 1 \) and \( z_k^2 \geq z_i^2 \) for all \( k \in [n] \).

In the following lemma, we verify that the objective function has the strong convexity property on the region \( \mathcal{R}_1 \).

Lemma 5.2 Let \( \gamma > 0 \) be given and suppose that \( \beta \geq 2(1 + \gamma) \rho n \). Then, for all \( \mathbf{z} \in \mathcal{R}_1 \) and all \( \mathbf{v} \in T_2 \mathcal{M} \cap \mathbb{S}^{n-1} \) it holds that \( H_f(\mathbf{z})[\mathbf{v}] \geq \gamma \rho \).

Proof Let us set \( w_k := z_k^2 - \frac{1}{n} \). By the definition of \( \mathcal{R}_1 \), we have \( z_k^2 \in \left[ \frac{1}{2n}, \frac{3}{2n} \right] \) for all \( k \) and \( \mathbf{z} \in \mathcal{R}_1 \). Hence, it follows \( w_k^2 \leq 1/4n^2 \) for all \( k \) and due to \( \sum_{k \in [n]} w_k = 0 \), we obtain:

\[
\sum_{k \in [n]} z_k^4 = \sum_{k \in [n]} \left( w_k + \frac{1}{n} \right)^2 = \frac{1}{n} + \sum_{k \in [n]} w_k^2 \leq \frac{5}{4n}.
\]

Next, by Lemma 5.1 and \( \mathbf{z} \in \mathcal{R}_1 \), it holds that

\[
\min_{\mathbf{v} \in T_2 \mathcal{M} \cap \mathbb{S}^{n-1}} \sum_{k \in [n]} z_k^2 v_k^2 = \min_{k \in [n]} z_k^2 \geq \frac{1}{2n}.
\]

Combining the last results and using \( \mathbf{v}^T A \mathbf{v} - \mathbf{z}^T A \mathbf{z} \geq -\rho \) and (2.2), we have

\[
\min_{\mathbf{v} \in T_2 \mathcal{M} \cap \mathbb{S}^{n-1}} H_f(\mathbf{z})[\mathbf{v}] = \min_{\mathbf{v} \in T_2 \mathcal{M} \cap \mathbb{S}^{n-1}} \mathbf{v}^T [A + 6\beta \text{diag}(|\mathbf{z}|^2)] \mathbf{v} - 2\lambda
\]

\[
\geq -\rho + 6\beta \left[ \min_{\mathbf{v} \in T_2 \mathcal{M} \cap \mathbb{S}^{n-1}} \sum_{k \in [n]} z_k^2 v_k^2 \right] - 2\beta \|\mathbf{z}\|_4^4
\]

\[
\geq -\rho + \frac{2\beta}{n} \left[ \frac{3}{2} - \frac{5}{4} \right] \geq \gamma \rho
\]

for all \( \mathbf{z} \in \mathcal{R}_1 \).  

\( \square \)
The next lemma shows that the norm of the Riemannian gradient is strictly larger than zero—uniformly—on the region $R_2$.

**Lemma 5.3** Let $\beta \geq 8(1 + \frac{1}{n^2})(1 + \gamma)\rho n^{3/2}$ be given for some $\gamma > 0$. Then, for all $z \in R_2$, it holds that $\|\text{grad } f(z)\| \geq \frac{\sqrt{2}}{\sqrt{n}}\beta$.

**Proof** First, we split the norm of the Riemannian gradient $\text{grad } f(z)$ as follows

$$\|\text{grad } f(z)\| = 2\beta \|\text{diag}(|z|^2)z - \|z\|_4^4 \cdot z\| - \|Az - (z^T Az) \cdot z\|$$

$$= 2\beta \left|\|z\|_6^6 - \|z\|_4^8 - \|Az - (z^T Az) \cdot z\|\right|.$$

We now estimate the minuend and subtrahend in the latter expression separately. Let us set $z = \sum_k \alpha_k p_k$, where $\{p_k\}_k$ is the set of orthogonal eigenvectors of the matrix $A$ with corresponding eigenvalues $\lambda_1, \ldots, \lambda_n$. Then, it holds that $\sum_k \alpha_k^2 = 1$ and

$$\|Az - (z^T Az) \cdot z\|^2 = z^T A^2 z - (z^T Az)^2 = \sum_k \alpha_k^2 \lambda_k^2 - \left[\sum_k \alpha_k^2 \lambda_k\right]^2$$

$$= \sum_{k, \ell \in [n]} \alpha_k^2 \lambda_k^2 - \sum_{k, \ell \in [n]} \alpha_k^2 \alpha_\ell^2 \lambda_k \lambda_\ell$$

$$= \frac{1}{2} \sum_{k, \ell \in [n]} \alpha_k^2 \alpha_\ell^2 (\lambda_k - \lambda_\ell)^2 = \frac{1}{2} \rho^2 \sum_{k, \ell \in [n]} \alpha_k^2 \alpha_\ell^2 = \frac{1}{2} \rho^2. \quad (5.3)$$

We continue with bounding the term $\|z\|_6^6 - \|z\|_4^8$. As before using the spherical constraint $\sum_k z_k^2 = 1$, we obtain:

$$\|z\|_6^6 - \|z\|_4^8 = \sum_{k \in [n]} z_k^6 - \left[\sum_{k \in [n]} z_k^4 \right] = \frac{1}{2} \sum_{k, \ell \in [n]} z_k^2 z_\ell^2 (z_k^2 - z_\ell^2)^2.$$

In the following and without loss of generality, we assume that the components of $z$ are ordered and satisfy $z_1^2 \leq z_2^2 \leq \cdots \leq z_n^2$. Notice that $z \in R_2$ implies $z_1^2 \neq z_n^2$ and let us choose $t \in [n - 1]$ such that $z_t^2 \leq \frac{1}{2}(z_1^2 + z_n^2) \leq z_{t+1}^2$. Then, we have

$$1 = \sum_{k=1}^t z_k^2 + \sum_{k=t+1}^n z_k^2 \geq tz_1^2 + \frac{n-t-1}{2}(z_1^2 + z_n^2) + z_t^2,$$

which yields

$$t \geq \frac{(n-1)(z_n^2 + z_1^2) + 2z_n^2 - 2}{z_n^2 - z_1^2} \geq \frac{(n-1)z_1^2 + z_n^2 - 1}{z_n^2 - z_1^2}.$$
Consequently, we get
\[
\sum_{k, \ell \in [n]} z_k^2 z_\ell^2 (z_k^2 - z_\ell^2)^2 \geq z_1^2 \sum_{k \in [n]} z_k^2 (z_k^2 - z_1^2)^2 + \sum_{k \in [n]} z_k^2 (z_k^2 - z_n^2)^2
\]
\[
\geq \frac{1}{4} (z_n^2 - z_1^2)^2 \left[ z_1^2 + \sum_{k=t+1}^n z_k^2 + \sum_{k=1}^t z_k^2 \right]
\]
\[
= \frac{1}{4} (z_n^2 - z_1^2)^2 \left[ z_1^2 + (z_n^2 - z_1^2) \sum_{k=1}^t z_k^2 \right]
\]
\[
\geq \frac{1}{4} (z_n^2 - z_1^2)^2 \left[ z_1^2 + (z_n^2 - z_1^2) t z_1^2 \right] \geq \frac{n}{4} (z_n^2 - z_1^2)^2 z_1^4.
\]
Using \(z_1^2 \geq \frac{n-1}{4n^2} \) and \(\max_k |z_k^2 - \frac{1}{n}| = \max(|z_1^2 - \frac{1}{n}|, |z_n^2 - \frac{1}{n}|) \geq \frac{1}{2n} \), it follows
\[
2[\|\mathbf{z}\|_6^6 - \|\mathbf{z}\|_4^8] \geq \frac{n}{4} (z_n^2 - z_1^2)^2 z_1^4 \geq \frac{n}{4} \left( \frac{1}{2n} \right)^2 \left( \frac{n}{4n^2} \right)^2 = \left( \frac{n - 1}{16n^{5/2}} \right)^2
\]
and finally, combining the last results, we obtain
\[
\|\text{grad} f(\mathbf{z})\| \geq \frac{1}{\sqrt{2}} \left[ 2\beta \cdot \frac{n - 1}{16n^{5/2}} - \rho \right] \geq \frac{\gamma}{\sqrt{2}} \rho,
\]
as desired. \(\square\)

Finally, we show that we can always find a negative curvature direction if \(\mathbf{z}\) belongs to the set \(\mathcal{R}_3\).

**Lemma 5.4** Let \(\gamma > 0\) be arbitrary and let \(\beta \geq 2(1 + \gamma) pn\) be given. Then, for all \(\mathbf{z} \in \mathcal{R}_3\), there exists \(\mathbf{v} \in T_{\mathbf{z}} \mathcal{M} \cap \mathbb{S}^{n-1}\) such that \(H_f(\mathbf{z})[\mathbf{v}] \leq -\gamma \rho\).

**Proof** Using the bound \(\mathbf{v}^T A \mathbf{v} - \mathbf{z}^T A \mathbf{z} \leq \rho\) for \(\mathbf{z}, \mathbf{v} \in \mathbb{S}^{n-1}\) and Lemma 5.1, it follows
\[
\min_{\mathbf{v} \in T_{\mathbf{z}} \mathcal{M} \cap \mathbb{S}^{n-1}} H_f(\mathbf{z})[\mathbf{v}] \leq \rho + 2\beta \left[ \frac{3n}{n - 1} (1 - z_0^2) z_0^2 - \|\mathbf{z}\|_4^4 \right],
\]
where \(z_0^2 := \min_{k \in [n]} z_k^2\). Moreover, due to \(\|\mathbf{z}\|_4^4 \geq \frac{1}{n} \|\mathbf{z}\|_4^4 = \frac{1}{n}\) and \(\mathbf{z} \in \mathcal{R}_3\), we can infer
\[
\min_{\mathbf{v} \in T_{\mathbf{z}} \mathcal{M} \cap \mathbb{S}^{n-1}} H_f(\mathbf{z})[\mathbf{v}] \leq \rho + 2\beta \left[ \frac{3n}{n - 1} z_0^2 - \frac{1}{n} \right] \leq \rho - \frac{\beta}{2n} \leq -\gamma \rho,
\]
which finishes the proof of Lemma 5.4. \(\square\)

Combining the last lemmata, we can derive the following strict-saddle property.
Theorem 5.1 Suppose that the coefficient $\beta$ satisfies $\beta \geq \frac{8n}{n-1}(1 + \gamma)\rho n^{3/2}$ for some given $\gamma > 0$. Then, the function $f$ has the $(C_\gamma, \rho, \frac{\gamma}{\sqrt{2}}\rho, C_\gamma \rho)$-strict-saddle property with $C_\gamma := \frac{4}{n-1}(1 + \gamma)n^{3/2} - 1$.

As a consequence of the strict-saddle property, we can prove that each component of $\mathcal{R}_1$ contains exactly one local minimizer when $\beta$ is sufficiently large. In the next lemma, we first discuss uniqueness of local minimizer if the Riemannian Hessian is positive definite on a certain subset of the sphere.

Lemma 5.5 Let $v \in (0, 1]$ and $z_0 \in S^{n-1}$ be given and let us define $\mathcal{R}_v := \{ z \in S^{n-1} : \|z - z_0\|^2 \leq v \}$. If the Riemannian Hessian $H_f$ of $f$ is positive definite on $\mathcal{R}_v$, i.e., if we have

$$v^T \text{Hess } f(z)v = H_f(z)[v] > 0, \quad \forall \ v \in T_z \mathcal{M}\setminus\{0\}, \quad \forall \ z \in \mathcal{R}_v,$$

then the problem $\min_{z \in \mathcal{R}_v} f(z)$ has at most one local minimizer.

Proof Suppose that there exist two different local minima $z_1, z_2$ of $f$ in the set $\mathcal{R}_v$. Let us consider the geodesic curve $\ell : \mathbb{R} \to S^{n-1}$ on the sphere connecting $z_1$ and $z_2$. As shown in [2, Example 5.4.1], the curve $\ell$ can be represented as follows

$$\ell(t) := z_1 \cos(t) + v \sin(t), \quad v := \frac{z_2 - (z_2^T z_1)z_1}{\|z_2 - (z_2^T z_1)z_1\|} \in T_{z_1} \mathcal{M} \cap S^{n-1},$$

where $T \in (0, 2\pi)$ is chosen such that $\ell(T) = z_2$. Multiplying $\ell(T) = z_2$ with $v^T$ from the left yields $\sin(T) = v^T z_2 \geq 0$ and hence, we have $T \in (0, \pi]$. Then, for all $t \in (0, T)$, it follows $\ell(t)^T \ell'(t) = 0$, $\|\ell'(t)\| = 1$, and

$$\|\ell(t) - z_0\|^2 = 2 - 2 \left( z_0^T z_1 \cdot \cos(t) + z_0^T v \cdot \sin(t) \right) = 2 - 2M \cos(t + \theta),$$

where $M = \sqrt{(z_0^T z_1)^2 + (z_0^T v)^2}$ and $\theta$ is chosen such that $\cos(\theta) = z_0^T z_1 / M$, $\sin(\theta) = -z_0^T v / M$. Since $\ell(0)$ and $\ell(T)$ are both elements of $\mathcal{R}_v$, we know that

$$\|z_0 - \ell(0)\|^2 \leq v \leq 1, \quad \|z_0 - \ell(T)\|^2 \leq v \leq 1,$$

which lead to $\cos(\theta), \cos(\theta + T) > 0$. Due to $T \in (0, \pi]$, we have $\theta, \theta + T \in (-\frac{\pi}{2} + 2k\pi, \frac{\pi}{2} + 2k\pi)$ for some $k \in \mathbb{Z}$. Utilizing the landscape of $t \mapsto \cos(t)$ on those intervals, the maximal value of $[0, T] \ni t \mapsto \|z_0 - \ell(t)\|^2$ is achieved at one of the endpoints $t = 0$ or $t = T$. Hence, it holds that $\|z_0 - \ell(t)\|^2 \leq v$ and $\ell(t) \in \mathcal{R}_v$ for all $t \in [0, T]$.

The special form of the curve $\ell$ yields $\ell''(t) = -\ell(t)$ for all $t \in \mathbb{R}$ and thus, the second-order derivative of the continuous function $g(t) := f(\ell(t))$ is given by...
Corollary 5.1 If \( \beta > 4\rho n^2 \), the problem (1.1) has at least \( 2^n \) local minima. Furthermore, if \( \beta > \frac{18n^3}{n-1}\rho \), then the problem (1.1) has exactly \( 2^n \) local minima.

