Journal of Machine Learning Research 24 (2023) 1-47 Submitted 9/19; Revised 10/23; Published 11/23

Scale Invariant Power Iteration

Cheolmin Kim* CHEOLMKIM@U.NORTHWESTERN.EDU
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL, 60208, USA

Youngseok Kim* YOUNGSEOK@UCHICAGO.EDU
Department of Statistics
University of Chicago
Chicago, IL, 60637, USA

Diego Klabjan D-KLABJAN@NORTHWESTERN.EDU
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL, 60208, USA

Editor: Zaid Harchaoui

Abstract

We introduce a new class of optimization problems called *scale invariant problems* that cover interesting problems in machine learning and statistics and show that they are efficiently solved by a general form of power iteration called *scale invariant power iteration (SCI-PI)*. SCI-PI is a special case of the generalized power method (GPM) (Journée et al., 2010) where the constraint set is the unit sphere. In this work, we provide the convergence analysis of SCI-PI for scale invariant problems which yields a better rate than the analysis of GPM. Specifically, we prove that it attains local linear convergence with a generalized rate of power iteration to find an optimal solution for scale invariant problems. Moreover, we discuss some extended settings of scale invariant problems and provide similar convergence results. In numerical experiments, we introduce applications to independent component analysis, Gaussian mixtures, and non-negative matrix factorization with the KL-divergence. Experimental results demonstrate that SCI-PI is competitive to application specific state-of-the-art algorithms and often yield better solutions.

Keywords: scale invariance, power iteration, optimization, convergence analysis, machine learning applications

1. Introduction

We study a new class of optimization problems called *scale invariant problems* having the form of

$$\text{maximize} \quad f(x) \quad \text{subject to} \quad x \in \partial\mathcal{B}_d \triangleq \{x \in \mathbb{R}^d : \|x\|_2 = 1\}, \tag{1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a twice continuously differentiable, *scale invariant function*. We say that a function f is scale invariant, which is rigorously defined later in Definition 1, if its geometric surface is invariant under constant multiplication of x. Many important

*The two authors contributed equally to this paper.

optimization problems in statistics and machine learning can be formulated as scale invariant problems, for instance, L_p-norm kernel PCA and maximum likelihood estimation of mixture proportions, to name a few. Moreover, as studied herein, independent component analysis (ICA, Example 3), Gaussian mixture models (GMM, Example 4), and non-negative matrix factorization with the Kullback-Leibler divergence (KL-NMF, Example 5) can be formulated as extended settings of scale invariant problems where the objective function is a sum of scale invariant functions or the problem is scale invariant with respect to a subset of variables while the other variables are fixed.

Since $\partial\mathcal{B}_d$ is not a convex set, scale invariant problems are in general non-convex optimization problems. Nevertheless, some instances can be efficiently solved, for instance, the leading eigenvector problem (Golub and Van Loan, 2012). Power iteration (Muntz, 1913; Mises and Pollaczek-Geiringer, 1929) is an algorithm to find the leading eigenvector of a diagonalizable matrix A. In power iteration, $x_{k+1} \leftarrow Ax_k/\|Ax_k\|_2$ is repeatedly applied until some stopping criterion is satisfied. Since no hyperparameter is required, this update rule is practical but at the same time it attains global linear convergence with the rate of $|\lambda_2|/|\lambda_1|$ where $|\lambda_i|$ is the i^{th} largest absolute eigenvalue of A (Wilkinson, 1965; Golub and Van Loan, 2012). The linear convergence property of power iteration has been extended to many applications. However, theoretical understanding of when and how such algorithms enjoy this attractive convergence property of power iteration is limited. For example, for convex f, a general form of power iteration called *generalized power method (GPM)* (Journée et al., 2010) has been shown to attain only global sublinear convergence rate of $O(1/\epsilon)$, not generalizing the appealing linear convergence of power iteration. For historical development of power iteration, see Golub and Van der Vorst (2000); Tapia et al. (2018).

In this work, we present scale invariant problems that generalize the leading eigenvector problem in the sense that any stationary point x^* of (1) satisfying $\nabla f(x^*) = \lambda^* x^*$ for some λ^* is an eigenvector of $\nabla^2 f(x^*)$. By this property, scale invariant problems can be seen as the leading eigenvector problem near a local optimum x^*, so we can expect that a general form of power iteration would work well for them. By swapping the objective function and the constraint, we obtain a geometrically interpretable dual problem with the goal of finding the closest point w to the origin from the constraint $f(w) = 1$. By mapping an iterate x_k to the dual space, taking a descent step in the dual space and mapping it back to the original space, we geometrically derive scale invariant power iteration (SCI-PI), which replaces Ax_k with $\nabla f(x_k)$ in power iteration. SCI-PI is the same algorithm as GPM applied to the unit sphere constraint. However, we improve the convergence rate of GPM for scale invariant problems showing that the algorithm attains local linear convergence with a generalized rate of power iteration when initialized close to it. To the best of our knowledge, this is the first work exploiting the properties of scale invariant problems. Also, this is the first linear convergence result of GPM for general optimization problems. This improvement is significant since with linear convergence, the iteration complexity to attain an ϵ-optimal solution reduces from $O(1/\epsilon)$ to $O(1/\log(1/\epsilon))$. Moreover, under some mild conditions, we provide an explicit expression regarding the initial condition on $\|x_0 - x^*\|_2$ to ensure convergence.

In the extended settings (Section 4), we discuss three variants of (1). In the first setting, we consider a sum of scale invariant functions (Subsection 4.1) as an objective function. This setting covers a Kurtosis-based ICA and can be solved by SCI-PI with similar convergence guarantees. Second, we consider a block version of scale invariant problems (Subsection 4.2)

which covers KL-NMF and the Burer-Monteiro factorization of semi-definite programs. To solve this block scale invariant problem, we present a block version of SCI-PI and show that it attains linear convergence in a two-block case. Lastly, we consider partially scale invariant problems (Subsection 4.3) which include general mixture problems such as GMM. For this partially scale invariant problems, we present an alternating algorithm based on SCI-PI and gradient ascent along with its convergence analysis. In numerical experiments, we benchmark the proposed algorithms against state-of-the-art methods for ICA, KL-NMF, and GMM. The experimental results show that our algorithms are computationally competitive and result in better solutions in "most" if we do not beat in all herein studied cases.

In summary, this work has the following contributions.

1. We introduce scale invariant problems which cover interesting examples in statistics and machine learning. By the eigenvector property (Proposition 4), they resemble the leading eigenvector problem near a local optimum x^*.

2. For scale invariant problems, we prove that SCI-PI (a special form of GPM) converges to a local maximum x^* at a logarithmic rate when initialized close to x^*. This generalizes the attractive convergence property of power iteration. Moreover, we introduce three extended settings of scale invariant problems along with solution algorithms and their convergence analyses.

3. We report numerical experiments including a novel reformulation of KL-NMF to a block scale invariant problem. The experimental results demonstrate that SCI-PI is not only computationally competitive to state-of-the-art methods but also often yield better solutions.

The paper is organized as follows. In Section 2, we define scale invariance and present interesting properties of scale invariant problems including an eigenvector property and a dual formulation. We then provide a geometric derivation of SCI-PI and a convergence analysis in Section 3. The extended settings are discussed in Section 4 and we report the numerical experiments in Section 5. We finish the introduction with literature review and a notation paragraph.

1.1 Related Works

Power Iteration The global linear convergence property of power iteration is analogous to that of gradient descent for convex optimization. Therefore, many variants including coordinate-wise (Lei et al., 2016), momentum (Xu et al., 2018), online (Boutsidis et al., 2015; Garber et al., 2015), stochastic (Oja, 1982), stochastic variance-reduced (Shamir, 2015, 2016; Xu et al., 2018; Kim and Klabjan, 2020b), and truncated (Yuan and Zhang, 2013; Han and Liu, 2014) power iterations have been developed, drawing a parallel literature to gradient descent for convex optimization. We discover a class of optimization problems which can be locally seen as the leading eigenvector problem, and prove that they can be efficiently solved by a general form of power iteration.

Generalized Power Method (GPM) GPM (Journée et al., 2010) is an iterative algorithm that finds the next iterate by projecting the gradient at the current iterate to the constraint set. GPM has been applied to statistical problems such as sparse principal

component analysis (PCA) (Journée et al., 2010; Luss and Teboulle, 2013) and L_1-norm kernel PCA (Kim and Klabjan, 2020a). While GPM has a general form of power iteration, its convergence analysis does not extend the attractive convergence property of power iteration. For example, only global sublinear convergence has been shown for convex f. We generalize the local linear convergence property of power iteration to scale invariant problems.