Proof Without loss of generality, we can assume \( \lambda_n(A) = \lambda_n = 0 \). We first prove that there exists at least one local minimizer in each component of the region \( \mathcal{R}_1 \) if \( \beta > 4\rho n^2 \). Let \( \sigma \in \{-1, +1\}^n \) be a given binary vector and let us define the sets \( B := \{ z \in S^{n-1} : \| z \|^2 - \frac{1}{n} \leq \frac{1}{2n} \} \) and \( \mathcal{P}_\sigma := \prod_{k \in [n]} \sigma_k \mathbb{R}_+ \). We now show that for each possible choice of \( \sigma \) the set \( B \cap \mathcal{P}_\sigma \) contains a local minimizer of \( f \). Let us note that, for all \( z \in S^{n-1} \), the condition \( \| z \|^2 - \frac{1}{n} \leq \frac{1}{2n} \) is equivalent to \( \| z \|^2 \leq \frac{1}{\beta} + \frac{1}{4n^2} \). This observation can be used to establish \( z_k \neq 0 \) for all \( z \in B \cap \mathcal{P}_\sigma \) and consequently, the sets \( B \cap \mathcal{P}_\sigma \) and \( B \cap \mathcal{P}_v \) do not intersect for different binary vectors \( \sigma \neq v \). In particular, if there exists \( k \) with \( z_k = 0 \), then it follows \( \| z \|^4 \geq \frac{1}{n-1} \| z \|^2 = \frac{1}{n-1} \) which yields a contradiction to \( \| z \|^4 \leq \frac{1}{n} + \frac{1}{4n^2} \). Now, for all \( z \in D := \{ z \in S^{n-1} : \| z \|^2 = \frac{1}{n} + \frac{1}{4n^2} \} \), we have

\[
f(z) \geq \beta \left[ \frac{1}{2} \frac{1}{n} + \frac{1}{4n^2} \right] \tag{5.4}
\]

and setting \( z_\sigma = \sigma / \sqrt{n} \in B \cap \mathcal{P}_\sigma \), we obtain

\[
f(z_\sigma) = \frac{1}{2} z_\sigma^T A z_\sigma + \frac{\beta}{2n} \leq \frac{\rho}{2} + \frac{\beta}{2n} < \beta \left[ \frac{1}{n} + \frac{1}{4n^2} \right]. \tag{5.5}
\]

Hence, the global minimizer \( y \) of the problem \( \min_{z \in B \cap \mathcal{P}_\sigma} f(z) \) satisfies \( \| y \|^4 \leq \frac{1}{n} + \frac{1}{4n^2} \). This implies that the Lagrange multiplier associated with the constraint \( \| z \|^4 \leq \frac{1}{n} + \frac{1}{4n^2} \) is zero, and the corresponding KKT conditions for \( y \) reduce to \( \text{grad} \ f(y) = 0 \). Due to \( B \cap \mathcal{P}_\sigma \subset \mathcal{R}_1 \cap \mathcal{P}_\sigma \), the Riemannian Hessian \( \text{Hess} \ f(y) \) is positive definite on the tangent space \( T_y \mathcal{M} \setminus \{0\} \) and thus, by Lemma 2.1, the point \( y \) is one of at least \( 2^n \) isolated local minima of problem (1.1).

Next, we consider the case when \( \beta > \frac{18n^3}{n-1}\rho \). We introduce the following refined versions of the \( \mathcal{R}_1 \) and \( \mathcal{R}_2 \):

\[
\tilde{\mathcal{R}}_1 = \bigcup_{\sigma \in \{-1, +1\}^n} \left\{ z \in S^{n-1} : \| z - \frac{1}{\sqrt{n}} \sigma \| \leq \frac{2}{9\sqrt{n}} \right\},
\]

\[
\tilde{\mathcal{R}}_2 = \bigcap_{\sigma \in \{-1, +1\}^n} \left\{ z \in S^{n-1} : \| z - \frac{1}{\sqrt{n}} \sigma \| \geq \frac{2}{9\sqrt{n}}, \min_{k \in [n]} z_k^2 \geq \frac{n-1}{4n^2} \right\}.
\]
It can be easily seen that the set $\bar{R}_1$ consists of $2^n$ non-intersecting components and that the three regions $\bar{R}_1$, $\bar{R}_2$, and $R_3$ cover the sphere $S^{n-1}$. Now, let $z \in \bar{R}_1$ be arbitrary. Then, there exists $\sigma \in \{\pm 1\}^n$ such that

$$\left| z_k - \frac{\sigma_k}{\sqrt{n}} \right| \leq \frac{2}{9 \sqrt{n}} \quad \text{and} \quad \left| z_k + \frac{\sigma_k}{\sqrt{n}} \right| \leq \frac{20}{9 \sqrt{n}}, \quad \forall \ k \in [n].$$

Hence, it follows $\|z^2 - n^{-1} 1\|_\infty \leq \frac{40}{81 n} < \frac{1}{2n}$, which implies $\bar{R}_1 \subset R_1$. Thus, the strong convexity property also holds on $\bar{R}_1$. We now prove that the Riemannian gradient is lower bounded on $\bar{R}_2$. For every $z \in \bar{R}_2$, there exists $\sigma \in \{\pm 1\}^n$ such that $\sigma z_k = |z_k|$ for all $k \in [n]$. Consequently, we obtain

$$\|z^2 - n^{-1} 1\|_\infty = \max_{k \in [n]} \left| z_k - \frac{\sigma_k}{\sqrt{n}} \right| \geq \frac{1}{\sqrt{n}} \left( \frac{2}{9 n \sqrt{n}} \right)^2 \left( \frac{n - 1}{4 n^2} \right)^2 = \left( \frac{n - 1}{36 n^3} \right)^2.$$

Thus, the norm of the Riemannian gradient is lower bounded by

$$\|\nabla f(z)\| \geq \frac{1}{\sqrt{2}} \left[ 2\beta \sqrt{2[\|z\|_6^6 - \|z\|_4^8]} - \rho \right] \geq \frac{1}{\sqrt{2}} \left( \beta \cdot \frac{n - 1}{18 n^3} - \rho \right) > 0$$

and all local minima have to be located in the set $\bar{R}_1$. Applying Lemma 5.5, each connected component of $\bar{R}_1$ contains at most one local minimizer and hence, problem (1.1) has exactly $2^n$ local minima. \hfill $\square$

Although the local minima of problem (1.1) are not unique, all the local minima have similar objective function values.

**Theorem 5.2** Suppose that $\beta > \frac{18 n^3}{n-1} \rho$. Then, it follows

$$f(y) - \min_{z \in S^{n-1}} f(z) \leq \frac{1}{18 n} \cdot \left[ \min_{z \in S^{n-1}} f(z) - \lambda_n(A) \right], \quad (5.6)$$

for all local minimizer $y \in S^{n-1}$ of problem (1.1) where $\lambda_n(A)$ denotes the smallest eigenvalue of the matrix $A$.

**Proof** Without loss of generality, we can assume that the smallest eigenvalue of the matrix $A$ is zero. Then, for all $z \in S^{n-1}$, we have $f(z) \geq \frac{\beta}{4} \|z\|_4^4 \geq \frac{\beta}{2n}$. According to the analysis in Corollary 5.1, each component $B \cap P_\sigma$, $\sigma \in \{\pm 1\}^n$, contains exactly one local minimizer $y$ of problem (1.1). Due to the estimates (5.4) and (5.5), there is no global minimizer of the restricted problem $\min_{z \in B \cap P_\sigma} f(z)$ on the boundary of $B \cap P_\sigma$. Hence, all global minimizer of the restricted problem $\min_{z \in B \cap P_\sigma} f(z)$ are also
local minimizer of the original problem (1.1). Thus, $y$ is the unique global minimizer of $\min_{z \in B \cap P_n} f(z)$. Together with (5.5), this yields

$$f(y) \leq f(z) \leq \frac{\rho}{2} + \frac{\beta}{2n} \leq \left[1 + \frac{n\rho}{\beta}\right] f(z), \quad \forall z \in \mathbb{S}^{n-1}. $$

Finally, the estimate (5.6) can be established via minimizing the latter expression with respect to $z$ and using the bound on $\beta$. The general case for $\lambda_n \neq 0$ can be obtained via shifting $f$, i.e., by setting $f(x) \to f(x) + \lambda_n, x \in \mathbb{S}^{n-1}$. \qed

In the remainder of this section, we present an example demonstrating that the bound $\beta \approx C\rho n^{3/2}$ and the dependence on $n^{3/2}$ cannot be further improved and that the strict-saddle property is violated whenever a smaller coefficient is chosen.

**Example 1** Let $C > 0$ and $\epsilon > 0$ be given constants and suppose $\beta = Cn^{3/2-\epsilon}$. In the following, we construct a specific matrix $A \in \mathbb{R}^{n \times n}$ and a point $z \in \mathbb{S}^{n-1}$ such that the three conditions of the strict-saddle property do not hold at $z$. We set

$$z_1 = \left[\frac{1}{3n-2}\right]^\frac{1}{2}, \quad z_k = \left[\frac{3}{3n-2}\right]^\frac{1}{2}, \quad k \geq 2, $$

$$u = 2\beta \|z\|_4^4 \cdot z - 2\beta \text{diag}(|z|^2)z, \quad \text{and} \quad A = \alpha ww^T + zu^T + uz^T, $$

where

$$w_1 = -\left[\frac{3n-3}{3n-2}\right]^\frac{1}{2}, \quad w_k = \left[\frac{1}{(3n-2)(n-1)}\right]^\frac{1}{2}, \quad k \geq 2, \quad \alpha = -\frac{16\beta}{(3n-2)^2}. $$

Then, we have $\|z\| = \|w\| = 1, \|z\|_4 = (9n-8)/(3n-2)^2$, and

$$\|u\|^2 = 4\beta^2 [\|z\|_6^6 - \|z\|_4^8] = 4\beta^2 \left[\frac{1 + 27(n-1)}{(3n-2)^3} - \|z\|_4^8\right] = \frac{48(n-1)}{(3n-2)^4} \cdot \beta^2. $$

Furthermore, it holds that $u^T z = w^T z = 0$ which implies $A z = u$ and grad $f(z) = 0$. The eigenvalues of the matrices $\alpha ww^T$ and $zu^T + uz^T$ are $0, \ldots, 0, \alpha$ and $\|u\|, 0, \ldots, 0, -\|u\|$, respectively. Thus, by Weyl’s inequality, it follows

$$\rho = \lambda_1(A) - \lambda_n(A) \leq 2\|u\| - \alpha = 8\beta \cdot \frac{2 + \sqrt{3n-3}}{(3n-2)^2} \leq \tilde{C} n^{-\epsilon} = O(n^{-\epsilon}), $$

for some constant $\tilde{C}$. Let us now consider an arbitrary vector $v \in T_z M \cap \mathbb{S}^{n-1}$. We have $\sum_{k=2}^n v_k^2 = 1 - v_1^2, v_1 + \sqrt{3} \sum_{k=2}^n v_k = 0, \text{and}$

$$v^T w = v_1 w_1 - \frac{1}{\sqrt{3}} \left[\frac{1}{(3n-2)(n-1)}\right]^\frac{1}{2} v_1 = -\left[\frac{3n-2}{3n-3}\right]^\frac{1}{2} \cdot v_1. $$

\[\text{ Springer}\]
Suppose that the gap between the two smallest eigenvalues of the matrix $A$ satisfies $\delta := \lambda_{n-1} - \lambda_n > 0$ and let $\beta, \gamma > 0$ be given. Then, for all $z \in R_1$ and all $v \in T_z M \cap S^{n-1}$, it follows $H_f(z)[v] \geq \gamma \beta$. 

**Lemma 5.6** Suppose that the gap between the two smallest eigenvalues of the matrix $A$ satisfies $\delta := \lambda_{n-1} - \lambda_n > 0$ and let $\beta, \gamma > 0$ be given. Then, for all $z \in R_1$ and all $v \in T_z M \cap S^{n-1}$, it follows $H_f(z)[v] \geq \gamma \beta$.

Hence, we obtain

$$H_f(z)[v] = \alpha (v^T w)^2 + 6\beta v^T \text{diag}(|z|^2) v - 2\beta \|z\|^4$$  

$$= \frac{3n - 2}{3n - 3} \alpha v_1^2 - \frac{12\beta}{3n - 2} v_2^2 + \frac{18\beta}{3n - 2} - 2\beta \|z\|^4.$$ 

Since $H_f(z)[v]$ is linear in $v_1^2$, its maximum and minimum value are attained at the boundary of the range interval $[v, \bar{v}]$ of $v_1^2$. Notice that we have $\nu = 0$ and $\bar{v}$ can be found by discussing the optimization problem

$$\min_{v_1} \nu_1 \quad \text{s.t.} \quad v_1 + \sqrt{3} \sum_{k=2}^n v_k = 0, \quad \|v\| = 1$$

and its associated KKT conditions. In particular, it can be shown that $\bar{v} = (3n - 3)/(3n - 2)$. In the case $v_1^2 = 0$, we obtain $H_f(z)[v] > 0$ and for $v_1^2 = \bar{v}$, we have

$$H_f(z)[v] = \alpha - \frac{36\beta(n - 1)}{(3n - 2)^2} + \frac{18\beta}{3n - 2} - \frac{2\beta(9n - 8)}{(3n - 2)^2} = \alpha + \frac{16\beta}{(3n - 2)^2} = 0.$$ 

Consequently, we can infer $\min_{v \in T_z M \cap S^{n-1}} H_f(z)[v] = 0$ and $H_f(z)[v] \geq 0$ for all $v \in T_z M \cap S^{n-1}$, which shows that none of the conditions of the strict-saddle property hold at $z$. Thus, the order $n^{3/2}$ cannot be improved in the deterministic case.

**5.2 Small interaction coefficient**

In the following, we discuss the geometric landscape of problem (1.1) for small interaction coefficients. We additionally assume that there is a positive spectral gap $\delta = \lambda_{n-1} - \lambda_n > 0$ between the two smallest eigenvalues of the matrix $A$. As shown by Marvcenko [58], this condition holds with probability 1 when $A$ is a Gaussian random matrix.

Let us recall that the eigenvalue decomposition of $A$ is given by $A = P \Lambda P^T$, where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ and $P = (p_1, p_2, \ldots, p_n)$ is an orthogonal matrix. Similar to the method used in Sect. 5.1, we now divide $S^{n-1}$ into three sub-regions:

1. (Strong convexity) $R_1 := \{z \in S^{n-1} : z = \sum_{k=1}^{n} \alpha_k p_k, \alpha_k^2 \geq \frac{(2+\gamma) \beta + \rho}{\delta + \rho}\}.$
2. (Large gradient) $R_2 := \{z \in S^{n-1} : \|Az\|^2 - (z^T Az)^2 \geq \frac{\gamma^2}{\delta + \rho} \beta^2\}.$
3. (Negative curvature) $R_3 := \{z \in S^{n-1} : z = \sum_{k=1}^{n} \alpha_k p_k, \alpha_k^2 \leq \frac{\delta - (4+\gamma) \beta}{\delta + \rho}\}.$

In Theorem 5.3, it will be shown that those three regions can cover the whole sphere when $\beta$ is sufficiently small. An exemplary illustration of the sets $R_1,R_3$ is given in Fig. 2. We first show that the Riemannian Hessian is uniformly positive definite on $R_1$.

**Lemma 5.6** Suppose that the gap between the two smallest eigenvalues of the matrix $A$ satisfies $\delta := \lambda_{n-1} - \lambda_n > 0$ and let $\beta, \gamma > 0$ be given. Then, for all $z \in R_1$ and all $v \in T_z M \cap S^{n-1}$, it follows $H_f(z)[v] \geq \gamma \beta$. 

 Springer
Fig. 2  Plot of the sets $\mathcal{R}_1$–$\mathcal{R}_3$ for the problem (5.1). a Depicts the regions introduced in Sect. 5.2 for a small interaction coefficient $\beta = 0.2$. The overlap of the sets $\mathcal{R}_1$–$\mathcal{R}_2$ and $\mathcal{R}_2$–$\mathcal{R}_3$ is shown in green. The set $\mathcal{R}_1$ is the union of the yellow and the two surrounding green areas, while $\mathcal{R}_2$ is the union of all green and light blue areas. The region $\mathcal{R}_3$ coincides with the union of the dark blue sets and the enclosing green area. b Shows the regions introduced in Sect. 5.1 for large $\beta$. Here, the (disjoint) yellow, turquoise, and dark blue areas directly correspond to the sets $\mathcal{R}_1$, $\mathcal{R}_2$, and $\mathcal{R}_3$, respectively. The red point marker again depicts the location of saddle points. Non-filled and filled diamond markers are used for local and global minima. Local and global maxima are marked by non-filled and filled squares.

**Proof** Let $v \in T_zM \cap S^{n-1}$ be arbitrary with $v = \sum_{k \in [n]} v_k p_k$. Then, we have $\sum_{k \in [n]} \alpha_k v_k = 0$ and $\sum_{k \in [n]} v_k^2 = 1$ and the Cauchy inequality implies

$$|v_n \alpha_n| = \left| \sum_{k=1}^{n-1} v_k \alpha_k \right| \leq \left[ \sum_{k=1}^{n-1} v_k^2 \right]^{\frac{1}{2}} \left[ \sum_{k=1}^{n-1} \alpha_k^2 \right]^{\frac{1}{2}} = \sqrt{(1 - v_n^2)(1 - \alpha_n^2)}.$$

Thus, it follows $v_n^2 \alpha_n^2 \leq (1 - v_n^2)(1 - \alpha_n^2)$ and $\alpha_n^2 \leq 1 - v_n^2$ and, due to $\|z\|_4 \leq 1$, we obtain

$$H_f(z)[v] = v^T Av - z^T Az + 6\beta \sum_{k=1}^{n} z_k^2 v_k^2 - 2\beta\|z\|_4^4 \geq \sum_{k=1}^{n} \lambda_k (v_k^2 - \alpha_k^2) - 2\beta$$

$$\geq \sum_{k=1}^{n-1} (\lambda_n + \delta) v_k^2 + \lambda_n v_n^2 - \sum_{k=1}^{n-1} (\lambda_n + \rho) \alpha_k^2 - \lambda_n \alpha_n^2 - 2\beta$$

$$= (1 - v_n^2)\delta - (1 - \alpha_n^2)\rho - 2\beta$$

$$\geq \alpha_n^2 \delta - (1 - \alpha_n^2)\rho - 2\beta = \alpha_n^2 (\delta + \rho) - \rho - 2\beta \geq \gamma \beta,$$

as desired. \qed

Next, we prove that the norm of the Riemannian gradient is strictly larger than zero on the set $\mathcal{R}_2$.

**Lemma 5.7** For all $z \in \mathcal{R}_2$, it holds that $\|\text{grad } f(z)\|_2 \geq \gamma \beta$.
Proof As in the proof of Lemma 5.3, we have
\[
\left\| \left( \text{diag}(|z|^2) - \|z\|_4^4 \cdot I_n \right) z \right\|^2 = \|z\|_6^6 - \|z\|_4^8 \leq \|z\|_6^6 (1 - \|z\|_4^2) \leq \frac{1}{4} \left[ \frac{3}{4} \right]^3,
\]
where the last estimate follows from the fact that the mapping \( x \mapsto x^6 - x^8 \) attains its global maximum at \( x = \pm \sqrt[3]{\frac{3}{2}} \). Hence, we obtain
\[
\|\text{grad} f(z)\|_2 = \| (A - z^T A z \cdot I_n) z + 2\beta \left( \text{diag}(|z|^2) - \|z\|_4^4 \cdot I_n \right) \cdot z \|
\geq \| (A - z^T A z \cdot I_n) z \| - 2\beta \left( \text{diag}(|z|^2) - \|z\|_4^4 \cdot I_n \right) \cdot z \|
\geq \| (A - z^T A z \cdot I_n) z \| - \frac{3\sqrt{3}}{8} \beta = \left[ \frac{2}{3} - \frac{3\sqrt{3}}{8} \right] \beta + \gamma \beta \geq \gamma \beta.
\]

Finally, for points in the region \( R_3 \), we construct directions along which the curvature of the objective function is strictly negative.

Lemma 5.8 Suppose that the gap between the two smallest eigenvalues of the matrix \( A \) satisfies \( \delta := \lambda_{n-1} - \lambda_n > 0 \) and let \( \gamma > 0 \) be given. If \( \beta \leq (4 + \gamma)^{-1} \delta \), then for all \( z \in R_3 \) there exists \( v \in T_z M \cap S^{n-1} \) such that \( H_f(z)(v) \leq - \gamma \beta \).

Proof By the Cauchy’s inequality and using the estimate \( |3x - x^2| \leq 2, x \in [0, 1] \), we have
\[
\sum_{k \in [n]} 3z_k^2 v_k^2 - z_k^4 \leq 3\|z\|_4^2 \|v\|_4^2 - \|z\|_4^4 \leq 3\|z\|_4^2 - \|z\|_4^4 \leq 2.
\]
Next, we choose a specific direction \( v = \sum_{k \in [n]} v_k p_k \) that satisfies \( H_f(z)(v) \leq - \gamma \beta \), \( z^T v = \sum_{k \in [n]} \alpha_k v_k = 0 \) and \( \|v\|^2 = \sum_{k \in [n]} v_k^2 = 1 \).

Case 1. \( \alpha_n = 0 \). Let us set \( v_n = 1 \) and \( v_k = 0 \) for all \( k \in [n-1] \). Then, we obtain
\[
H_f(z)(v) = v^T A v - z^T A z + 2\beta \sum_{k=1}^n (3z_k^2 v_k^2 - z_k^4)
\leq \sum_{k=1}^n \lambda_k (v_k^2 - \alpha_k^2) + 4\beta = \lambda_n - \sum_{k=1}^{n-1} \lambda_k \alpha_k^2 + 4\beta
\leq \lambda_n - (\lambda_n + \delta) \sum_{k=1}^{n-1} \alpha_k^2 + 4\beta = -\delta + 4\beta \leq -\gamma \beta,
\]
where the last inequality immediately follows from \( \beta \leq \frac{\delta}{4+\gamma} \).
**Case 2.** \(0 < \alpha_n^2 \leq \frac{\delta-(4+\gamma)\beta}{\delta+\rho}\). In this situation, we set

\[v_k = \eta \alpha_k, \quad \forall \ k \in [n-1], \quad \nu_n = -\sqrt{1 - \alpha_n^2}, \quad \eta = \frac{\alpha_n}{\sqrt{1 - \alpha_n^2}}.\]

With this choice, we have \(z^T v = \eta (1 - \alpha_n^2) - \alpha_n \sqrt{1 - \alpha_n^2} = 0\) and \(\|v\|^2 = (\eta^2 + 1)(1 - \alpha_n^2) = 1\), i.e., it holds that \(v \in TzM \cap S^{n-1}\). Similar to the calculations in the proof of Lemma 5.6, we now get

\[H_f(z)[v] = v^T A v - z^T A z + 2\beta \sum_{k=1}^{n} (3z_k^2 v_k^2 - z_k^4)\]

\[\leq \sum_{k=1}^{n} \lambda_k (v_k^2 - \alpha_k^2) + 4\beta \]

\[\leq \sum_{k=1}^{n-1} ((\lambda_n + \rho)v_k^2 - (\lambda_n + \delta)\alpha_k^2) + \lambda_n (v_n^2 - \alpha_n^2) + 4\beta\]

\[= \rho \cdot \sum_{k=1}^{n-1} v_k^2 - \delta \cdot \sum_{k=1}^{n-1} \alpha_k^2 + 4\beta = \rho (1 - v_n^2) - \delta (1 - \alpha_n^2) + 4\beta\]

\[= (\delta + \rho)\alpha_n^2 - \delta + 4\beta \leq -\gamma \beta.\]

Combining the latter two cases, we can conclude the proof of Lemma 5.8. \(\square\)

We now verify that \(f\) has the strict-saddle property whenever \(\beta\) is chosen sufficiently small.

**Theorem 5.3** Suppose that the gap between the two smallest eigenvalues of the matrix \(A\) satisfies \(\delta := \lambda_{n-1} - \lambda_n > 0\) and let \(\gamma > 0\) be given. If \(\beta \leq \left[2(\frac{\gamma}{3} + \gamma) + (\frac{\gamma}{3} + \gamma) \delta \right]^{-1} \delta =: b_\gamma\), then \(f\) has the \((\gamma \beta, \gamma \beta, \gamma \beta)\)-strict-saddle property.

**Proof** By Lemmata 5.6–5.8, we know that the function \(f\) satisfies the strong convexity, large gradient, and negative curvature property on the three different set \(\mathcal{R}_1, \mathcal{R}_2,\) and \(\mathcal{R}_3,\) respectively. To finish the proof, we need to show that those regions actually cover the whole sphere \(S^{n-1}\). Combining these observations, we can then conclude that \(f\) has the \((\gamma \beta, \gamma \beta, \gamma \beta)\)-strict-saddle property.

In order to prove \(\mathcal{R}_1 \cup \mathcal{R}_2 \cup \mathcal{R}_3 = S^{n-1}\), we only need to verify that for all \(z = \sum_{k \in [n]} \alpha_k p_k \in S^{n-1}\) with

\[\frac{\delta - (4+\gamma)\beta}{\delta + \rho} \leq \alpha_n^2 \leq \frac{(2+\gamma)\beta + \rho}{\delta + \rho},\]

we have \(\|Az\|^2 - (z^T Az)^2 \geq (\frac{\gamma}{3} + \gamma)^2 \beta^2\). Using the bounds (5.7), it follows

\[\square\]
\[ \| Az \|^2 - (z^T A z)^2 \geq \frac{1}{2} \sum_{k,j \in [n]} \alpha_k^2 \alpha_j^2 (\lambda_k - \lambda_j)^2 \]

where the first identity was established in the proof of Lemma 5.3. Rearranging the terms in the latter estimate, we see that our claim is satisfied if

\[ \ell(\beta) := \left[ (\frac{3}{4} + \gamma)^2 (\delta + \rho)^2 - (4 + \gamma)(2 + \gamma) \right] \beta^2 + 2(3 + \gamma)\delta^3 \beta - \delta^4 \leq 0 \]

for all \( \beta \leq b_\gamma \). Since the unique nonnegative zero of the quadratic polynomial \( \ell \) is given by

\[ \bar{\beta} = \left[ 3 + \gamma + \sqrt{1 + (\frac{3}{4} + \gamma)^2 (1 + \frac{\rho}{\delta})^2} \right]^{-1} \delta, \]

we can finish the proof by noticing \( \bar{\beta} \geq b_\gamma \).

Finally, as a counterpart of Corollary 5.1, we can establish the uniqueness of local minima as a consequence of the strict-saddle property.

**Corollary 5.2** Under the conditions of Theorem 5.3, the problem (1.1) has exactly two local minima which are also global minima.

**Proof** Note that all local minima locate in \( \mathcal{R}_1 \) and that \( \mathcal{R}_1 \) consists of two symmetrical non-intersecting subsets. We now consider one of the subsets \( \tilde{\mathcal{R}}_1 := \{ z \in S^{n-1} : z = \sum_{k \in [n]} \alpha_k p_k, \alpha_n \geq \sqrt{\nu} \} \) where \( \nu := (\delta + \rho)^{-1}((2 + \gamma)\beta + \rho) \). Using \( \| z - p_n \|^2 = 2 - 2\alpha_n \), for \( z \in \tilde{\mathcal{R}}_1 \), an equivalent definition of this subset is given by

\[ \tilde{\mathcal{R}}_1 = \left\{ z \in S^{n-1} : \| z - p_n \|^2 \leq 2 - 2\sqrt{\nu} \right\}. \]

In order to apply Lemma 5.5, we need to verify \( 2 - 2\sqrt{\nu} \leq 1 \) or, equivalently, \( \nu \geq \frac{1}{4} \). However, due to \( \gamma, \beta > 0 \) and \( \rho \geq \delta > 0 \), we have \( \nu \geq \frac{\rho}{\delta + \rho} \geq \frac{1}{2} > \frac{1}{4} \). Similarly, the second subset can be represented by \( \tilde{\mathcal{R}}_2 := \{ z \in S^{n-1} : \| z + p_n \|^2 \leq 2 - 2\sqrt{\nu} \} \). Then, by Lemma 5.5, there exist exactly two equivalent local minima which are also global minima of problem (1.1).