Block Power Iteration A block version of power iteration has been developed to solve the phase synchronization problem (Boumal, 2016). If the problem consists of a single block and the shift parameter is set to zero, this algorithm specializes to power iteration. Under some conditions on the measurement noise and the initial iterate, it attains linear convergence to a global solution (Liu et al., 2017). To solve the Burer-Monteiro factorization (Burer and Monteiro, 2003) of semi-definite programs (Vandenberghe and Boyd, 1996), Erdogdu et al. (2022) developed a block coordinate maximization (BCM) algorithm. By iteratively sampling a block and applying power iteration to it, BCM attains local linear convergence as well as global sublinear convergence. However, the linear convergence property of block power iteration has not been extended to more general settings. In this work, we prove that block variants of SCI-PI attain linear convergence for block and partially scale invariant problems.

Alternating Minimization Alternating algorithms have been developed for many applications such as k-means clustering (MacQueen, 1967), Gaussian mixture model (Bishop, 2006), dictionary learning (Olshausen and Field, 1997; Aharon et al., 2006), matrix completion (Candès and Recht, 2009), matrix factorization (Lee and Seung, 2001), and finding a point in the intersection of two closed sets (Lewis et al., 2009). Beck (2015) studied alternating minimization and proved that it achieves sublinear convergence for convex programming. For optimization problems with a separable convex objective function and a linear constraint, alternating direction method of multipliers (ADMM) (Boyd et al., 2011) has been shown to attain linear convergence (Hong and Luo, 2017). However, due to the exact minimization step, ADMM can incur high per iteration cost. Instead of performing exact minimization, our algorithms alternatively apply simple steps to update blocks but at the same time they achieve local linear convergence.

Manifold Optimization Viewed as an optimization problem on the real projective plane, a scale invariant problem can be reformulated to an equivalent problem in the embedding space. The reformulated problem is unconstrained in the embedding space but it has a highly non-convex structure, e.g., the maximization of the Rayleigh quotient. In order to solve the reformulated problem, general algorithms for unconstrained non-convex optimization such as gradient and Newton methods with line search, and trust region method (Absil et al., 2009) can be employed. Rather than working in the embedding space, we focus on a generalization of power iteration.

Gauge Optimization A gauge function which is a nonnegative, convex, and positively homogeneous function that vanishes at the origin is a multiplicatively scale invariant function. Gauge functions include norms and pseudonorms as special cases and generalize the notion of a norm. The *gauge program* that minimizes a gauge function over a convex set is introduced in (Freund, 1987) and further studied in (Friedlander et al., 2014). The literature on gauge optimization is mainly about developing and studying dual problems. Conversely, we develop a simple numerical algorithm that solves the primal problem.

1.2 Notation

Let $\mathbb{R}$ and $\mathbb{R}^+$ denote the set of real numbers and the set of non-negative real numbers, respectively. Let d be the dimension of the optimization variable x. Let $\mathbb{R}^d$ denote the set of d-dimensional real vectors and f be a function from $\mathbb{R}^d$ to $\mathbb{R}$. We denote the gradient and Hessian of a function f as ∇f and $\nabla^2 f$. Let u and v be functions from $\mathbb{R}$ to $\mathbb{R}^+$ and $\mathbb{R} \setminus \{0\}$ to $\mathbb{R}$, representing multiplicative and additive factor functions, respectively, and let p be the degree of a scale invariant function, which equals to the degree of homogeneity for a multiplicative scale invariant function and 0 for an additive scale invariant function. We use $(\lambda_i, \mathbf{v}_i)$ and $(s_i, \mathbf{u}_i)$ to represent eigen-pairs. The j^{th} element of $\mathbf{v}_i$ is denoted as $\mathbf{v}_{i,j}$. Let k be the iteration index and we denote the sequences of iterates and function values by $\{x_k\}_{k=0,1,\cdots}$ and $\{f(x_k)\}_{k=0,1,\cdots}$, respectively. Lastly, we let $\odot$, $\oslash$ and $(\cdot)^{\odot 2}$ denote element-wise product, division and square, respectively and let $\mathbb{1}_n \in \mathbb{R}^n$ denote the vector of n ones.

2. Scale Invariant Problems

Before presenting properties of scale invariant problems, we first define scale invariant functions.

Definition 1 *We say that a function $f : \mathbb{R}^d \to \mathbb{R}$ is multiplicatively scale invariant if it satisfies*

$$f(cx) = u(c)f(x) \tag{2}$$

for some even function $u : \mathbb{R} \to \mathbb{R}^+$ with $u(0) = 0$. Also, we say that $f : \mathbb{R}^d \setminus \{0\} \to \mathbb{R}$ is additively scale invariant if it satisfies

$$f(cx) = f(x) + v(c) \tag{3}$$

for some even function $v : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ with $v(1) = 0$.

The following proposition characterizes the exact form of u and v for continuous f.

Proposition 2 *If a continuous function $f \neq 0$ satisfies (2) with a multiplicative factor u, then we have*

$$u(c) = |c|^p \tag{4}$$

for some $p > 0$. Also, if a continuous function f satisfies (3) with an additive factor v, then we have

$$v(c) = \log_a |c| \tag{5}$$

for some a such that $0 < a$ and $a \neq 1$.

Using the explicit forms of u and v in Proposition 2, we establish derivative-based properties of scale invariant functions below.

Proposition 3 *Suppose that f is twice differentiable. If f satisfies (2) with a multiplicative factor $u(c) = |c|^p$, we have*

$$c\nabla f(cx) = |c|^p \nabla f(x), \quad \nabla f(x)^T x = pf(x), \quad \nabla^2 f(x)x = (p-1)\nabla f(x). \tag{6}$$

Also, if f satisfies (3) *with an additive factor $v(c) = \log_a |c|$, we have*

$$c\nabla f(cx) = \nabla f(x), \quad \nabla f(x)^T x = \frac{1}{\log(a)}, \quad \nabla^2 f(x)x = -\nabla f(x). \tag{7}$$

Proposition 3 states that a scale invariant function f satisfies that $\nabla^2 f(x)x$ is a scalar multiple of $\nabla f(x)$. Let

$$\mathcal{L}(\lambda, x) = f(x) + \lambda\left(1 - \|x\|_2\right).$$

be the Lagrange function of (1) and (λ^*, x^*) be a stationary point satisfying $\nabla\mathcal{L}(\lambda, x) = 0$ such that

$$\nabla f(x^*) = \lambda^* x^*, \quad \|x^*\|_2 = 1. \tag{8}$$

In the next proposition, we derive an eigenvector property which states that for any stationary point (λ^*, x^*) of (1) satisfying (8), x^* is an eigenvector of $\nabla^2 f(x^*)$.

Proposition 4 *Suppose that f is twice differentiable and let (λ^*, x^*) be a stationary point of* (1) *satisfying* (8)*. If f satisfies* (2) *with $u(c) = |c|^p$, then we have*

$$\nabla^2 f(x^*)x^* = (p-1)\lambda^* x^*.$$

Also, if f satisfies (3) *with $v(c) = \log_a |c|$, then we have*

$$\nabla^2 f(x^*)x^* = -\lambda^* x^*.$$

In both cases, x^* is an eigenvector of $\nabla^2 f(x^*)$. Moreover, if λ^* is greater than the largest eigenvalue of $\nabla^2 f(x^*)(I - x^*(x^*)^T)$, then x^* is a local maximum to (1)*.*

Proof If f is multiplicative scale invariant with the degree of p, by Proposition 3, we have

$$\nabla^2 f(x^*)x^* = (p-1)\nabla f(x^*) = (p-1)\lambda^* x^*.$$

Also, by Proposition 3, if f is additive scale invariant f, we have

$$\nabla^2 f(x^*)x^* = -\nabla f(x^*) = -\lambda^* x^*.$$

Therefore, in both cases, a stationary point x^* is an eigenvector of $\nabla^2 f(x^*)$.