**6 Estimation of the Kurdyka–Łojasiewicz exponent**

In this section, we estimate the Kurdyka–Łojasiewicz (KL) exponent of problem (1.1). Specifically, we want to find the largest \( \theta \in (0, \frac{1}{2}] \) such that for all stationary points \( z \)
of problem (1.1), the Łojasiewicz inequality,

$$|f(y) - f(z)|^{1-\theta} \leq \eta_z \|\text{grad } f(y)\|, \quad \forall y \in B(z, \delta_z) \cap \mathbb{C}S^{n-1},$$

(6.1)

holds with some constants $\delta_z, \eta_z > 0$. Any $\theta$ satisfying (6.1) is called KL exponent of problem (1.1).

As already mentioned, the Łojasiewicz inequality (6.1) plays a fundamental role in nonconvex optimization and is frequently utilized to analyze the local convergence properties of nonconvex optimization methods [6–8,15,16,50,53,62]. In [6], Attouch and Bolte derived an abstract KL-framework based on the Łojasiewicz inequality that allows to establish global convergence and local convergence rates for general optimization approaches satisfying certain function reduction and asymptotic step size safe-guard conditions. In particular, if $\theta \geq \frac{1}{2}$, then the corresponding iterates can be shown to converge linearly. Otherwise, if $\theta \in (0, \frac{1}{2})$, the iterates converge at a sublinear rate $O(t^{-\frac{1-\theta}{2\theta}})$. By introducing the auxiliary problem

$$\min_{z \in \mathbb{C}^n} \hat{f}(z), \quad \hat{f}(z) := \begin{cases} f(z) & \text{if } z \in \mathbb{C}S^{n-1}, \\ +\infty & \text{otherwise}, \end{cases}$$

(6.2)

the original problem (1.1) can be treated as the minimization of an extended real-valued, proper, and lower semicontinuous function which allows to apply existing results and the rich KL theory for nonsmooth problems, see, e.g., [12–14,47,55].

In the nonsmooth setting, the Riemannian gradient, appearing in (6.1), is typically substituted by the nonsmooth slope of $\hat{f}$ which is based on the Fréchet and limiting subdifferential of $\hat{f}$. In our case, if problem (6.2) is restricted to the real space $\mathbb{R}^n$, the limiting subdifferential and nonsmooth slope of $\hat{f}$ at $z \in S^{n-1}$ can be expressed as

$$\partial \hat{f}(z) = \{\nabla f(z) + \mu z : \mu \in \mathbb{R}\}$$

and

$$\arg\min_{g \in \partial \hat{f}(z)} \|g\| = (I - zz^T)\nabla f(z) = \text{grad } f(z).$$

Hence, the Riemannian-type Łojasiewicz inequality (6.1) coincides with the standard notion and KL framework used in nonsmooth optimization. Our goal is now to show that the largest KL exponent of (1.1) is at least $\frac{1}{4}$ under suitable conditions.

Throughout this section, we assume that $z \in \mathbb{C}S^{n-1}$ is a stationary point of problem (1.1). Furthermore, $y \in \mathbb{C}S^{n-1}$ denotes a neighboring point of $z$ and we set $\Delta = y - z$. We now collect and present some preparatory notations and computational results that will be used in the following derivations. Since $z$ is a stationary point of problem (1.1), we have

$$Hz = Az + 2\beta \text{diag}(|z|^2)z - 2\lambda z = 0, \quad 2\lambda = z^*Az + 2\beta \|z\|^4_4,$$

(6.3)
where $H := A + 2\beta \text{diag}(|z|^2) - 2\lambda I$, see, e.g., (2.6). Moreover, as proved in (2.7), it holds that

$$f(y) - f(z) = \frac{1}{2}y^*Hy + \frac{\beta}{2}\|\tau\|^2, \quad \tau_k := |y_k|^2 - |z_k|^2, \quad k \in [n]. \quad (6.4)$$

Due to (2.1) and (2.2), the norm of the Riemannian gradient can be expressed as follows

$$\|\text{grad} f(y)\|^2 = \frac{1}{2}\|P_y^\perp[A + 2\beta \text{diag}(|y|^2)]y\|^2 = \frac{1}{2}\|P_y^\perp[H + 2\beta \text{diag}(\tau)]y\|^2$$

$$= \frac{1}{2}\|[H + 2\beta \text{diag}(\tau)]\|y\|^2 - \frac{1}{2}(y^*[H + 2\beta \text{diag}(\tau)]y)^2$$

$$= \frac{1}{2}y^*H^2y - \frac{1}{2}(y^*Hy)^2 + 2\beta^2y^*\text{diag}(|\tau|^2)y - 2\beta^2(y^*\text{diag}(\tau)y)^2$$

$$+ 2\beta y^*H\text{diag}(\tau)y - 2\beta(y^*Hy)(y^*\text{diag}(\tau)y), \quad (6.5)$$

where $P_y^\perp = I_n - yy^*$ is the orthogonal projection onto the space $[\text{span} \{ y \}]^\perp$.

In the following, we will work with the polar representations $z_k = r_k e^{i\theta_k}$, $y_k = t_k e^{i\phi_k}$ for $r_k$, $t_k \geq 0$, $\theta_k, \phi_k \in [0, 2\pi]$, and all $k \in [n]$. It then holds that $|y_k|^2 = t_k^2$, $|z_k|^2 = r_k^2$, and $t_k = t_k^2 - r_k^2$ for all $k$. We further introduce the index sets

$$A := \{ k : z_k = 0 \}, \quad I := \{ k : z_k \neq 0 \} \quad (6.6)$$

and $r_+ := \min_{k \in I} r_k > 0$. Notice that we have $\tau_k = t_k^2 \geq 0$ for all $k \in A$.

We first show that the Łojasiewicz inequality holds with $\theta = \frac{1}{3}$ at those stationary points where $H = 0$ (we also refer to the remark after this lemma).

**Lemma 6.1** Suppose $z$ is an arbitrary point in $\mathbb{CS}^{n-1}$. Then, the inequality

$$\|\tau\|^{2(1-\theta)} \leq \eta_z \|P_y^\perp\text{diag}(\tau)y\| \quad \forall y \in B(z, \delta_z) \cap \mathbb{CS}^{n-1} \quad (6.7)$$

holds with exponent $\theta = \frac{1}{3}$ and for some constants $\eta_z, \delta_z > 0$. Here, $\tau$ is defined as in (6.4).

**Proof** In the case $\tau = 0$, we have $\|\tau\| = \|P_y^\perp\text{diag}(\tau)y\| = 0$ and consequently, the inequality (6.7) holds trivially for all $\theta \in [0, 1)$. Let us first assume $A \neq \emptyset$. Recalling $\Delta = y - z$, a straightforward calculation yields $|\Delta_k|^2 = t_k^2 - 2r_k t_k \cos(\theta_k - \phi_k) + r_k^2 \geq (t_k - r_k)^2$ and hence, it follows

$$|\tau_k| = (t_k + r_k)|t_k - r_k| \leq (2r_k + |\Delta_k|)|\Delta_k|, \quad \forall k \in I, \quad (6.8)$$
and $\tau_k = t_k^2 = |\Delta_k|^2$ for all $k \in A$. Using $\sum_{k \in I} \tau_k = -\sum_{k \in A} \tau_k = -\|	au_A\|_1$, the estimates $\|	au_A\|_2^2 \leq |A|\|	au_A\|_2^2$ and

$$\frac{1}{2} \sum_{k, j \in I} (\tau_k - \tau_j)^2 = |I| \sum_{k \in I} \tau_k^2 - \left[ \sum_{k \in I} \tau_k \right]^2$$

$$= |I| \|	au_I\|^2 - \left[ \sum_{k \in A} \tau_k \right]^2 = |I| \|	au_I\|^2 - \|	au_A\|_1^2, \quad (6.9)$$

and setting $m := |I| \geq 1$, we obtain

$$\|	au\|^2 \leq \frac{1}{2m} \sum_{k, j \in I} (\tau_k - \tau_j)^2 + \frac{n}{m} \|	au_A\|^2$$

$$\leq \frac{n}{m} \left[ \sum_{k, j \in I} |\tau_k - \tau_j|^2 + \|	au_A\|^2 \right]. \quad (6.10)$$

Furthermore, it holds that

$$2\|P^- y^* \text{diag}(\tau) y\|^2 = 2y^* \text{diag}(|\tau|^2) y - 2(y^* \text{diag}(\tau) y)^2$$

$$= 2 \left[ \sum_{k, j \in I} t_k^2 t_j^2 - \left( \sum_{k \in I} t_k^2 \tau_k \right)^2 \right] = \sum_{k, j = 1}^n t_k^2 t_j^2 (\tau_k - \tau_j)^2$$

$$= \sum_{k, j \in A} \tau_k \tau_j (\tau_k - \tau_j)^2 + \sum_{k, j \in I} t_k^2 t_j^2 (\tau_k - \tau_j)^2$$

$$+ 2 \sum_{k \in A} \sum_{j \in I} \tau_k t_j^2 (\tau_k - \tau_j)^2, \quad (6.11)$$

where the identity in the second line is a direct consequence of the derivation in (5.3).

Next, defining $\delta_z := r^2_+ / 6 < 1$ and using (6.8), we have

$$\|	au_I\|_\infty \leq \max_{k \in I} (2r_k + |y_k - z_k| |y_k - z_k|) \leq 3\delta_z \leq r^2_+ / 2. \quad (6.12)$$

for all $y \in B(z, \delta_z)$. Moreover, the last condition implies $|y_k|^2 = t_k^2 \geq r^2_+ / 2$ for all $k \in I$ and thus, using $\|	au_I\|^2 = \|	au\|^2 - \|	au_A\|^2$, it holds that

$$2 \sum_{k \in A} \sum_{j \in I} \tau_k t_j^2 (\tau_k - \tau_j)^2 \geq r^2_+ \sum_{k \in A} \sum_{j \in I} (t_k^3 - 2t_k^2 \tau_j + \tau_k \tau_j^2)$$

$$= r^2_+ \sum_{k \in A} \left[ m \tau_k^3 - 2\tau_k^2 \left( \sum_{j \in I} \tau_j \right) + \tau_k \|	au_I\|^2 \right]$$
\[
= r_+^2 \sum_{k \in \mathcal{A}} \left[ m \tau_k^3 + 2 \tau_k^2 \left( \sum_{j \in \mathcal{A}} \tau_j \right) + \tau_k(\|\tau\|^2 - \|\tau_\mathcal{A}\|^2) \right] \\
= r_+^2 m \|\tau_\mathcal{A}\|_3^3 + r_+^2 \|\tau_\mathcal{A}\|_1 \|\tau\|^2 + \|\tau_\mathcal{A}\|^2 \geq r_+^2 m \|\tau_\mathcal{A}\|_3^3.
\]  

(6.13)

In addition, since \( \tau_k \geq 0 \) for all \( k \in \mathcal{A} \), we can infer \( \sum_{k,j \in \mathcal{A}} \tau_k \tau_j (\tau_k - \tau_j)^2 \geq 0 \). Due to \( \|\tau\|_\infty \leq 1/2 \), we have \( |\tau_k - \tau_j| \leq 1 \) for all \( k, j \in \mathcal{I} \) and consequently, it follows \( \sum_{k,j \in \mathcal{I}} (\tau_k - \tau_j)^2 \geq \sum_{k,j \in \mathcal{I}} |\tau_k - \tau_j|^3 \). Together and using the estimates \( \|w\|^3 \leq \sqrt{\rho} \|w\|^3 \), \( w \in \mathbb{C}^p \), and (6.10), we finally obtain

\[
\| P_\mathcal{I} \text{diag}(\tau) y \|^2 \geq \frac{r_+^2}{2} \min \left\{ \frac{r_+^2}{4}, m \right\} \left[ \sum_{k,j \in \mathcal{I}} |\tau_k - \tau_j|^3 + \|\tau_\mathcal{I}\|^3 \right] \\
\geq \frac{r_+^2}{2\sqrt{m^2 + n - m}} \frac{\min \left\{ r_+^2/4, m \right\}}{m} \left[ \sum_{k,j \in \mathcal{I}} |\tau_k - \tau_j|^2 + \|\tau_\mathcal{I}\|^2 \right]^{3/2} \\
\geq c (m/n)^{3/2} \|\tau\|^3.
\]

Thus, the Łojasiewicz-type inequality (6.7) holds with \( \theta = \frac{1}{4} \) and \( \eta_\mathcal{I} = (\frac{n}{m})^{1.5} c^{-0.5} \).

We then consider the case when \( \mathcal{A} = \emptyset \). Similar to the inequality (6.12), defining \( \delta_\mathcal{I} = r_+^2/6 \) leads to \( r_+^2 \geq r_+^2/2 \) for all \( k \in [n] \). Combining this with equality (6.11), we have

\[
2 \| P_\mathcal{I} \text{diag}(\tau) y \|^2 = \sum_{k,j \in [n]} \delta_\mathcal{I} \tau_k^2 \tau_j^2 (\tau_k - \tau_j)^2 \geq \frac{r_+^4}{4} \sum_{k,j \in [n]} (\tau_k - \tau_j)^2.
\]

Then, utilizing (6.10), we immediately obtain

\[
\|\tau\|^2 \leq \frac{1}{2m} \sum_{k,j \in \mathcal{I}} (\tau_k - \tau_j)^2 \leq \frac{4}{mr_+^4} \| P_\mathcal{I} \text{diag}(\tau) y \|^2.
\]

Hence, the Łojasiewicz-type inequality (6.7) holds with \( \theta = \frac{1}{2} \) and \( \eta_\mathcal{I} = 2m^{-1/2} r_+^{-2} \).

\[\square\]

**Remark 5** If \( z \in \mathbb{C}S^{n-1} \) is a stationary point of problem (1.1) with \( H = 0 \), then we have \( f(y) - f(z) = (\beta/2)\|\tau\|^2 \) and \( \|\nabla f(z)\| = \sqrt{2}\beta\| P_\mathcal{I} \text{diag}(\tau) y \| \) and thus, Lemma 6.1 implies that the Łojasiewicz inequality (6.1) holds with \( \theta = \frac{1}{4} \). Moreover, this result can be used to show that the KL exponent can not be larger than \( \frac{1}{4} \) for general stationary points.

**Remark 6** More generally, using the same analysis as in the proof of Lemma 6.1, it can be shown that the Łojasiewicz-type inequality (6.7) holds with exponent \( \theta = \frac{1}{2} \) along directions \( y \in B(z, \delta_\mathcal{I}) \cap \mathbb{C}S^{n-1} \) with \( y_\mathcal{A} = 0 \).

\[\square\] Springer
Next, we prove that the largest KL exponent is at least $\frac{1}{4}$ when $A$ is diagonal.

**Theorem 6.1** Let $A = \text{diag}(a) \in \mathbb{C}^{n \times n}$, $a \in \mathbb{R}^n$, be a diagonal matrix. Then, the largest KL exponent of problem (1.1) is at least $\frac{1}{4}$.

**Proof** In the case $\tau = 0$, using $H\Delta = Hy$, we have $f(y) - f(z) = \frac{1}{2}\Delta^*H\Delta$ and $2\|\text{grad } f(y)\|^2 = \Delta^*H^2\Delta - (\Delta^*H\Delta)^2$. Here, we set $\Delta = y - z$ as usual. We now decompose $\Delta$ as follows

$$\Delta = u + v, \quad Hu = 0, \quad \|Hv\| \geq \sigma_-(H)\|v\|, \quad \text{and } u^*v = 0,$$

where $\sigma_-(H)$ denotes smallest positive singular value of $H$. Moreover, let $\sigma_+(H) \geq 0$ be the maximum singular value of $H$. It holds that

$$f(y) - f(z) \leq \frac{\sigma_+(H)}{2}\|v\|^2, \quad \|\text{grad } f(y)\|^2 \geq \frac{\sigma_-(H)^2}{2}\|v\|^2 - \frac{\sigma_+(H)^2}{2}\|v\|^4.$$

Thus, due to $\|v\|^2 \leq \|\Delta\|^2$, the inequality (6.1) is satisfied with exponent $\theta = \frac{1}{4}$.

Next, we consider the general case $\tau \neq 0$. In this situation, we have $A_{[A]} = 0$, $H_{[A]} = A_{[A]} = 0$ and the stationarity conditions imply

$$a_k + 2\beta|z_k|^2 - 2\lambda = 0 \quad \forall k \in \mathcal{I} \implies H_{[\mathcal{I}]} = 0. \quad (6.14)$$

Hence, due to $\tau_k = |\Delta_k|^2 = |y_k|^2 = t_k^2$ for all $k \in A$ and (6.14), it follows

$$y^*\text{diag}(\tau)Hy = \sum_{k \in [n]} (a_k + 2\beta|z_k|^2 - 2\lambda)\tau_k|y_k|^2 = \sum_{k \in A} (a_k - 2\lambda)t_k^2, \quad (6.15)$$

and $y^*H^\ell y = \sum_{k \in A} (a_k - 2\lambda)\tau_k|y_k|^2$ for $\ell \in \mathbb{N}$. Utilizing Young’s inequality, $t_k^2 \leq 1$, and $\tau_k = t_k^2$ for all $k \in A$, we obtain

$$|y^*Hy \cdot y^*\text{diag}(\tau)y| = |y^*Hy| \left\| \tau_{[A]} \right\|^2 + \sum_{\ell \in \mathcal{I}} \left| t_{[\ell]} \tau_k \right| \leq |y^*Hy| \left[ \left\| \tau_{[A]} \right\|^2 + \left\| \tau_{[\mathcal{I}]} \right\|_1 \right]\leq \left[ \frac{|y^*Hy|q^p}{q} + \frac{\|\tau_{[A]}\|^{2p}}{p} \right] + \left[ \frac{|y^*Hy|^q}{q} + \frac{\|\tau_{[\mathcal{I}]}\|_1^p}{p} \right]$$

$$= \frac{2|y^*Hy|^q}{q} + \frac{\|\tau_{[A]}\|^{2p} + \|\tau_{[\mathcal{I}]}\|_1^p}{p} \quad (6.16)$$

for some $p > 1$ and $q = 1 + \frac{1}{p-1}$.

Let us now introduce the index set $B := \{k \in A : a_k - 2\lambda \neq 0\}$ and let us define $h_- := \min_{k \in B} |a_k - 2\lambda|$ and $h_+ := \max_{k \in B} |a_k - 2\lambda|$. Then, due to (6.15) and using the representation of $y^*H^\ell y$, $\ell \in \mathbb{N}$, it follows

$$y^*\text{diag}(\tau)Hy \geq -h_+ \sum_{k \in B} |y_k|^2 \tau_k = -h_+\|y_B\|_4^4,$$
\( y^* H^2 y \geq h_+^2 \|y_B\|^2 \), and \(|y^* H y| \leq h_+ \|y_B\|^2 \). Combining the latter inequalities with (6.16) and applying (6.5), we can infer

\[
\|\text{grad} \ f(y)\|^2 \\
= \frac{1}{2} y^* H^2 y - \frac{1}{2} (y^* H y)^2 + 2\beta^2 y^* \text{diag}(|\tau|^2) y - 2\beta^2 (y^* \text{diag}(\tau) y)^2 \\
+ 2\beta y^* H \text{diag}(\tau) y - 2\beta (y^* H y) (y^* \text{diag}(\tau) y) \\
\geq \frac{h_+^2}{2} \|y_B\|^2 - h_+^2 \|y_B\|^4 + 2\beta^2 y^* \text{diag}(|\tau|^2) y - 2\beta^2 (y^* \text{diag}(\tau) y)^2 \\
- 2\beta h_+ \|y_B\|^4 - 4\beta h_+^q \|y_B\|^2 - \frac{2\beta}{p} [\|A\|^{2p} + \|\tau_I\|_p^p] \\
= \frac{h_+^2}{2} \|y_B\|^2 - \alpha(\|y_B\|^2) + 2\beta^2 y^* \text{diag}(|\tau|^2) y - 2\beta^2 (y^* \text{diag}(\tau) y)^2 \\
- \frac{2\beta}{p} [\|A\|^{2p} + \|\tau_I\|_p^p] \\
\geq \frac{h_+^2}{2} \|y_B\|^2 - \alpha(\|y_B\|^2) + \frac{\beta^2 r_+^4}{4} \sum_{k, j \in I} |\tau_k - \tau_j|^2 + \beta^2 r_+^2 m \|A\|^3 \\
- 2\beta p^{-1} [\|A\|^{2p} + \|\tau_I\|_p^p] \\
\geq h_+^2 \|y_B\|^2 - \alpha(\|y_B\|^2) + \frac{\beta^2 r_+^4}{4} \sum_{k, j \in I} |\tau_k - \tau_j|^2 + \beta^2 r_+^2 m \|A\|^3 \\
- 2\beta p^{-1} [\|A\|^{2p} + \|\tau_I\|_p^p] \\
\geq h_+^2 \|y_B\|^2 - \alpha(\|y_B\|^2) + \frac{\beta^2 r_+^4}{4} \sum_{k, j \in I} |\tau_k - \tau_j|^2 + \beta^2 r_+^2 m \|A\|^3 \\
- 2\beta p^{-1} [\|A\|^{2p} + \|\tau_I\|_p^p] \\
\geq h_+^2 \|y_B\|^2 - \alpha(\|y_B\|^2) + \frac{\beta^2 r_+^4}{4} \sum_{k, j \in I} |\tau_k - \tau_j|^2 + \beta^2 r_+^2 m \|A\|^3 \\
- 2\beta p^{-1} [\|A\|^{2p} + \|\tau_I\|_p^p]
\]

for \( \|y_B\| \to 0 \) and \( m := |I| \). Assuming that \( y \) is sufficiently close to \( z \), then the last step in the latter inequality is a consequence of the estimates (6.11) and (6.13) in Lemma 6.1. Next, by (6.9), we have

\[
\|\tau_I\|_1^4 \leq m^2 \|\tau_I\|^4 \leq \left[ \frac{1}{2} \sum_{k, j \in I} |\tau_k - \tau_j|^2 + |A| \|A\|^2 \right]^2 \\
\leq \left[ \sum_{k, j \in I} |\tau_k - \tau_j|^2 \right]^2 + 2(n - m)^2 \|A\|^4 \\
\]

Setting \( p = 4 \) and using (6.10), we can choose \( \|\Delta\| \) (and thus \( \|\tau\| \) and \( \|y_B\| \)) sufficiently small, such that

\[
\|\text{grad} \ f(y)\|^2 \geq \eta_1 \left[ \|y_B\|^2 + \sum_{k, j \in I} |\tau_k - \tau_j|^2 + \|A\|^3 \right]^{3/2} \\
\geq \eta_2 \left[ \|y_B\|^2 + \sum_{k, j \in I} |\tau_k - \tau_j|^2 + \|A\|^2 \right]^{3/2} \\
\]
and \( f(y) - f(z) \leq \eta_3 \|y_B\|^2 + \sum_{k,j \in I} |\tau_k - \tau_j|^2 + \|\tau_A\|^2 \) for suitable \( \eta_1, \eta_2, \eta_3 > 0 \).

This shows that the Łojasiewicz inequality is satisfied with \( \theta = \frac{1}{4} \). \( \Box \)

Finally, we derive a lower bound of the largest KL exponent in the real case for global minimizers characterized by the positive semidefiniteness condition in Theorem 2.1.

**Theorem 6.2** Suppose \( A \in \mathbb{R}^{n \times n} \) is a symmetric matrix and \( z \) is a stationary point of problem (1.1) satisfying

\[
H \succeq 0,
\]

where \( H \) is defined in (2.6). Then, the largest KL exponent of (1.1) at \( z \) is at least \( \frac{1}{4} \).

**Proof** The proof of Theorem 6.2 is provided in Appendix A. \( \Box \)

### 7 Conclusions

In this paper, we analyze the geometric properties of a class of quartic–quadratic optimization problems under a single spherical constraint. When the matrix \( A \) in the quadratic form is diagonal, the stationary points and local minima can be fully characterized and we show that the minimization problem does not possess any spurious local minima. Furthermore, a closed-form expression for global minimizer is available which is based on the projection onto the \( n \)-simplex. If \( A \) is a rank-one matrix, a similar analysis can be performed and we derive characteristic properties of associated local minima and uniqueness of global minima up to a certain phase shift. We verify that the problem satisfies a Riemannian-type strict-saddle property in the real space when the interaction coefficient is at least of order \( O(n^{3/2}) \) or sufficiently small which corresponds to the case where either the quartic or the quadratic part is the leading term of the objective function. Finally, we estimate the largest Kurdyka–Łojasiewicz exponent \( \theta \) of problem (1.1) and show that \( \theta \) is at least \( \frac{1}{4} \) for all stationary points \( z \) if \( A \) is diagonal or if the problem is restricted to the real space and \( z \) fulfills a certain global optimality condition.

**Acknowledgements** The authors are grateful to Prof. Adrian Lewis, the Associate Editor and the anonymous reviewers for their valuable comments and suggestions that helped to improve the quality of our manuscript.

### A Proof of Theorem 6.2

**Proof** As in the proof of Theorem 6.1, the verification of Theorem 6.2 is mainly based on proper decompositions of \( \Delta = y - z \) and \( \text{diag}(\tau)y \) that allow to derive appropriate bounds of \( |f(y) - f(z)| \) and \( \|\nabla f(y)\| \) if \( y \) is close to \( z \) or if, equivalently, \( \Delta \) is sufficiently small. In particular, we will discuss three cases that will allow to simplify
and estimate the expressions for $|f(y) - f(z)|$ and $\|\nabla f(y)\|$ via identifying the different leading terms.

Without loss of generality we assume $\beta = 1$. Let $y \in \mathbb{S}^{n-1}$ be arbitrary and let us set $\Delta = y - z$, and

$$
\gamma_1 = \sum_{k \in I} z_k^3 \Delta_k, \quad \gamma_2 = \sum_{k \in I} z_k^2 \Delta_k^2, \quad \gamma_3 = \sum_{k \in I} z_k \Delta_k^3, \quad \gamma_4 = \|\Delta\|^4. \quad (A.1)
$$

Based on the representation $\nabla f(y) = P_y^+[H + 2\text{diag}(\tau)]y$, we now introduce the following decompositions

$$
2\text{diag}(\tau)y = w + c_1 y, \quad w = 2P_y^+ \text{diag}(\tau)y, \quad c_1 = 2y^T \text{diag}(\tau)y
$$

$$
\Delta = u + v, \quad H u = 0, \quad u^T v = 0, \quad \|Hv\| \geq \sigma_-(H)\|v\|, \quad (A.2)
$$

where $\sigma_-(H)$ denotes the smallest positive singular value of $H$. Using this decomposition, (6.5), and $Hy = H\Delta$, we can express the norm of the Riemannian gradient as follows

$$
\|\nabla f(y)\|^2 = \|P_y^+[H + 2\text{diag}(\tau)]y\|^2 = \|P_y^+[Hy + w + c_1 y]\|^2
$$
$$
= \|P_y^+[H\Delta + w]\|^2 = \|H\Delta + w\|^2 - (y^T H\Delta + y^T w)^2
$$
$$
= \|H\Delta + w\|^2 - (\Delta^T H \Delta)^2 = \|Hv + w\|^2 - (v^T H v)^2. \quad (A.3)
$$

Let $\lambda_+(H)$ be the largest eigenvalue of $H$. Then, by definition of $v$, we obtain

$$
|v^T H v| \leq \lambda_+(H)\|v\|^2 \leq \lambda_+(H)\sigma_-(H)^{-2}\|Hv\|^2 =: \tilde{\sigma}^{-1} \cdot v^T H^2 v. \quad (A.4)
$$

Moreover, Lemma 6.1 yields

$$
\|\tau\|^3 \leq \eta_1\|w\| \quad (A.5)
$$

for some constant $\eta_1 > 0$ and for all $y \in \mathbb{S}^{n-1}$ sufficiently close to $z$. Throughout the proof, we will also repeatedly use the following facts:

$$
2y^T \Delta = \|y\|^2 + \|\Delta\|^2 - \|y - \Delta\|^2 = \|\Delta\|^2, \quad (A.6)
$$

$$
z^T \Delta = y^T \Delta - \|\Delta\|^2 = -\frac{1}{2}\|\Delta\|^2.
$$

Let $\epsilon \in (0, 1]$ be an arbitrary, small positive constant. We now discuss three different cases.