Suppose that λ^* is greater than the largest eigenvalue of $\nabla^2 f(x^*)(I - x^*(x^*)^T)$. For any h satisfying $h^T x^* = 0$, we have

$$\begin{aligned} h^T \nabla^2_{xx}\mathcal{L}(x^*, \lambda^*)h &= h^T\left(\nabla^2 f(x^*) - \lambda^*(I - x^*(x^*)^T)\right)h \\ &= h^T\nabla^2 f(x^*)(I - x^*(x^*)^T)h - \lambda^*\|h\|_2^2 < 0. \end{aligned}$$

Since the second-order sufficient condition is satisfied, x^* is a local maximum. ■

Proposition 4 states that a stationary point x^* is an eigenvector of $\nabla^2 f(x^*)$. Note that the Lagrange multiplier λ^* is not necessarily an eigenvalue corresponding to x^*. The eigenvalue

corresponding to x^* is $(p-1)\lambda^*$ if f is multiplicatively scale invariant or $-\lambda^*$ if f is additively scale invariant. The second-order sufficient condition for local optimality requires that the Lagrange multiplier λ^* rather than the eigenvalue corresponding to x^* is greater than the largest eigenvalue of $\nabla^2 f(x^*)(I - x^*(x^*)^T)$. Due to this eigenvector property, scale invariant problems can be considered as a generalization of the leading eigenvector problem. Next, we introduce a dual formulation of scale invariant problems.

Proposition 5 *Suppose that the objective function f is continuous and either multiplicatively scale invariant with a positive optimal value and a multipicative factor $u(c) = |c|^p$ such that $p > 0$ or additively scale invariant having an additive factor $v(c) = \log_a |c|$ such that $a > 1$. Then, solving (1) is equivalent to solving the following optimization problem*

$$\text{minimize} \quad \|w\|_2 \quad \text{subject to} \quad f(w) = 1. \tag{9}$$

In other words, if x^* is an optimal solution to (1), then $w^* = x^*/f(x^*)^{1/p}$ (multiplicative) or $w^* = a^{1-f(x^*)}x^*$ (additive) is an optimal solution to (9). Conversely, if w^* is an optimal solution to (9), $x^* = w^*/\|w^*\|_2$ is an optimal solution to (1).

For a multiplicatively scale invariant f having a negative optimal value and a multiplicative factor $u(c) = |c|^p$ such that $p < 0$, we can derive a similar reformulation by replacing $f(w) = 1$ with $f(w) = -1$. On the other hand, for an additively scale invariant f having an additive factor $v(c) = \log_a |c|$ such that $0 < a < 1$, we obtain a maximization problem with the same objective function and constraint. The dual formulation (9) has a nice geometric interpretation that an optimal solution w^* is the closest point to the origin from $\{w : f(w) = 1\}$. We use this understanding to derive SCI-PI in Section 3.

Lastly, we introduce two well-known examples of scale invariant problems in machine learning and statistics.

Example 1 (L_p-norm Kernel PCA) *Given data vectors $a_i \in \mathbb{R}^d$ and a mapping Φ, L_p-norm PCA considers*

$$\text{maximize} \quad \frac{1}{n}\sum_{i=1}^n \|\Phi(a_i)^T x\|_p^p \quad \text{subject to} \quad x \in \partial\mathcal{B}_d \tag{10}$$

where the objective function satisfies property (2) with $u(c) = |c|^p$.

Example 2 (Estimation of Mixture Proportions) *Given a design matrix $L \in \mathbb{R}^{n\times d}$ satisfying $L_{ij} \geq 0$, the problem of estimating mixture proportions seeks to find a vector π of mixture proportions on the probability simplex $\mathcal{S}^d = \left\{\pi : \sum_{j=1}^d \pi_j = 1,\ \pi \geq 0\right\}$ that solves*

$$\text{maximize} \quad \frac{1}{n}\sum_{i=1}^n \log\left(\sum_{j=1}^d L_{ij}\pi_j\right) \quad \text{subject to} \quad \pi \in \mathcal{S}^d. \tag{11}$$

By reparametrizing π_j by x_j^2, we obtain an equivalent optimization problem

$$\text{maximize} \quad \frac{1}{n}\sum_{i=1}^n \log\left(\sum_{j=1}^d L_{ij}x_j^2\right) \quad \text{subject to} \quad x \in \partial\mathcal{B}_d, \tag{12}$$

which now satisfies property (3) with $v(c) = 2\log|c|$.

The reformulation idea in Example 2 implies that any simplex-constrained problem with scale invariant f can be reformulated to a scale invariant problem.

3. Scale Invariant Power Iteration

In this section, we provide a geometric derivation of SCI-PI to find a local optimal solution of (1). The algorithm is developed using the geometric interpretation of the dual formulation (9) as illustrated in Figure 1. Starting with an iterate $x_k \in \partial\mathcal{B}$, we obtain a dual iterate w_k by mapping x_k to the constraint $f(w) = 1$. Given w_k, we identify the hyperplane l_k on which the current iterate w_k lies and is tangent to $f(w) = 1$. After identifying the equation of l_k, we find the closest point z_k to the origin from l_k and obtain a new dual iterate w_{k+1} by mapping z_k to the constraint $f(w) = 1$. Finally, we obtain a new primal iterate x_{k+1} by mapping w_{k+1} back to the set $\partial\mathcal{B}_d$.

Now, we develop an algorithm based on the above idea. For derivation of the algorithm, we assume that an objective function f is differentiable and satisfies either (2) with $u(c) = |c|^p$ where $p > 0$ and $f(x) > 0$ for all $x \in \partial\mathcal{B}$ or (3) with $v(c) = \log_a|c|$ where $1 < a$. Under these conditions, a scalar mapping from x_k to w_k can be well defined as $w_k = x_k/f(x_k)^{1/p}$ or $w_k = a^{1-f(x_k)}x_k$, respectively. Let $w_k = c_k x_k$. Since w_k is on the constraint $f(w) = 1$, the normal vector of the hyperplane l_k is $\nabla f(w_k)$. Therefore, we can write down the equation of the hyperplane l_k as $\{w : \nabla f(w_k)^T(w - w_k) = 0\}$. Note that z_k is a scalar multiple of $\nabla f(w_k)$ where the scalar can be determined from the requirement that z_k is on l_k. Since w_{k+1} is the projection of z_k, it must be a scalar multiple of the normal vector $y_k = \nabla f(w_k)$. Therefore, we can write w_{k+1} as $w_{k+1} = d_k y_k$. Finally, by projecting w_{k+1} to $\partial\mathcal{B}_d$, we obtain

$$x_{k+1} = \frac{w_{k+1}}{\|w_{k+1}\|_2} = \frac{d_k y_k}{\|d_k y_k\|_2} = \frac{y_k}{\|y_k\|_2} = \frac{\nabla f(w_k)}{\|\nabla f(w_k)\|_2} = \frac{\nabla f(c_k x_k)}{\|\nabla f(c_k x_k)\|_2} = \frac{\nabla f(x_k)}{\|\nabla f(x_k)\|_2}$$

where the last equality follows from Proposition 3. The update rule is the linear optimization oracle on $\partial\mathcal{B}_d$. In the sense that SCI-PI finds an optimal solution by solving a sequence of linear optimization problems, it is similar to the Frank-Wolfe algorithm (also called conditional gradients) and online linear prediction algorithms (Huang et al., 2017). Summarizing all the above, we obtain SCI-PI presented in Algorithm 1.

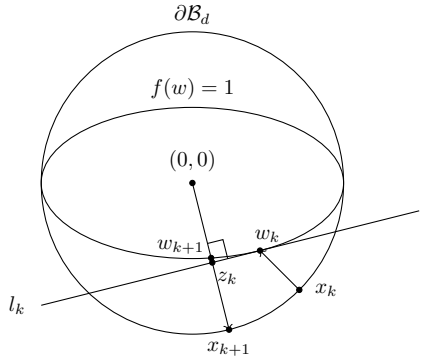

Figure 1: Geometric derivation of SCI-PI

Algorithm 1 SCI-PI

Output: initial point $x_0 \in \partial\mathcal{B}_d$
$k \leftarrow 0$
while $\nabla f(x_k) \neq 0$ **do**
  $x_{k+1} \leftarrow \dfrac{\nabla f(x_k)}{\|\nabla f(x_k)\|_2}$
  $k \leftarrow k + 1$
end while
Output: x_k

Note that in Figure 1 $\{\|w_k\|_2\}_{k=0,1,\dots}$ is non-increasing if the sublevel set $\{w \,|\, f(w) \leq 1\}$ is convex. Since all sublevel sets of a quasi-convex function are convex, we can expect that SCI-PI yields an ascending step if f is quasi-convex (not necessarily scale invariant). See Proposition 11 in Appendix B for the convergence to a stationary point for quasi-convex f. If f is not quasi-convex, the sequence $\{f(x_k)\}_{k=0,1,\dots}$ is not necessarily increasing, making

it hard to analyze global convergence. Exploiting the eigenvector property, we study local convergence of SCI-PI for scale invariant f below.