Case 1. $\|w\| \geq (1 + \epsilon)\|Hv\|$ or $\|w\| \leq (1 - \epsilon)\|Hv\|$. We first assume $w \neq 0$ and $\epsilon < 1$. Since the function $x \mapsto \varphi(x) := x + 1/x$ is monotonically decreasing on the interval $(0, 1 - \epsilon]$ and monotonically increasing on $[1 + \epsilon, \infty)$, it follows

$$
\varphi(x) \geq 1 - \epsilon + \frac{1}{1 - \epsilon} \quad x \in (0, 1 - \epsilon] \quad \text{and} \quad \varphi(x) \geq 1 + \epsilon + \frac{1}{1 + \epsilon} \quad x \geq 1 + \epsilon
$$
Thus, we obtain
\[
\frac{\|w\|}{\|Hv\|} + \frac{\|Hv\|}{\|w\|} \geq \min \left\{ 1 - \epsilon + \frac{1}{1 - \epsilon}, 1 + \epsilon + \frac{1}{1 + \epsilon} \right\} = \frac{(1 + \epsilon)^2 + 1}{1 + \epsilon},
\]
which further implies
\[
\|Hv + w\|^2 \geq \|w\|^2 + \|Hv\|^2 - 2\|w\|\|Hv\| = \left[ 1 - 2 \left( \frac{\|w\|}{\|Hv\|} + \frac{\|Hv\|}{\|w\|} \right)^{-1} \right] (\|w\|^2 + \|Hv\|^2) \geq \frac{\epsilon^2}{2} (\|w\|^2 + \|Hv\|^2).
\]

Next, we choose \(\delta_z > 0\) sufficiently small such that \(\|v\|^2 \leq \epsilon^2 \tilde{\sigma} / (10\lambda_+(H))\) and \(|v^THv| \leq 1\). (This is possible due to the decomposition \((A.2)\) and \(|v| \leq |\Delta|\)). Using \((A.4), (A.5)\), the estimate \([\frac{1}{2}|a + b|^3/2 \leq |a|^{3/2} + |b|^{3/2}] / \sqrt{2}, a, b \in \mathbb{R}\), and setting \(\tilde{\eta}_1 := \min\{\eta_{1,1}^2, \frac{\delta}{2}\}\), it follows
\[
\|\text{grad } f(y)\|^2 \geq \frac{\epsilon^2}{5} (\|w\|^2 + \|Hv\|^2) - \lambda_+(H)\|v\|^2|v^THv|
\]
\[
\geq \frac{\epsilon^2}{10} (2\|w\|^2 + 2\|Hv\|^2) - \lambda_+(H) \cdot \epsilon^2 \tilde{\sigma} / (10\lambda_+(H)) \cdot \tilde{\sigma}^{-1} \|Hv\|^2
\]
\[
= \frac{\epsilon^2}{10} \left( 2\|w\|^2 + \|Hv\|^2 \right) \geq \frac{\epsilon^2}{5} \min \left\{ \frac{1}{\tilde{\eta}_1^2}, \frac{\tilde{\sigma}}{2} \right\} \left[ \|\tau\|^3 + |v^THv| \right]
\]
\[
\geq \frac{\epsilon^2\tilde{\eta}_1}{5} \left[ \|\tau\|^3 + |v^THv| \right] \geq \frac{\epsilon^2\tilde{\eta}_1}{10} \left[ \|\tau\|^2 + v^THv \right] \frac{3}{2}
\]
\[
= \frac{\epsilon^2\tilde{\eta}_1 \sqrt{2}}{5} |f(y) - f(z)|^2,
\]
where the last equality is a consequence of \((6.4)\). Thus, we can infer that the largest KL exponent of problem \((1.1)\) at \(z\) is at least \(\frac{1}{4}\).

**Case 2.** \((2 - \epsilon)r_2^2 \|\Delta\| \geq \|\Delta\|^2\) or \(\gamma_1 \leq (2 - \epsilon)r_2^2 \|\Delta\|^2 \|\Delta\|^{-2}\). First, utilizing the identity \(\Delta^T y = \frac{1}{2} \|\Delta\|^2\) in \((A.6)\), we have \(P_y^\perp \Delta = \Delta - \frac{1}{2} \|\Delta\|^2 y\). Defining \(t_\Delta := \Delta^T P_y^\perp \Delta = \|\Delta\|^2 - \frac{1}{4} \|\Delta\|^4\), we will now work with the following additional decompositions
\[
H\Delta = \Delta^T H\Delta \cdot y + c_2 P_y^\perp \Delta + w_1, \quad c_2 = \frac{1 - \frac{1}{2} \|\Delta\|^2}{t_\Delta} \Delta^T H\Delta,
\]
\[
w = c_3 P_y^\perp \Delta + w_2, \quad c_3 = \frac{1}{t_\Delta} \left[ 2\Delta^T \text{diag}(\tau) y - \|\Delta\|^2 y^T \text{diag}(\tau) y \right].
\]
We note that due to the choice of \( c_2 \) and \( c_3 \) and (A.2), the vectors \( w_1 \) and \( w_2 \) are orthogonal to \( y \) and \( \Delta \). Hence, by (A.3), it holds that

\[
\| \text{grad} \ f(y) \|^2 = \| H \Delta \|^2 + 2 w^T H \Delta + \| w \|^2 - (\Delta^T H \Delta)^2 \\
= (c_2^2 + 2 c_2 c_3 + c_3^2) \| P_y \Delta \|^2 + \| w_1 \|^2 + 2 w_1^T w_2 + \| w_2 \|^2 \\
= (c_2 + c_3)^2 \cdot t_\Delta + \| w_1 + w_2 \|^2 \geq (c_2 + c_3)^2 \cdot t_\Delta.
\]

Recalling the definitions introduced in (A.1) and using \( \tau_k = (y_k - z_k)(y_k + z_k) = \Delta_k(\Delta_k + 2z_k) \) and \( y_k = \Delta_k + z_k \), we can express \( t_\Delta \cdot c_3 \) via

\[
t_\Delta c_3 = 2 \sum_{k \in [n]} \Delta_k^2(\Delta_k + 2z_k)(\Delta_k + z_k) \\
- \| \Delta \|^2 \sum_{k \in [n]} \Delta_k(\Delta_k + 2z_k)(z_k + \Delta_k)^2 \\
= \sum_{k \in [n]} 2[\Delta_k^4 + 3 \Delta_k^3 z_k + 2 \Delta_k^2 z_k^2] \\
- \| \Delta \|^2 [\Delta_k^4 + 4 \Delta_k^3 z_k + 5 \Delta_k^2 z_k^2 + 2 \Delta_k z_k^3] \\
= (2 - \| \Delta \|^2) \gamma_4 + (6 - 4 \| \Delta \|^2) \gamma_3 + (4 - 5 \| \Delta \|^2) \gamma_2 - 2 \| \Delta \|^2 \gamma_1. \quad (A.7)
\]

Notice that we have \( \gamma_2 \geq r_+^2 \| \Delta_T \|^2 \) and \( \gamma_1 \leq 1 \) provided that \( \| \Delta \| \leq 1 \). Now, if \((2 - \epsilon)r_+^2 \| \Delta_T \| \geq \| \Delta \|^2 \), we obtain

\[
(4 - 5 \| \Delta \|^2) \gamma_2 - 2 \| \Delta \|^2 \gamma_1 \geq (4 - 5 \| \Delta \|^2)r_+^2 \| \Delta_T \|^2 - 2 \| \Delta \|^2 \\
\geq (2 - 5 \| \Delta \|^2)r_+^2 \| \Delta_T \|^2.
\]

and hence, it follows

\[
t_\Delta \cdot c_3 \geq \gamma_4 + \epsilon r_+^2 \| \Delta_T \|^2 + o(\| \Delta_T \|^2) \geq \| \Delta \|^2 + \frac{\epsilon r_+^2}{2} \| \Delta_T \|^2 \quad (A.8)
\]

for \( \Delta \) sufficiently small. Otherwise, if \( \gamma_1 \geq (2 - \epsilon)r_+^2 \| \Delta_T \|^2 \| \Delta \|^{-2} \), then we also have \((4 - 5 \| \Delta \|^2) \gamma_2 - 2 \| \Delta \|^2 \gamma_1 \geq (2 - 5 \| \Delta \|^2)r_+^2 \| \Delta_T \|^2 \) and thus, (A.8) holds in both sub-cases. Consequently, due to the positive semidefiniteness of \( H \), (A.8), and

\[
\| \Delta \|^2 = -2 \Delta^T z = -2 \Delta^T \varepsilon_T z \leq 2 \| \Delta_T \| \| z_T \| \leq 2 \| \Delta_T \|, \quad (A.9)
\]

we can infer

\[
\| \text{grad} \ f(y) \| \geq |c_2 + c_3| \sqrt{t_\Delta} = \left[ \left( 1 - \frac{1}{2} \| \Delta \|^2 \right) \Delta^T H \Delta + t_\Delta c_3 \right]^{1/2} \leq \left[ \Delta^T H \Delta + 2 \| \Delta \|^4 + \epsilon r_+^2 \| \Delta_T \|^2 \right] \cdot (2 \| \Delta \|)^{-1}
\]
\[ \begin{align*}
\geq & \left[ \Delta^T H \Delta + 2\|\Delta\|^4 + \epsilon r_+^2 \|\Delta z\|^2 \right]^{\frac{3}{2}} \cdot \frac{\epsilon^{\frac{1}{4}} \sqrt{r_+} \|\Delta z\|}{2\|\Delta\|} \\
\geq & \eta_2 [\Delta^T H \Delta + 2\|\Delta\|^4 + \epsilon r_+^2 \|\Delta z\|^2]^{\frac{3}{2}},
\end{align*} \]

where \( \eta_2 := \epsilon^{\frac{1}{4}} \sqrt{r_+} / 2\sqrt{2} \) and if \( \Delta \) is chosen sufficiently small. Here, we also used the estimates \( t_\Delta = \|\Delta\|^2 - \|\Delta\|^4 / 4 \leq \|\Delta\|^2 \) and \( 1 - \|\Delta\|^2 / 2 \geq 1 / 2 \) in the second inequality. The third inequality follows from \( [\Delta^T H \Delta + 2\|\Delta\|^4 + \epsilon r_+^2 \|\Delta z\|^2]^{\frac{3}{2}} \geq \epsilon^{\frac{1}{4}} \sqrt{r_+} \|\Delta z\| \). Next, utilizing (6.4) and \( |\gamma_k| \leq 2|\Delta_k| \) for all \( k \in \mathcal{I} \), we finally obtain

\[ |f(y) - f(z)| = \frac{1}{2} [\Delta^T H \Delta + \|\tau\|^2] \]

\[ \leq \frac{1}{2} \left[ \Delta^T H \Delta + \|\Delta \|_{4}^{4} + 4\|\Delta z\|^2 \right] \leq \eta_3 \|\text{grad} f(y)\|^\frac{3}{2} \]

for some constant \( \eta_3 > 0 \) and for all \( y \) sufficiently close to \( z \). Hence, the largest KL exponent is at least \( \frac{1}{4} \) in this case.

**Case 3.** \( (1 - \epsilon)\|Hv\| \leq \|w\| \leq (1 + \epsilon)\|Hv\|, \) \( \gamma_1 \geq (2 - \epsilon)r_+^2 \|\Delta z\|^2 \|\Delta\|^{-2} \) and \( (2 - \epsilon)r_+^2 \|\Delta z\| \leq \|\Delta\|^2 \). In this case, the inequality (A.9) implies \( \|\Delta z\| = \Theta (\|\Delta\|^2) \) and setting \( v = [(2 - \epsilon)r_+^2]^{-1} \), the terms \( \gamma_i \) for \( i = 1, \ldots, 3 \) can be estimated as follows

\[ (4\nu)^{-1} \|\Delta\|^2 \leq (2 - \epsilon)r_+^2 \|\Delta z\|^2 \|\Delta\|^{-2} \leq \gamma_1 \leq \|\Delta z\|_1 \leq |\mathcal{I}|v\|\Delta\|^2, \]

\[ \frac{r_+^2}{4} \|\Delta\|^4 \leq r_+^2 \|\Delta z\|^2 \leq \gamma_2 \leq \|\Delta z\|^2 \leq v^2 \|\Delta\|^4, \]

\[ -v^3 \|\Delta\|^6 \leq -\|\Delta z\|_3 \leq \gamma_3 \leq \|\Delta z\|^3 \leq v^3 \|\Delta\|^6. \] (A.10)

Together with \( \gamma_4 = \|\Delta\|^4 \), this shows that

\[ \gamma_1 = \Theta (\|\Delta\|^2), \quad \gamma_2 = \Theta (\|\Delta\|^4), \quad \gamma_3 = O (\|\Delta\|^6), \quad \gamma_4 = \Theta (\|\Delta\|^4) \] (A.11)

for all \( \Delta \). Let us set \( m := |\mathcal{I}| \) and define \( \sigma_k = \Delta_k z_k + \frac{1}{2m} \|\Delta\|^2 \) for all \( k \in \mathcal{I} \) and \( \sigma_k = 0 \) for all \( k \in \mathcal{A} \). Then, by (A.6), we have

\[ \sum_{k \in [n]} \sigma_k = 0, \quad \|\sigma\|_1 \geq \sum_{k \in \mathcal{I}} z_k^2 \sigma_k = \gamma_1 + \frac{1}{2m} \|\Delta\|^2, \] (A.12)

and \( \|\sigma\| = \Theta (\|\Delta\|^2) \). We now express \( \|w\|^2 \) in terms of \( \gamma_1, \gamma_2, \gamma_3, \) and \( \gamma_4 \). Specifically, by utilizing (A.2), (A.11) and by mimicking the derivation of (A.7), we obtain

\[ \frac{1}{4} \|w\|^2 = \|P_y \text{diag}(\tau)y\|^2 = y^T \text{diag}(|\tau|^2)y - (y^T \text{diag}(\tau)y)^2 \]

\[ = \sum_{k \in [n]} (z_k + \Delta_k)^2 \Delta_k^2 (\Delta_k + 2z_k)^2 - (\|\Delta\|^4 + 4\gamma_3 + 5\gamma_2 + 2\gamma_1)^2 \]

\[ \square \] Springer
On the geometric analysis of a quartic–quadratic... 467

Furthermore, by (6.4) and (A.1), it holds that

\[ \sum_{k,j \in I} 2 \Delta_k^2 \Delta_j^2 z_k^2 + \sum_{k,j \in I} 2 \Delta_k^2 \Delta_j^2 z_k^2 z_j^2 - 4 \left( \sum_{k \in I} \Delta_k^3 z_k \right)^2 + O(\|\Delta\|^6) \]

\[ = 2 \sum_{k,j \in I} \left[ \Delta_k^2 \Delta_j^2 - 2 \Delta_k^3 z_k^2 \Delta_k \Delta_j + z_k^2 \Delta_j^2 \right] + O(\|\Delta\|^6) \]

\[ = 2 \sum_{k,j \in I} z_k^2 z_j^2 (\sigma_k - \sigma_j)^2 + O(\|\Delta\|^6). \]

We notice that the higher order terms \( \sum_{k \in I} \Delta_k^3 \Delta_k^{-1} z_k^3 \Delta_k^{-1} \), \( \ell \in \{0, 1, 2\} \), can be discussed as in (A.10) and are all bounded by \( O(\|\Delta\|^6) \). Finally, applying (A.12) and \( |z_k| \leq 1 \) and \( z_k^2 \geq r_+^2 \) for all \( k \in I \), we obtain

\[ \sum_{k,j \in I} z_k^2 z_j^2 (\sigma_k - \sigma_j)^2 \geq r_+^4 \sum_{k,j \in I} (\sigma_k - \sigma_j)^2 = 2m r_+^4 \|\sigma\|^2 \]

and \( \sum_{k,j \in I} z_k^2 z_j^2 (\sigma_k - \sigma_j)^2 \leq 2m \|\sigma\|^2 \) which implies \( \|w\| = \Theta(\|\Delta\|^2) \) and

\[ \|Hv\| = \Theta(\|\Delta\|^2) \quad (A.13) \]

for \( \Delta \to 0 \). As a consequence, by (A.4), we also get

\[ |v^T H v| = O(\|\Delta\|^4). \quad (A.14) \]

Furthermore, by (6.4) and (A.1), it holds that

\[ 2|f(y) - f(z)| \leq |\Delta^T H \Delta| + \|\tau\|^2 = |v^T H v| + \sum_{k \in [n]} \left[ \Delta_k^2 + 2 z_k \Delta_k \right]^2 \]

\[ = |v^T H v| + \|\Delta\|^4 + 4 \gamma_3 + 4 \gamma_2 = O(\|\Delta\|^4) \quad (A.15) \]

for \( \Delta \to 0 \). For some index sets \( K, J \subset [n] \), let \( H_{K,J} \in \mathbb{R}^{|K| \times |J|} \) denote the submatrix \( H_{K,J} = (H_{kj})_{k \in K, j \in J} \). Notice that due to the positive semidefiniteness of \( H \), we have \( H_{II} \geq 0 \) and \( H_{AA} \geq 0 \). Moreover, due to (A.2), (A.3), and \( |v^T H v| = O(\|\Delta\|^4) \), we obtain
\[ \| \text{grad } f(y) \|^2 = \| H \Delta + w \|^2 + O(\| \Delta \|^8) \]
\[ = \| H_A \Delta + w_A \|^2 + \| H_I \Delta + w_I \|^2 + O(\| \Delta \|^8) \]
\[ = \| H_{II}^T \Delta_I + H_{AA} \Delta_A + 2\text{diag}(\| \Delta_A \|^2) \Delta_A - c_1 \Delta_A \|^2 \]
\[ + \| H_{II} \Delta_I + H_{IA} \Delta_A + w_I \|^2 + O(\| \Delta \|^8). \] (A.16)

where \( c_1 = 2y^T \text{diag}(\tau) y \) was defined in (A.2) and we used the identities \( z_k = 0 \) for \( k \in A \) and \( w_A = 2\text{diag}(\tau_A) y_A - c_1 y_A = 2\text{diag}(\| \Delta_A \|^2) \Delta_A - c_1 \Delta_A \). We set
\[ h := H_{II} \Delta_I + H_{IA} \Delta_A, \quad g_1 := h + w_I, \]
\[ g_2 := H_{II}^T \Delta_I + H_{AA} \Delta_A + 2\text{diag}(\| \Delta_A \|^2) \Delta_A - c_1 \Delta_A \] (A.17)

and, let \( \eta_4, \eta_5 > 0 \) and \( \mu \in (0, \frac{1}{2}) \) be given constants. Next, we discuss two separate sub-cases.

**Sub-case 3.1.** \( \Delta_I^T h \geq -\eta_4 \| \Delta \|^{4+\mu} \) or \( \Delta_I^T h \leq -\eta_5 \| \Delta \|^{4-\mu} \). Let us first assume \( \Delta_I^T h \geq -\eta_4 \| \Delta \|^{4+\mu} \). Following the derivation of (A.7) and using (A.6), (A.2), and (A.11), we obtain \( \frac{1}{2} c_1 = \| \Delta \|^4 + 4\gamma_3 + 5\gamma_2 + 2\gamma_1 = \Theta(\| \Delta \|^2), \Delta_{II}^T y_I = -\frac{1}{2} \| \Delta \|^2, \) and
\[ \Delta_I^T w_I = 2\Delta_{II}^T \text{diag}(\tau_I) y_I - c_1 \Delta_{II}^T y_I = 2\| \Delta_I \|^4 + 6\gamma_3 + 4\gamma_2 + \frac{1}{2} \| \Delta \|^2 c_1, \]
which yields
\[ \Delta_I^T g_1 \geq -\eta_4 \| \Delta \|^{4+\mu} + 2\| \Delta_I \|^4 + 4\gamma_2 + \| \Delta \|^2[\| \Delta \|^4 + 5\gamma_2 + 2\gamma_1] \]
\[ + 6\gamma_3 + 4\| \Delta \|^2 \gamma_3 \]
\[ \geq \Theta(\| \Delta \|^4). \]

Similarly, in the case \( \Delta_I^T h \leq -\eta_5 \| \Delta \|^{4-\mu} \) and if \( \| \Delta \| \) is sufficiently small, we get
\[ \Delta_I^T g_1 \leq -\frac{\eta_5}{2} \| \Delta \|^{4-\mu}. \]

Combining both cases, we can infer \( \| \Delta_I \| \| g_1 \| \geq \| \Delta_I^T g_1 \| \geq \eta_6 \| \Delta \|^{4} \) for some \( \eta_6 > 0 \) and for all \( y \) sufficiently close to \( b \). By the assumptions of case 3, this implies \( \| g_1 \| \geq \nu^{-1} \eta_6 \| \Delta \|^{2} \) and hence, by (A.16), we have \( \| \text{grad } f(y) \| \geq \Theta(\| \Delta \|^2) \). Considering (A.15), the Łojasiewicz inequality holds with \( \theta = \frac{1}{2} \) in this sub-case.

**Sub-case 3.2.** \(-\eta_5 \| \Delta \|^{4-\mu} \leq \Delta_I^T h \leq -\eta_4 \| \Delta \|^{4+\mu} \). Due to \( \Delta_I^T H_{II} \Delta_I \geq 0 \), this directly yields \( \Delta_I^T H_{II} \Delta_A = \Delta_I^T h - \Delta_I^T H_{II} \Delta_I < 0 \). Moreover, we can estimate that
\[ \Delta_I^T H_{II} \Delta_A = \Delta_I^T h - \Delta_I^T H_{II} \Delta_I \geq -\eta_5 \| \Delta \|^{4-\mu} - \lambda_-(H_{II}) \| \Delta_I \|^2 \]
\[ \geq -\eta_5 \| \Delta \|^{4-\mu} - \lambda_-(H_{II}) \cdot (2 - \epsilon)^{-2} r_+^{-4} \| \Delta \|^4 \geq -2\eta_5 \| \Delta \|^{4-\mu}, \]
where \( \lambda_-(H_{II}) \) is the maximal eigenvalue of \( H_{II} \). We note that the last inequality holds when \( \| \Delta \| \) is small enough, since \( \eta_5, r_+, \epsilon, H_{II} \) do not depend on \( \Delta \) and can be
considered as constants when $\Delta \to 0$. Also, we have
\[
\Delta^T_A H_{AA} \Delta_A \geq \Delta^T_A (H^T_{II} \Delta_I + H_{AA} \Delta_A) \geq \Delta^T H \Delta = v^T H v.
\] (A.18)

Utilizing the positive semidefiniteness of $H$, it holds that
\[
\|Hv\|^2 = \left(H^\frac{1}{2} v\right)^T H \left(H^\frac{1}{2} v\right) \leq \lambda_+(H) \|H^\frac{1}{2} v\| = \lambda_+(H) \cdot v^T H v.
\]

Hence, the estimates $\|Hv\| = \Theta(\|\Delta\|^2)$ in (A.13) and $|v^T H v| = O(\|\Delta\|^4)$ in (A.14) lead to $v^T H v = \Theta(\|\Delta\|^3)$. Consequently, if $\Delta^T_A H_{AA} \Delta_A \geq \eta_7 \|\Delta\|^{4-2\mu}$ for some constant $\eta_7 > 0$, then we can infer
\[
\Delta^T_A g_2 = \Delta^T_A (H^T_{II} \Delta_I + H_{AA} \Delta_A) + 2\|\Delta_A\|_2^4 - c_1 \|\Delta_A\|^2 \\
\geq \eta_7 \|\Delta\|^{4-2\mu} - 2\eta_5 \|\Delta\|^{4-\mu} + O(\|\Delta\|^4) \geq \frac{\eta_7}{2} \|\Delta\|^{4-2\mu}
\]
for $\Delta \to 0$. As in sub-case 3.1, this allows to show $\|\text{grad} f(y)\| \geq \Theta(\|\Delta\|^{3-2\mu})$ and thus, by (A.15), the Łojasiewicz inequality holds with $\theta = \frac{4+2\mu}{4} \geq \frac{1}{4}$. Finally, let us consider $\eta_8 \|\Delta\|^4 \leq \Delta^T_A H_{AA} \Delta_A \leq \eta_7 \|\Delta\|^{4-2\mu}$, where $\eta_8 > 0$ is chosen such that $\Delta^T H \Delta \geq \eta_8 \|\Delta\|^4$ and let us define the final decompositions
\[
\Delta_A = \psi_1 + \xi_1, \quad \psi_1 \in \text{null } H_{AA}, \quad \xi_1 \in [\text{null } H_{AA}]^\perp,
\]
\[
\text{diag}(\|\Delta_A\|^2) \Delta_A = \psi_2 + \xi_2, \quad \psi_2 \in \text{null } H_{AA}, \quad \xi_2 \in [\text{null } H_{AA}]^\perp,
\]
where $\text{null } M$ is the null space of a matrix $M$. We then have $\|\xi_1\| = \Theta(\|H_{AA} \xi_1\|) = \Theta(\sqrt{\xi^T H_{AA} \xi_1}) = O(\|\Delta\|^{2-\mu})$ and $\|\psi_1\| = O(\|\Delta\|)$. Notice that such decompositions exist due to the symmetry of $H_{AA}$.

Since $H$ is positive semidefinite, we can show that $\text{null } H_{AA} \subset \text{null } H_{II}$. If $H_{II} = 0$, then this claim is certainly true. Otherwise, if we assume that the statement is false, the set $S = \text{null } H_{AA} \cap [\text{null } H_{II}]^\perp$ does not only contain zero and we can select $\psi \in S \setminus \{0\}$ and $\xi \in [\text{null } H_{II}]^\perp \setminus \{0\} = [\text{ran } H_{II}] \setminus \{0\}$. Then it holds that
\[
\left[\begin{array}{c}
\xi^T \\
\psi^T
\end{array}\right]
\left[
\begin{array}{cc}
H_{II} & H_{IA} \\
H_{TI} & H_{AA}
\end{array}\right]
\left[
\begin{array}{c}
\xi \\
\psi
\end{array}\right] = \xi^T H_{II} \xi + 2a \psi^T H_{II} \xi \geq 0, \quad \forall a \in \mathbb{R}.
\]
But since $[H_{II} \psi]^T \xi \neq 0$, we can choose a such that $\xi^T H_{II} \xi + 2a \psi^T H_{II} \xi < 0$, which is a contradiction.

Hence and due to $H_{II}^T H_{II} \Delta_I \in \text{ran } H_{II}^T H_{II} = [\text{null } H_{II}]^\perp$, we can infer $H_{II}^T H_{II} \Delta_I + H_{AA} \Delta_A \in [\text{null } H_{AA}]^\perp$. This implies that $g_2$ can be written as $g_2 = g_3 + d$ where $g_3 \in [\text{null } H_{AA}]^\perp$ and $d = 2\psi_2 - c_1 \psi_1 \in \text{null } H_{AA}$. If $\|d\| \geq \frac{\eta_8}{2} \|\Delta\|^3$, we obtain $\|\text{grad } f(y)\| \geq \|g_2\| \geq \|d\| \geq \frac{\eta_8}{2} \|\Delta\|^3$ by (A.16) and the Łojasiewicz inequality is satisfied with $\theta = \frac{1}{3}$ due to (A.15). Otherwise, if $\|d\| \leq \frac{\eta_8}{2} \|\Delta\|^3$, it follows.
where we applied the estimates $\|\xi_2\| \leq \|A\Delta\|_6^3 \leq \|A\|^3$ and $c_1 = \Theta(\|A\|^2)$. Using (A.18), this shows

$$\Delta_t A g_2 = \Delta_t^T (H_{A\Delta}^T \Delta + H_{AA} \Delta_A) + \|A\|_4^4 - c_1 \|A\|_2^2 \geq \|A\|_T^4 - \frac{\eta_8}{2} \|A\|_T^4 + O(\|A\|_T^{5-\mu}) \geq \frac{\eta_8}{4} \|A\|_T^4$$

for $\Delta \to 0$. Thus, we have $\|\nabla f(y)\| \geq \Theta(\|A\|^3)$ and as before we can infer that the Łojasiewicz inequality holds with $\theta = \frac{1}{4}$ in this case. 