Theorem 6 *Let f be a scale invariant, twice continuously differentiable function on an open set containing $\partial\mathcal{B}_d$. Let x^* be a local maximum such that $\nabla f(x^*) = \lambda^* x^*$ and $(\lambda_i, \mathbf{v}_i)$ be an eigen-pair of $\nabla^2 f(x^*)$ with $x^* = \mathbf{v}_1$. If $\lambda^* > \bar{\lambda}_2 = \max_{2\le i\le d} |\lambda_i|$, then there exists some $\delta > 0$ such that under the initial condition $\|x_0 - x^*\|_2 < \delta$, the sequence of iterates $\{x_k\}_{k=0,1,\cdots}$ generated by SCI-PI satisfies*

$$\|x_k - x^*\|_2^2 \le \prod_{t=0}^{k-1} \left(\frac{\bar{\lambda}_2}{\lambda^*} + \gamma_t\right)^2 \|x_0 - x^*\|_2^2,$$

where

$$\frac{\bar{\lambda}_2}{\lambda^*} + \gamma_t < 1 \text{ for all } t \ge 0 \text{ and } \lim_{k\to\infty} \gamma_k = 0.$$

Moreover, if $\nabla_j f = \partial f/\partial x_j$ has a continuous Hessian H_j on an open set containing $\mathcal{B}_d \triangleq \{x \in \mathbb{R}^d : \|x\|_2 \le 1\}$, we can explicitly write δ as

$$\delta(\lambda^*, \bar{\lambda}_1, \bar{\lambda}_2, M) = \min\left\{ \frac{\sqrt{2}\lambda^*}{\lambda^* + |\lambda^* - \bar{\lambda}_1 - M|}, \frac{\sqrt{2}(\lambda^* - \bar{\lambda}_2)}{\lambda^* - \bar{\lambda}_2 + M + |\lambda^* - \bar{\lambda}_1 - M|} \right\}$$

where $\bar{\lambda}_1 = |\lambda_1|$ and

$$M = \max_{x\in\partial\mathcal{B}_d, y^1,\cdots,y^d\in\mathcal{B}_d} \sqrt{\textstyle\sum_{i=1}^d (x^T G_i(y^1,\cdots,y^d)x)^2}, \quad G_i(y^1,\cdots,y^d) = \textstyle\sum_{j=1}^d \mathbf{v}_{i,j} H_j(y^j).$$

Proof Since $\nabla^2 f(x^*)$ is real and symmetric, without loss of generality, we assume that $\{\mathbf{v}_1, \cdots, \mathbf{v}_d\}$ form an orthogonal basis in $\mathbb{R}^d$.

Since f is twice continuously differentiable on an open set containing $\partial\mathcal{B}_d$, for $x \in \partial\mathcal{B}_d$, using the Taylor expansion of $\nabla f(x)^T \mathbf{v}_i$ at x^*, we have

$$\nabla f(x)^T \mathbf{v}_i = \nabla f(x^*)^T \mathbf{v}_i + (x - x^*)^T \nabla^2 f(x^*) \mathbf{v}_i + R_i(x) \tag{13}$$

where

$$R_i(x) = o(\|x - x^*\|_2). \tag{14}$$

From $\nabla f(x^*) = \lambda^* x^*$ and $x^* = \mathbf{v}_1$, we have

$$\begin{aligned} \nabla f(x)^T \mathbf{v}_1 &= \nabla f(x^*)^T x^* + (x - x^*)^T \nabla^2 f(x^*) x^* + R_1(x) \\ &= \lambda^* - \lambda_1(1 - x^T x^*) + R_1(x) \\ &= \lambda^* + \alpha(x) \end{aligned} \tag{15}$$

where

$$\alpha(x) = -\lambda_1(1 - x^T x^*) + R_1(x). \tag{16}$$

On the other hand, for $2 \leq i \leq d$, due to $\nabla f(x^*) = \lambda^* x^*$, $x^* = \mathbf{v}_1$ and the orthogonality of $\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$, we have

$$\nabla f(x^*)^T \mathbf{v}_i = \lambda^* (x^*)^T \mathbf{v}_i = 0. \tag{17}$$

From (13), this results in

$$\nabla f(x)^T \mathbf{v}_i = \lambda_i x^T \mathbf{v}_i + R_i(x). \tag{18}$$

Using (18) and the definition of $\bar{\lambda}_2$, we have

$$\begin{aligned} \sum_{i=2}^{d} \left(\nabla f(x)^T \mathbf{v}_i\right)^2 &= \sum_{i=2}^{d} \left[\lambda_i^2 (x^T \mathbf{v}_i)^2 + 2\lambda_i (x^T \mathbf{v}_i) R_i(x) + (R_i(x))^2\right] \\ &\leq \bar{\lambda}_2^2 \sum_{i=2}^{d} (x^T \mathbf{v}_i)^2 + 2\bar{\lambda}_2 \sum_{i=2}^{d} |x^T \mathbf{v}_i| |R_i(x)| + \sum_{i=2}^{d} \left(R_i(x)\right)^2. \end{aligned} \tag{19}$$

By the Cauchy Schwartz inequality, we have

$$\bar{\lambda}_2 \sum_{i=2}^{d} |x^T \mathbf{v}_i| |R_i(x)| \leq \bar{\lambda}_2 \sqrt{\sum_{i=2}^{d} (x^T \mathbf{v}_i)^2} \sqrt{\sum_{i=2}^{d} \left(R_i(x)\right)^2},$$

which results in

$$\begin{aligned} \sum_{i=2}^{d} \left(\nabla f(x)^T \mathbf{v}_i\right)^2 &\leq \bar{\lambda}_2^2 \sum_{i=2}^{d} (x^T \mathbf{v}_i)^2 + 2\bar{\lambda}_2 \sqrt{\sum_{i=2}^{d} (x^T \mathbf{v}_i)^2} \sqrt{\sum_{i=2}^{d} \left(R_i(x)\right)^2} + \sum_{i=2}^{d} \left(R_i(x)\right)^2 \\ &= \left(\bar{\lambda}_2 \sqrt{\sum_{i=2}^{d} (x^T \mathbf{v}_i)^2} + \beta(x)\right)^2 \end{aligned} \tag{20}$$

where

$$\beta(x) = \sqrt{\sum_{i=2}^{d} \left(R_i(x)\right)^2}. \tag{21}$$

Using $1 - x^T x^* = o(\|x - x^*\|_2)$ and (14) for (16) and (21), we have

$$\alpha(x) = o(\|x - x^*\|_2), \quad \beta(x) = o(\|x - x^*\|_2). \tag{22}$$

From $x \in \partial \mathcal{B}_d, x^* = \mathbf{v}_1$, and the fact that $\{\mathbf{v}_1, \cdots, \mathbf{v}_d\}$ forms an orthogonal basis in $\mathbb{R}^d$, we have

$$\sum_{i=2}^{d} (x^T \mathbf{v}_i)^2 = 1 - (x^T \mathbf{v}_1)^2 = 1 - (x^T x^*)^2 \leq 2(1 - x^T x^*) = \|x - x^*\|_2^2. \tag{23}$$

Therefore, by (15), (20), (22), (23), and Lemma 12, we obtain the first part of the desired result.

On the other hand, if $\nabla_j f$ has a continuous Hessian H_j, by Lemma 14, we have

$$\sum_{i=1}^{d}(R_i(x))^2 \leq \frac{1}{4}M^2\|x-x^*\|_2^4. \tag{24}$$

Using (24) for (16) and (21), we have

$$\begin{aligned}\alpha(x) &= -\lambda_1(1-x_k^T x^*) + R_1(x) \geq -(M+|\lambda_1|)(1-x^T x^*),\\ \beta(x) &= \sqrt{\sum_{i=2}^{d}(R_i(x))^2} \leq \frac{M}{2}\|x-x^*\|_2^2.\end{aligned} \tag{25}$$

Therefore, using (15), (20), (23), (25) and Lemma 12 with

$$A=\lambda^*,\ B=M+|\lambda_1|,\ C=0,\ D=\bar{\lambda}_2,\ E=0,\ F=M,$$

we obtain the desired result. ■

Theorem 6 states local convergence of SCI-PI with an asymptotic rate of $\lambda^*/\bar{\lambda}_2$. Note that the assumption that the Lagrange multiplier λ^* corresponding to a local maximum x^* satisfies $\lambda^* > \bar{\lambda}_2 = \max_{2\leq i\leq d}|\lambda_i|$ holds for all strict local maxima if f is convex, multiplicatively scale invariant with $p \geq 1$ since $\lambda_i \geq 0$ for all i and $\lambda^* = (p-1)\lambda_1$, according to Proposition 4. However, in general, not all local maxima satisfy this assumption since it is stronger than the second-order sufficient condition stated as $\lambda^* > \max_{2\leq i\leq d}\lambda_i$ in Proposition 4. Nevertheless, by adding $\sigma\|x\|^2$ for some $\sigma > 0$ to the objective function f, we can always enforce $\lambda^* > \bar{\lambda}_2$. Conversely, by adding $\sigma\|x\|^2$ for some $\sigma < 0$, we may improve the convergence rate as done by shifted power iteration (Golub and Van Loan, 2012). The convergence rate of γ_k is $o(1)$ for twice continuously differentiable f. If $\nabla_j f$ has a continuous Hessian, we further have $\gamma_k = O(\|x_k - x^*\|_2)$. (For the derivations of the convergence rate of γ_k, see the proofs of Lemmas 12 and 14.)

The non-convexity of the objective function hinders the attainment of global guarantees for SCI-PI. While Theorem 6 establishes a condition on the initial iterate that guarantees local convergence, finding an initial point sufficiently close to a global optimal solution is a challenging task. To ensure global guarantees, additional conditions on problem-specific parameters are necessary. For example, mild assumptions on problem parameters in affine phase retrieval render the objective function strongly convex, leading to global optimality (Huang and Xu, 2022). Nevertheless, local convergence remains an appealing property since random initialization often provides a satisfactory starting point (Chen et al., 2019).

Reduction to power iteration For the leading eigenvector problem to find the leading eigenvector of a positive semi-definite matrix $A \succeq 0$, the objective function is $f(x) = \frac{1}{2}x^T Ax$, and thus SCI-PI specializes to power iteration. Let λ_i be the i^{th} largest eigenvalue of A. The condition $\lambda^* > \bar{\lambda}_2$ in Theorem 6 is interpreted as the positive eigen-gap $\lambda_1 - \lambda_2 > 0$ assumption. The convergence result in Theorem 6 not only matches the convergence rate of λ_1/λ_2 but also restores the initial condition $\delta < \sqrt{2}$ or $x_0^T x^* > 0$ of power iteration since $M = 0$.

Comparision to generalized power method Under the spherical constraint, generalized power method (Journée et al., 2010) has the same update rule as SCI-PI. Generalized power method has shown to attain sublinear convergence for convex f. Local linear convergence in Theorem 6 has been not shown for generalized power method. This convergence result established for a scale invariant objective function with the spherical constraint is extended to various settings in the next section.

4. Extended Settings

4.1 Sum of Scale Invariant Functions

Consider a sum of scale invariant functions of the form $f(x) = \sum_{i=1}^{m} f_i^M(x) + \sum_{j=1}^{n} f_j^A(x)$ where f_i^M is multiplicatively scale invariant with $u_i(c) = |c|^{p_i}$ and f_j^A is additively scale invariant with $v_j(c) = \log_{a_j}|c|$. Note that this does not imply that f is scale invariant in general. Thus, a stationary point x^* satisfying $\nabla f(x^*) = \lambda^* x^*$ is not necessarily an eigenvector of $\nabla^2 f(x^*)$. Instead, a stationary point x^* is an eigenvector of $F(x^*)$ defined as

$$F(x) = \sum_{i=1}^{m} \left(\frac{1}{p_i - 1}\right) \nabla^2 f_i^M(x) - \sum_{j=1}^{n} \nabla^2 f_j^A(x).$$

since

$$\nabla f(x) = \sum_{i=1}^{m} \nabla f_i^M(x) + \sum_{j=1}^{n} \nabla f_j^A(x) = F(x)x$$

by Proposition 3. Here is an example that involves a sum of scale invariant functions.

Example 3 (Kurtosis-based ICA) *Given a pre-processed data matrix $W \in \mathbb{R}^{n\times d}$, Kurtosis-based ICA (Hyvärinen and Oja, 2000) solves*

$$\text{maximize} \quad \frac{1}{n}\sum_{i=1}^{n}\left[(w_i^T x)^4 - 3\right]^2 \quad \text{subject to} \quad x \in \partial\mathcal{B}_d. \tag{26}$$

The objective function f is a sum of scale invariant functions.

We present a local convergence analysis of SCI-PI for a sum of scale invariant functions as follows.

Theorem 7 *Let f be a sum of scale invariant functions and twice continuously differentiable on an open set containing $\partial\mathcal{B}_d$. Let x^* be a local maximum such that $\nabla f(x^*) = \lambda^* x^*$ and $\{\mathbf{v}_1, \cdots, \mathbf{v}_d\}$ be a set of eigenvectors of $F(x^*)$ with $x^* = \mathbf{v}_1$. If $\lambda^* > \bar{\lambda}_2 = \|\nabla^2 f(x^*)(I - x^*(x^*)^T)\|_2$, then there exists some $\delta > 0$ such that under the initial condition $\|x_0 - x^*\|_2 < \delta$, the sequence of iterates $\{x_k\}_{k=0,1,\cdots}$ generated by SCI-PI satisfies*

$$\|x_{k+1} - x^*\|_2^2 \leq \prod_{t=0}^{k-1}\left(\frac{\bar{\lambda}_2}{\lambda^*} + \gamma_t\right)^2 \|x_0 - x^*\|_2^2,$$

where

$$\frac{\bar{\lambda}_2}{\lambda^*} + \gamma_t < 1 \text{ for all } t \geq 0 \text{ and } \lim_{k\to\infty} \gamma_k = 0.$$

Moreover, if $\nabla_j f = \partial f/\partial x_j$ has a continuous Hessian H_j on an open set containing $\mathcal{B}_d$, we can explicitly write δ as

$$\delta(\lambda^*, \bar{\lambda}_1, \bar{\lambda}_2, M) = \frac{\sqrt{2}(\lambda^* - \bar{\lambda}_2)}{\lambda^* + M + \bar{\lambda}_1 + |\lambda^* - M|}$$

where $\bar{\lambda}_1 = \sqrt{2} \cdot \|\nabla^2 f(x^*)x^*\|_2$ and

$$M = \max_{x\in\partial\mathcal{B}_d, y^1,\cdots,y^d\in\mathcal{B}_d} \sqrt{\textstyle\sum_{i=1}^d (x^T G_i(y^1,\cdots,y^d)x)^2}, \quad G_i(y^1,\cdots,y^d) = \textstyle\sum_{j=1}^d \mathbf{v}_{i,j} H_j(y^j).$$

Proof By the stationarity condition, a local optimal solution x^* is an eigenvector of $F(x^*)$. Since $F(x^*)$ is real and symmetric, without loss of generality, we assume that $\{\mathbf{v}_1, \cdots, \mathbf{v}_d\}$ form an orthogonal basis in $\mathbb{R}^d$.

Since f is twice continuously differentiable on an open set containing $\partial\mathcal{B}_d$, for $x \in \partial\mathcal{B}_d$, using the Taylor expansion of $\nabla f(x)^T\mathbf{v}_i$ at x^*, we have

$$\nabla f(x)^T\mathbf{v}_i = \nabla f(x^*)^T\mathbf{v}_i + (x - x^*)^T\nabla^2 f(x^*)\mathbf{v}_i + R_i(x) \tag{27}$$

where

$$R_i(x) = o(\|x - x^*\|_2). \tag{28}$$

Using (27) with $i = 1$ and $\nabla f(x^*) = \lambda^* x^*$, we obtain

$$\nabla f(x)^T\mathbf{v}_1 = \lambda^*(x^*)^T\mathbf{v}_1 + (x - x^*)^T\nabla^2 f(x^*)\mathbf{v}_1 + R_1(x) = \lambda^* + \alpha(x) \tag{29}$$

where

$$\alpha(x) = (x - x^*)^T\nabla^2 f(x^*)\mathbf{v}_1 + R_1(x). \tag{30}$$

Using (27) and $\nabla f(x^*) = \lambda^* x^*$ for $2 \leq i \leq d$, we have

$$\nabla f(x)^T\mathbf{v}_i = \lambda^*(x^*)^T\mathbf{v}_i + (x - x^*)^T\nabla^2 f(x^*)\mathbf{v}_i + R_i(x) = (x - x^*)^T\nabla^2 f(x^*)\mathbf{v}_i + R_i(x),$$

resulting in

$$\sum_{i=2}^d (\nabla f(x)^T\mathbf{v}_i)^2 = \sum_{i=2}^d \left((x - x^*)^T\nabla^2 f(x^*)\mathbf{v}_i + R_i(x)\right)^2. \tag{31}$$