**References**

1. Absil, P.A., Mahony, R., Andrews, B.: Convergence of the iterates of descent methods for analytic cost functions. SIAM J. Optim. 16(2), 531–547 (2005)
2. Absil, P.A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton (2009)
3. Adhikari, S.K.: Numerical solution of the two-dimensional Gross–Pitaevskii equation for trapped interacting atoms. Phys. Lett. A 265, 91–96 (2000)
4. Anandkumar, A., Ge, R.: Efficient approaches for escaping higher order saddle points in non-convex optimization. In: 29th Annual Conference on Learning Theory. Proceedings of Machine Learning Research, vol. 49, pp. 81–102 (2016)
5. Arora, S., Ge, R., Ma, T., Moitra, A.: Simple, efficient, and neural algorithms for sparse coding. J. Mach. Learn. Res. (2015)
6. Attouch, H., Bolte, J.: On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. Math. Program. 116(1–2), 5–16 (2009)
7. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka–Lojasiewicz inequality. Math. Oper. Res. 35(2), 438–457 (2010)
8. Attouch, H., Bolte, J., Svaiter, B.F.: Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss–Seidel methods. Math. Program. 137(1–2, Ser. A), 91–129 (2013)
9. Bandeira, A.S., Boumal, N., Singer, A.: Tightness of the maximum likelihood semidefinite relaxation for angular synchronization. Math. Program. 163(1–2), 145–167 (2017)
10. Bandeira, A.S., Boumal, N., Voroninski, V.: On the low-rank approach for semidefinite programs arising in synchronization and community detection. In: Conference on Learning Theory, pp. 361–382 (2016)
11. Bao, W., Cai, Y.: Mathematical theory and numerical methods for Bose–Einstein condensation. Kinetic. Relat. Models 6(1), 1–135 (2012)
12. Bolte, J., Daniilidis, A., Lewis, A.: The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. SIAM J. Optim. 17(4), 1205–1223 (2006)
13. Bolte, J., Daniilidis, A., Lewis, A.: A nonsmooth Morse–Sard theorem for subanalytic functions. J. Math. Anal. Appl. 321(2), 729–740 (2006)
14. Bolte, J., Daniilidis, A., Lewis, A., Shiota, M.: Clarke subgradients of stratifiable functions. SIAM J. Optim. 18(2), 556–572 (2007)
15. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization or nonconvex and nonsmooth problems. Math. Program. 146(1–2), 459–494 (2014)
16. Bonettini, S., Loris, I., Porta, F., Prato, M., Rebegoldi, S.: On the convergence of a line search based proximal-gradient method for nonconvex optimization. Inverse Probl. 33(5) (2017)
17. Boumal, N.: Nonconvex phase synchronization. SIAM J. Optim. 26(4), 2355–2377 (2016)
18. Cai, Y., Zhang, L., Bai, Z., Li, R.C.: On an eigenvector-dependent nonlinear eigenvalue problem. SIAM J. Matrix Anal. Appl. 39(3), 1360–1382 (2018)
19. Candès, E.J., Li, X.: Solving quadratic equations via phaseless when there are as many equations as unknowns. Found. Comput. Math. 14(5), 1017–1026 (2014)
20. Candès, E.J., Li, X., Soltanolkotabi, M.: Phase retrieval via wirtinger flow: theory and algorithms. IEEE Trans. Inf. Theory 61(4), 1985–2007 (2015)
21. Candès, E.J., Strohmer, T., Voroninski, V.: Phaselift: exact and stable signal recovery from magnitude measurements via convex programming. Commun. Pure Appl. Math. 66(8), 1241–1274 (2013)
22. Candès, E.J., Tao, T.: The power of convex relaxation: near-optimal matrix completion. IEEE Trans. Inf. Theory 56(5), 2053–2080 (2010)
23. Cartis, C., Gould, N.I.M., Toint, P.L.: Second-order optimality and beyond: characterization and evaluation complexity in convexly constrained nonlinear optimization. Found. Comput. Math. 18(5), 1073–1107 (2018)
24. Chen, Y., Chi, Y., Fan, J., Ma, C.: Gradient descent with random initialization: fast global convergence for nonconvex phase retrieval. Math. Program. 176(1-2, Ser. B), 5–37 (2019)
25. Chen, Y., Chi, Y., Goldsmith, A.J.: Exact and stable covariance estimation from quadratic sampling via convex programming. IEEE Trans. Inf. Theory 61(7), 4034–4059 (2015)
26. Chi, Y., Lu, Y.M., Chen, Y.: Nonconvex optimization meets low-rank matrix factorization: an overview (2018). Preprint arXiv:1809.09573
27. D’Acunto, D., Kurdyka, K.: Explicit bounds for the Łojasiewicz exponent in the gradient inequality for polynomials. Ann. Pol. Math. 87, 51–61 (2005)
28. Dedieu, J.P.: Third- and fourth-order optimality conditions in optimization. Optimization 33(2), 97–104 (1995)
29. Edwards, M., Burnett, K.: Numerical solution of the nonlinear Schrödinger equation for small samples of trapped neutral atoms. Phys. Rev. A 51(2), 1382 (1995)
30. Forti, M., Nistri, P., Quincampoix, M.: Convergence of neural networks for programming problems via a nonsmooth Łojasiewicz inequality. IEEE Trans. Neural Netw. 17(6), 1471–1486 (2006)
31. Gao, B., Liu, X., Chen, X., Yuan, Y.: On the Łojasiewicz exponent of the quadratic sphere constrained optimization problem (2016). Preprint arXiv:1611.08781
32. Gao, B., Liu, X., Chen, X., Yuan, Y.: A new first-order algorithmic framework for optimization problems with orthogonality constraints. SIAM J. Optim. 28(1), 302–332 (2018)
33. García-Ripoll, J.J., Pérez-García, V.M.: Optimizing Schrödinger functionals using sobolev gradients: applications to quantum mechanics and nonlinear optics. SIAM J. Sci. Comput. 23(4), 1316–1334 (2001)
34. Ge, R., Huang, F., Jin, C., Yuan, Y.: Escaping from saddle points–online stochastic gradient for tensor decomposition. In: Conference Learning Theory, pp. 797–842 (2015)
35. Ge, R., Jin, C., Zheng, Y.: No spurious local minima in nonconvex low rank problems: a unified geometric analysis, vol. 70, pp. 1233–1242. In: Conference Machine Learning (2017)
36. Ge, R., Lee, J.D., Ma, T.: Matrix completion has no spurious local minimum. Adv. Neural Inf. Process. Syst. 2973–2981 (2016)
37. Ge, R., Ma, T.: On the optimization landscape of tensor decompositions. Adv. Neural Inf. Process. Syst. 3653–3663 (2017)
38. Griffin, A., Smoke, D.W., Stringari, S.: Bose–Einstein Condensation. Cambridge University Press, Cambridge (1996)
39. Gwoździewicz, J.: The Łojasiewicz exponent of an analytic function at an isolated zero. Comment. Math. Helv. 74(3), 364–375 (1999)
40. Hu, J., Jiang, B., Liu, X., Wen, Z.: A note on semidefinite programming relaxations for polynomial optimization over a single sphere. Sci. China Math. 59(8), 1543–1560 (2016)
41. Hu, J., Milzarek, A., Wen, Z., Yuan, Y.: Adaptive quadratically regularized Newton method for Riemannian optimization. SIAM J. Matrix Anal. Appl. 39(3), 1181–1207 (2018)
42. Jaganathan, K., Eldar, Y.C., Hassibi, B.: Phase retrieval: an overview of recent developments (2015). Preprint arXiv:1510.07713
43. Jain, P., Kar, P.: Non-convex optimization for machine learning. Found. Trends® Mach. Learn. 10(3–4), 142–336 (2017)
44. Keshavan, R.H., Montanari, A., Oh, S.: Matrix completion from a few entries. IEEE Trans. Inf. Theory 56(6), 2980–2998 (2010)
45. Keshavan, R.H., Montanari, A., Oh, S.: Matrix completion from noisy entries. J. Mach. Learn. Res. 11(Jul), 2057–2078 (2010)
46. Kreutz-Delgado, K.: The complex gradient operator and the CR-calculus. Preprint arXiv:0906.4835 (2009)
47. Kurdyka, K.: On gradients of functions definable in o-minimal structures. Ann. Inst. Fourier (Grenoble) 48(3), 769–783 (1998)
48. Lee, J.D., Panageas, I., Piliouras, G., Simchowitz, M., Jordan, M.I., Recht, B.: First-order methods almost always avoid strict saddle points. Math. Program. 176(1–2, Ser. B), 311–337 (2019)
49. Lee, J.D., Simchowitz, M., Jordan, M.I., Recht, B.: Gradient descent converges to minimizers. In: Conference Learning Theory, pp. 1246–1257 (2016)
50. Li, G., Pong, T.K.: Calculus of the exponent of Kurdyka–Łojasiewicz inequality and its applications to linear convergence of first-order methods. Found. Comput. Math. 18(5), 1199–1232 (2018)
51. Li, X., Zhu, Z., So, A.M.C., Vidal, R.: Nonconvex robust low-rank matrix recovery (2018). Preprint arXiv:1809.09237
52. Liang, S., Sun, R., Li, Y., Srikant, R.: Understanding the loss surface of neural networks for binary classification. In: International Conference on Machine Learning, pp. 2835–2843 (2018)
53. Liu, H., So, A.M.C., Wu, W.: Quadratic optimization with orthogonality constraint: explicit Łojasiewicz exponent and linear convergence of retraction-based line-search and stochastic variance-reduced gradient methods. Math. Program. 1–48 (2017)
54. Liu, H., Yue, M.C., Man-Cho So, A.: On the estimation performance and convergence rate of the generalized power method for phase synchronization. SIAM J. Optim. 27(4), 2426–2446 (2017)
55. Łojasiewicz, S.: Une propriété topologique des sous-ensembles analytiques réels. Les équations aux dérivées partielles 117, 87–89 (1963)
56. Luo, Z.Q.: Error bounds for analytic systems and their applications. Math. Program. 67(1–3), 1–28 (1994)
57. Luo, Z.Q., Sturm, J.F.: Error bounds for quadratic systems. In: High Performance Optimization, pp. 383–404. Springer (2000)
58. Marčenko, V.A., Pastur, L.A.: Distribution of eigenvalues for some sets of random matrices. Math. USSR-Sbornik 1(4), 457 (1967)
59. Merlet, B., Nguyen, T.N., et al.: Convergence to equilibrium for discretizations of gradient-like flows on Riemannian manifolds. Differ. Equ. 26(5/6), 571–602 (2013)
60. More, J.J.: Generalizations of the trust region problem. Optim. Methods Softw. 2(3–4), 189–209 (1993)
61. Murty, K.G., Kabadi, S.N.: Some NP-complete problems in quadratic and nonlinear programming. Math. Program. 39(2), 117–129 (1987)
62. Ochs, P., Chen, Y., Brox, T., Pock, T.: iPiano: inertial proximal algorithm for nonconvex optimization. SIAM J. Imaging Sci. 7(2), 1388–1419 (2014)
63. Panageas, I., Piliouras, G.: Gradient descent only converges to minimizers: non-isolated critical points and invariant regions. arXiv:1605.00405 (2016)
64. Penot, J.P.: Higher-order optimality conditions and higher-order tangent sets. SIAM J. Optim. 27(4), 2508–2527 (2017)
65. Pethick, C.J., Smith, H.: Bose–Einstein Condensation in Dilute Gases. Cambridge University Press, Cambridge (2002)
66. Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T.: The art of scientific computing. In: Numerical Recipes. Cambridge University Press, Cambridge (1986)
67. Recht, B.: A simpler approach to matrix completion. J. Mach. Learn. Res. 12(Dec), 3413–3430 (2011)
68. Schneider, R., Uschmajew, A.: Convergence results for projected line-search methods on varieties of low-rank matrices via Łojasiewicz inequality. SIAM J. Optim. 25(1), 622–646 (2015)
69. Shechtman, Y., Eldar, Y.C., Cohen, O., Chapman, H.N., Miao, J., Segev, M.: Phase retrieval with application to optical imaging: a contemporary overview. IEEE Signal Process. Mag. 32(3), 87–109 (2015)
70. Sorber, L., Barel, M.V., Lathauwer, L.D.: Unconstrained optimization of real functions in complex variables. SIAM J. Optim. 22(3), 879–898 (2012)
71. Sun, J.: When Are Nonconvex Optimization Problems Not Scary? Columbia University, New York (2016)
72. Sun, J., Qu, Q., Wright, J.: A geometric analysis of phase retrieval. In: IEEE International Symposium on Information Theory (ISIT), 2016, pp. 2379–2383. IEEE (2016)
On the geometric analysis of a quartic–quadratic...

73. Sun, J., Qu, Q., Wright, J.: Complete dictionary recovery over the sphere i: overview and the geometric picture. IEEE Trans. Inf. Theory 63(2), 853–884 (2017)
74. Sun, J., Qu, Q., Wright, J.: Complete dictionary recovery over the sphere ii: recovery by Riemannian trust-region method. IEEE Trans. Inf. Theory 63(2), 885–914 (2017)
75. Sun, J., Qu, Q., Wright, J.: A geometric analysis of phase retrieval. Found. Comput. Math. 18(5), 1131–1198 (2018)
76. Sun, R., Luo, Z.Q.: Guaranteed matrix completion via non-convex factorization. IEEE Trans. Inf. Theory 62(11), 6535–6579 (2016)
77. Thoai, N.V.: General quadratic programming. In: Essays and Surveys in Global Optimization GERAD 25th Anniversary Series, vol. 7, pp. 107–129. Springer, New York (2005)
78. Wen, Z., Yin, W.: A feasible method for optimization with orthogonality constraints. Math. Program. 142(1–2, Ser. A), 397–434 (2013)
79. Wu, X., Wen, Z., Bao, W.: A regularized Newton method for computing ground states of Bose–Einstein condensates. SIAM J. Sci. Comput. 73(1), 303–329 (2017)
80. Yang, W.H.: Error bounds for convex polynomials. SIAM J. Optim. 19(4), 1633–1647 (2008)
81. Yang, W.H., Zhang, L.H., Song, R.: Optimality conditions for the nonlinear programming problems on Riemannian manifolds. Pac. J. Optim. 10, 415–434 (2014)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Haixiang Zhang1 · Andre Milzarek2,3,4 · Zaiwen Wen5,6 · Wotao Yin7

Zaiwen Wen
wenzw@pku.edu.cn
Haixiang Zhang
haixiang@math.berkeley.edu
Andre Milzarek
andremilzarek@cuhk.edu.cn
Wotao Yin
wotaoyin@math.ucla.edu

1 Department of Mathematics, University of California, Berkeley, CA, USA
2 School of Data Science (SDS), Chinese University of Hong Kong, Shenzhen, China
3 Shenzhen Research Institute of Big Data (SRIBD), Shenzhen, China
4 Shenzhen Institute for Artificial Intelligence and Robotics for Society (AIRS), Shenzhen, China
5 Beijing International Center for Mathematical Research, Peking University, Beijing, China
6 Center for Data Science, Peking University, Beijing, China
7 Department of Mathematics, University of California, Los Angeles, CA, USA