From $x^* = \mathbf{v}_1$ and the fact that $\{\mathbf{v}_1, \cdots, \mathbf{v}_d\}$ forms an orthogonal basis in $\mathbb{R}^d$, we have

$$\begin{aligned}
\sum_{i=2}^d \left((x - x^*)^T\nabla^2 f(x^*)\mathbf{v}_i\right)^2 &= \|\nabla^2 f(x^*)(x - x^*)\|_2^2 - \left((x - x^*)^T\nabla^2 f(x^*)\mathbf{v}_1\right)^2 \\
&= (x - x^*)^T\nabla^2 f(x^*)\left(I - x^*(x^*)^T\right)\nabla^2 f(x^*)(x - x^*) \\
&= (x - x^*)^T\nabla^2 f(x^*)\left(I - x^*(x^*)^T\right)^2\nabla^2 f(x^*)(x - x^*).
\end{aligned}$$

Since

$$\begin{aligned}\|\nabla^2 f(x^*)\left(I - x^*(x^*)^T\right)^2 \nabla^2 f(x^*)\|_2 &= \|\left(I - x^*(x^*)^T\right)\nabla^2 f(x^*)\|_2^2 \\ &= \|\nabla^2 f(x^*)\left(I - x^*(x^*)^T\right)\|_2^2,\end{aligned}$$

we have

$$\sum_{i=2}^{d}\left((x-x^*)^T\nabla^2 f(x^*)\mathbf{v}_i\right)^2 \leq \bar{\lambda}_2^2\|x-x^*\|_2^2. \tag{32}$$

Also, using (32) and the Cauchy-Schwartz inequality, we obtain

$$\begin{aligned}\sum_{i=2}^{d}(x-x^*)^T\nabla^2 f(x^*)\mathbf{v}_i R_i(x) &\leq \sum_{i=2}^{d}|(x-x^*)^T\nabla^2 f(x^*)\mathbf{v}_i||R_i(x)| \\ &\leq \sqrt{\sum_{i=2}^{d}\left((x-x^*)^T\nabla^2 f(x^*)\mathbf{v}_i\right)^2}\sqrt{\sum_{i=2}^{d}R_i(x)^2} \\ &\leq \bar{\lambda}_2\|x-x^*\|_2\sqrt{\sum_{i=2}^{d}R_i(x)^2}.\end{aligned} \tag{33}$$

Using (32) and (33) for (31), we obtain

$$\sum_{i=2}^{d}(\nabla f(x)^T\mathbf{v}_i)^2 \leq \bar{\lambda}_2^2\|x-x^*\|_2^2 + 2\bar{\lambda}_2\|x-x^*\|_2\sqrt{\sum_{i=2}^{d}R_i(x)^2} + \sum_{i=2}^{d}R_i(x)^2,$$

resulting in

$$\sum_{i=2}^{d}(\nabla f(x)^T\mathbf{v}_i)^2 \leq \left(\bar{\lambda}_2\|x-x^*\|_2^2 + \beta(x)\right)^2 \tag{34}$$

where

$$\beta(x) = \sqrt{\sum_{i=2}^{d}R_i(x)^2}. \tag{35}$$

Using $(x-x^*)^T\nabla^2 f(x^*)\mathbf{v}_1 = o(\sqrt{\|x-x^*\|_2})$ and (28) for (30) and (35), we have

$$\alpha(x) = o(\sqrt{\|x-x^*\|_2}), \quad \beta(x) = o(\|x-x^*\|_2). \tag{36}$$

By (29), (34), (36), and Lemma 12, we obtain the first part of the desired result.

On the other hand, if $\nabla_j f$ has a continuous Hessian H_j, by Lemma 14, we have

$$\sum_{i=1}^{d}(R_i(x))^2 \leq \frac{1}{4}M^2\|x-x^*\|_2^4. \tag{37}$$

Using $|(x-x^*)^T\nabla^2 f(x^*)\mathbf{v}_1| \le \bar{\lambda}_1\sqrt{1-x_k^T x^*}$ and (37) for (30) and (35), this leads to

$$\begin{aligned}\alpha(x) &= (x-x^*)^T\nabla^2 f(x^*)\mathbf{v}_1 + R_1(x) \ge -\bar{\lambda}_1\sqrt{1-x_k^T x^*} - M(1-x_k^T x^*),\\ \beta(x) &= \sqrt{\sum_{i=2}^{d} R_i(x)^2} \le \frac{M}{2}\|x-x^*\|_2^2.\end{aligned} \tag{38}$$

By (29), (34), (38), and Lemma 13 with

$$A=\lambda^*,\ B=M,\ C=\bar{\lambda}_1,\ D=0,\ E=\bar{\lambda}_2,\ F=M,$$

we obtain

$$\begin{aligned}\delta(\lambda^*,\bar{\lambda}_1,\bar{\lambda}_2,M) &= \min\left\{\frac{\sqrt{2}\lambda^*}{\lambda^*+|\lambda^*-M|+\bar{\lambda}_1},\ \frac{\sqrt{2}(\lambda^*-\bar{\lambda}_2)}{\lambda^*+M+\bar{\lambda}_1+|\lambda^*-M|}\right\}\\ &= \frac{\sqrt{2}(\lambda^*-\bar{\lambda}_2)}{\lambda^*+M+\bar{\lambda}_1+|\lambda^*-M|},\end{aligned}$$

which completes the proof. ■

Note that $\bar{\lambda}_1$ has the additional $\sqrt{2}$ factor which comes from the fact that x^* is not necessarily an eigenvector of $\nabla^2 f(x^*)$. Nonetheless, the asymptotic convergence rate in Theorem 7 provides a generalization of the convergence rate in Theorem 6.

4.2 Block Scale Invariant Problems

Next, consider a class of optimization problems having the form of

$$\text{maximize}\quad f(x,y)\quad \text{subject to}\quad x\in\partial\mathcal{B}_{d_x},\ y\in\partial\mathcal{B}_{d_y}$$

where $f:\mathbb{R}^{d_x+d_y}\to\mathbb{R}$ is scale invariant in x for fixed y and vice versa. A stationary point $(x^*,\ y^*)$ satisfies $\nabla_x f(x^*,y^*)=\lambda^* x^*$ and $\nabla_y f(x^*,y^*)=s^* y^*$ for some $\lambda^*, s^*\in\mathbb{R}$, and x^* and y^* are eigenvectors of $\nabla^2_{xx} f(x^*,y^*)$ and $\nabla^2_{yy} f(x^*,y^*)$, respectively, according to Proposition 4. Some examples of block scale invariant problems are given next.

Example 4 (Semidefinite Programming (SDP) (Vandenberghe and Boyd, 1996)) *Let $A, X\in\mathbb{R}^{n\times n}$. Given an SDP problem*

$$\text{maximize}\quad \langle A,X\rangle\quad \text{subject to}\quad X_{ii}=1,\ i\in\{1,2,\cdots,n\},\ X\succeq 0,$$

the Burer-Monteiro approach (Burer and Monteiro, 2003) yields the following block scale invariant problem (Erdogdu et al., 2022)

$$\text{maximize}\quad \langle A,\sigma\sigma^T\rangle\quad \text{subject to}\quad \|\sigma_i\|_2=1,\ i\in\{1,2,\cdots,n\}$$

where σ_i denotes the i^{th} row of $\sigma\in\mathbb{R}^{n\times r}$.

Example 5 (Kullback-Leibler (KL) divergence NMF) *The KL-NMF problem (Févotte and Idier, 2011; Lee and Seung, 2001; Wang and Zhang, 2013) is defined as*

$$\begin{aligned} \text{minimize} \quad & D_{KL}(V\|WH) \triangleq \sum_{i=1}^{n}\sum_{j=1}^{m}\left[V_{ij}\log\frac{V_{ij}}{\sum_{q=1}^{r}W_{iq}H_{qj}} - V_{ij} + \sum_{q=1}^{r}W_{iq}H_{qj}\right] \\ \text{subject to} \quad & W_{iq}\geq 0,\ H_{qj}\geq 0,\ i=1,\cdots,n,\ j=1,\cdots,m,\ q=1,\cdots,r. \end{aligned} \tag{39}$$

Many popular algorithms (Lee and Seung, 2001; Lin, 2007) for the KL-NMF problem are based on alternating minimization of W and H. Since the objective function can be decomposed over j, given $W \geq 0$ and $j \in \{1,\cdots,m\}$, we consider a subproblem of the form

$$\text{minimize} \quad f_{KL}^{j}(h) = \sum_{i=1}^{n}\left[V_{ij}\log\frac{V_{ij}}{\sum_{q=1}^{r}W_{iq}h_j} - V_{ij} + \sum_{q=1}^{r}W_{iq}h_q\right] \ \text{ subject to } h_q \geq 0 \tag{40}$$

where $h_q = H_{qj}$. Note that the KL-NMF problem in the form of (39) is not a block scale invariant problem. However, using a novel reformulation, we show that the KL divergence NMF subproblem is indeed a scale invariant problem.

Lemma 8 *The KL-NMF subproblem* (40) *is equivalent to the following scale invariant problem*

$$\text{maximize} \quad -\sum_{i=1}^{n}V_{ij}\log\sum_{q=1}^{r}W_{iq}\bar{h}_q \quad \text{subject to} \quad \sum_{q=1}^{r}\bar{h}_q = 1, \quad \bar{h}_q \geq 0, \tag{41}$$

with the relationship $(\sum_{i=1}^{n}V_{ij})\bar{h}_q = (\sum_{i=1}^{n}W_{iq})h_q$.

Proof Since a log-linear function is concave, (40) is a convex problem in h. Consider the Lagrangian of the original problem

$$\mathcal{L}(h,\lambda) = f_{KL}^{j}(h) - \sum_{q=1}^{r}\lambda_q h_q, \quad \lambda \geq 0. \tag{42}$$

Let h^* be an optimal solution to (40) and λ^* be a vector in $\mathbb{R}^r$ satisfying the following first-order KKT conditions

$$\nabla f_{KL}^{j}(h^*) = \lambda_q^* \mathbb{1}_m, \quad \lambda_q^* h_q^* = 0, \quad q = 1,\cdots,r \tag{43}$$

where ∇f_{KL}^{j} denotes the derivative of f_{KL}^{j} with respect to h.

Since (43) implies $\sum_{q=1}^{r}h_q^*\lambda_q^* = 0$, we have

$$\sum_{q=1}^{r}h_q^*\lambda_q^* = \sum_{q=1}^{r}h_q^*\nabla f_{KL}^{j}(h^*) = -\sum_{i=1}^{n}\sum_{q=1}^{r}\frac{V_{ij}W_{iq}h_q^*}{\sum_{q'=1}^{r}W_{iq'}h_{q'}^*} + \sum_{i=1}^{n}\sum_{q=1}^{r}W_{iq}h_q^*,$$

resulting in

$$\sum_{i=1}^{n}V_{ij} = \sum_{i=1}^{n}\sum_{q=1}^{r}W_{iq}h_q^*. \tag{44}$$

Next, let

$$\text{minimize } f^j_{SCI}(h) = \sum_{i=1}^{n} V_{ij} \log \frac{V_{ij}}{\sum_{q=1}^{r} W_{iq} h_q} \text{ subject to } \sum_{i=1}^{n} V_{ij} = \sum_{i=1}^{n} \sum_{q=1}^{r} W_{iq} h_q,\ h_q \geq 0, \tag{45}$$

and let f^*_{KL} and f^*_{SCI} be the optimal objective values of (40) and (45), respectively. We prove the equivalence of (45) and (40) by the following arguments:

1. Since (45) has an additional constraint $\sum_{i=1}^{n} V_{ij} = \sum_{i=1}^{n} \sum_{q=1}^{r} W_{iq} h_q$ compared to (40), it always satisfies $f^*_{SCI} \geq f^*_{KL}$.

2. Since we have shown that $\sum_{i=1}^{n} V_{ij} = \sum_{i=1}^{n} \sum_{q=1}^{r} W_{iq} h^*_q$, a solution h^* of (40) is a feasible point of (45). This implies $f^*_{KL} \geq f^*_{SCI}$.

Now, we reparameterize h by $\bar{h}$ so that $\sum_{i=1}^{n} V_{ij} = \sum_{i=1}^{n} \sum_{q=1}^{r} W_{iq} h_q$ if and only if $\sum_{q=1}^{r} \bar{h}_q = 1$, which yields the relationship between two variables $\bar{h}_q = h_q(\sum_{i=1}^{n} W_{iq})/(\sum_{i=1}^{n} V_{ij})$. Note that (41) is a mixture proportion estimation problem (Example 2) and thus a scale invariant problem. ■

To solve block scale invariant problems, we consider an alternating maximization algorithm called *block SCI-PI*, which repeats

$$x_{k+1} \leftarrow \nabla_x f(x_k, y_k)/\|\nabla_x f(x_k, y_k)\|_2, \quad y_{k+1} \leftarrow \nabla_y f(x_k, y_k)/\|\nabla_y f(x_k, y_k)\|_2. \tag{46}$$

We present a local convergence result of block SCI-PI below.

Theorem 9 *Suppose that f is twice continuously differentiable on an open set containing $\partial\mathcal{B}_{d_x} \times \partial\mathcal{B}_{d_y}$ and let (x^*, y^*) be a local maximum satisfying*

$$\nabla_x f(x^*, y^*) = \lambda^* x^*,\ \lambda^* > \bar{\lambda}_2 = \max_{2 \leq i \leq d_x} |\lambda_i|,\ \nabla_y f(x^*, y^*) = s^* y^*,\ s^* > \bar{s}_2 = \max_{2 \leq j \leq d_y} |s_j|$$

where $(\lambda_i, \mathbf{v}_i)$ and $(s_j, \mathbf{u}_j)$ are eigen-pairs of $\nabla^2_{xx} f(x^*, y^*)$ and $\nabla^2_{yy} f(x^*, y^*)$, respectively with $x^* = \mathbf{v}_1$ and $y^* = \mathbf{u}_1$. If

$$\nu^2 = \|\nabla^2_{yx} f(x^*, y^*)\|_2^2 < (\lambda^* - \bar{\lambda}_2)(s^* - \bar{s}_2),$$

then for the sequence of iterates $\{(x_k, y_k)\}_{k=0,1,\cdots}$ generated by (46), there exists some $\delta > 0$ such that if $\Delta_0 < \delta$, then we have

$$\Delta_k^2 \leq \prod_{t=0}^{k-1} (\rho + \gamma_t)^2 \Delta_0^2 \text{ and } \lim_{k \to \infty} \gamma_k = 0$$

where

$$\Delta_k = \left\| \begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix} \right\|_2, \quad \rho = \frac{1}{2} \left[\frac{\bar{\lambda}_2}{\lambda^*} + \frac{\bar{s}_2}{s^*} + \sqrt{\left[\frac{\bar{\lambda}_2}{\lambda^*} - \frac{\bar{s}_2}{s^*} \right]^2 + \frac{4\nu^2}{\lambda^* s^*}} \right] < 1.$$

Proof From Lemma 15 with $x = x_k$, $y = y_k$, we have

$$\frac{\sum_{i=2}^{d_x}(\nabla_x f(x_k, y_k)^T \mathbf{v}_i)^2}{(\nabla_x f(x_k, y_k)^T \mathbf{v}_1)^2} \leq \left(\frac{\bar{\lambda}_2}{\lambda^*}\|x_k - x^*\|_2 + \frac{\nu}{\lambda^*}\|y_k - y^*\|_2 + \theta^x(x_k, y_k)\right)^2.$$

Since

$$x_{k+1} = \frac{\nabla_x f(x_k, y_k)}{\|\nabla_x f(x_k, y_k)\|_2},$$

we obtain

$$\|x_{k+1} - x^*\|_2 \leq \sqrt{\frac{\sum_{i=2}^{d_x}(\nabla_x f(x_k, y_k)^T \mathbf{v}_i)^2}{(\nabla_x f(x_k, y_k)^T \mathbf{v}_1)^2}} \leq \frac{\bar{\lambda}_2}{\lambda^*}\|x_k - x^*\|_2 + \frac{\nu}{\lambda^*}\|y_k - y^*\|_2 + \theta^x(x_k, y_k). \tag{47}$$

Using Lemma 15 for $x = y_k$, $y = x_k$ and the definition of y_{k+1}, we have

$$\|y_{k+1} - y^*\|_2 \leq \frac{\nu}{s^*}\|x_k - x^*\|_2 + \frac{\bar{s}_2}{s^*}\|y_k - y^*\|_2 + \theta^y(x_k, y_k). \tag{48}$$

Combining (47) and (48), we obtain

$$\begin{bmatrix}\|x_{k+1} - x^*\|_2 \\ \|y_{k+1} - y^*\|_2\end{bmatrix} \leq \begin{bmatrix}\frac{\bar{\lambda}_2}{\lambda^*} & \frac{\nu}{\lambda^*} \\ \frac{\nu}{s^*} & \frac{\bar{s}_2}{s^*}\end{bmatrix} \begin{bmatrix}\|x_k - x^*\|_2 \\ \|y_k - y^*\|_2\end{bmatrix} + \begin{bmatrix}\theta^x(x_k, y_k) \\ \theta^y(x_k, y_k)\end{bmatrix}. \tag{49}$$

Since $\rho < 1$ due to $\nu^2 < (\lambda^* - \bar{\lambda}_2)(s^* - \bar{s}_2)$, by Lemma 17, we obtain the desired result. ■

Being the spectral norm of the off-diagonal block of the Hessian at the local maximum (x^*, y^*), ν measures how much partial derivatives of one block of variables are affected by the other block of variables. If the objective function f is separable in x and y as in the case of the KL-NMF problem, ν becomes zero, and we have $\rho = \max\{\bar{\lambda}_2/\lambda^*, \bar{s}_2/s^*\}$. Note that ρ increases as ν increases. If ν^2 becomes larger than $(\lambda^* - \bar{\lambda}_2)(s^* - \bar{s}_2)$, the Jacobi update rule (46) may fail due to the interaction effects between x and y. On the other hand, the result of Theorem 6 can be restored by dropping x or y in Theorem 9. While we consider the two-block case here, the algorithm and the convergence analysis can be easily generalized to more than two blocks.

4.3 Partially Scale Invariant Problems

Lastly, we consider a class of optimization problems of the form

$$\text{maximize} \quad f(x, y) \quad \text{subject to} \quad x \in \partial\mathcal{B}_{d_x}$$

where $f(x, y) : \mathbb{R}^{d_x + d_y} \to \mathbb{R}$ is a scale invariant function in x for each $y \in \mathbb{R}^{d_y}$. A partially scale invariant problem has the form of (1) with respect to x once y is fixed. If x is fixed, we obtain an unconstrained optimization problem with respect to y. A stationary point (x^*, y^*) satisfies $\nabla_x f(x^*, y^*) = \lambda^* x^*$ and $\nabla_y f(x^*, y^*) = 0$ for some $\lambda^* \in \mathbb{R}$, and x^* is an eigenvector of $\nabla^2_{xx} f(x^*, y^*)$. Here is an example of partially scale invariant problems.

Example 6 (Gaussian Mixture Model (GMM)) *The GMM problem is defined as*

$$\text{maximize} \quad \sum_{i=1}^{n} \log \sum_{j=1}^{d} \pi_j \, \mathcal{N}(a_i; \mu_j, \Sigma_j) \quad \text{subject to} \quad \pi \in \mathcal{S}^d.$$

Note that the objective function is scale invariant for fixed μ_j and Σ_j, and μ_j is unconstrained. If we assume some structure on Σ_j, estimation of Σ_j can also be unconstrained. For general Σ_j, semi-positive definiteness is necessary for Σ_j.

To solve partially scale invariant problems, we consider an alternating maximization algorithm based on SCI-PI and the gradient method as

$$x_{k+1} \leftarrow \nabla_x f(x_k, y_k)/\|\nabla_x f(x_k, y_k)\|_2, \quad y_{k+1} \leftarrow y_k + \alpha \nabla_y f(x_k, y_k). \tag{50}$$

While the gradient method is used in (50), any method for unconstrained optimization can replace it. We present a convergence analysis of (50) below.

Theorem 10 *Suppose that $f(x, y)$ is scale invariant in x for each $y \in \mathbb{R}^{d_y}$, μ-strongly concave in y with an L-Lipschitz continuous $\nabla_y f(x, y)$ for each $x \in \partial\mathcal{B}_{d_x}$, and three-times continuously differentiable on an open set containing $\partial\mathcal{B}_{d_x} \times \mathbb{R}^{d_y}$. Let (x^*, y^*) be a local maximum satisfying*

$$\nabla f(x^*) = \lambda^* x^*, \ \lambda^* > \bar{\lambda}_2 = \max_{2 \le i \le d_x} |\lambda_i|$$

where $(\lambda_i, \mathbf{v}_i)$ is an eigen-pair of $\nabla^2_{xx} f(x^*, y^*)$ with $x^* = \mathbf{v}_1$. If

$$\nu^2 = \|\nabla^2_{yx} f(x^*, y^*)\|_2^2 < \mu(\lambda^* - \bar{\lambda}_2),$$

then for the sequence of iterates $\{(x_k, y_k)\}_{k=0,1,\cdots}$ generated by (50) with $\alpha = 2/(L+\mu)$, there exists some $\delta > 0$ such that if $\Delta_0 < \delta$, then we have

$$\Delta_k^2 \le \textstyle\prod_{t=0}^{k-1} (\rho + \gamma_t)^2 \, \Delta_0^2 \text{ and } \lim_{k\to\infty} \gamma_k = 0$$

where

$$\Delta_k = \left\| \begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix} \right\|_2, \quad \rho = \frac{1}{2} \left[\frac{\bar{\lambda}_2}{\lambda^*} + \frac{L-\mu}{L+\mu} + \sqrt{\left[\frac{\bar{\lambda}_2}{\lambda^*} - \frac{L-\mu}{L+\mu} \right]^2 + \frac{8\nu^2}{\lambda^*(L+\mu)}} \right] < 1.$$

Proof Using Lemma 15 for $x = x_k$, $y = y_k$ and the definition of x_{k+1}, we have

$$\|x_{k+1} - x^*\|_2 \le \frac{\bar{\lambda}_2}{\lambda^*} \|x_k - x^*\|_2 + \frac{\nu}{\lambda^*} \|y_k - y^*\|_2 + \theta^x(x_k, y_k). \tag{51}$$

By Lemma 16 with $x = x_k, y = y_k$, we also have

$$\|y_{k+1} - y^*\|_2 \le \left(\frac{2\nu}{L+\mu} \right) \|x_k - x^*\|_2 + \left(\frac{L-\mu}{L+\mu} \right) \|y_k - y^*\|_2 + \theta^y(x_k, y_k). \tag{52}$$

Combining (51) and (52), we obtain

$$\begin{bmatrix} \|x_{k+1} - x^*\|_2 \\ \|y_{k+1} - y^*\|_2 \end{bmatrix} \leq \begin{bmatrix} \frac{\bar{\lambda}_2}{\lambda^*} & \frac{\nu}{\lambda^*} \\ \frac{2\nu}{L+\mu} & \frac{L-\mu}{L+\mu} \end{bmatrix} \begin{bmatrix} \|x_k - x^*\|_2 \\ \|y_k - y^*\|_2 \end{bmatrix} + \begin{bmatrix} \theta^x(x_k, y_k) \\ \theta^y(x_k, y_k) \end{bmatrix}. \tag{53}$$

Note that since $\nu^2 < \mu(\lambda^* - \bar{\lambda}_2)$, the spectral radius ρ satisfies

$$\rho = \frac{1}{2}\left(\frac{\bar{\lambda}_2}{\lambda^*} + \frac{L-\mu}{L+\mu} + \sqrt{\left(\frac{\bar{\lambda}_2}{\lambda^*} - \frac{L-\mu}{L+\mu}\right)^2 + \frac{8\nu^2}{\lambda^*(L+\mu)}}\right) < 1.$$

Therefore, by Lemma 14, we obtain the desired result. ■

As in the result of Theorem 9, the rate ρ increases as ν increases and is equal to $\max\{\bar{\lambda}_2/\lambda^*, (L-\mu)/(L+\mu)\}$ when $\nu = 0$. Also, by dropping y, we can restore the convergence result of Theorem 6.

5. Numerical Experiments

We test the proposed algorithms on real-world data sets. All experiments are implemented on a standard laptop (2.6 GHz Intel Core i7 processor and 16GM memory) using the Julia programming language. Let us emphasize that scale invariant problems frequently appear in many important applications in statistics and machine learning. We select three important applications, KL-NMF, GMM and ICA. A description of the data sets is provided below and source codes are available at: `https://github.com/youngseok-kim/SCIPI-JMLR`.

5.1 Description of Data Sets

For KL-NMF (Section 5.2), we use four public real data sets available online* and summarized in Table 1. Waving Trees (WT) has 287 images, each having 160×120 pixels. KOS and NIPS are sparse, large matrices implemented for topic modeling. WIKI is a large binary matrix having values 0 or 1 representing the adjacency matrix of a directed graph. Here, sparsity represents the fraction of zero elements in a matrix.

Name	# of samples	# of features	# of nonzeros	Sparsity
WIKI	8,274	8,297	104,000	0.999
NIPS	1,500	12,419	280,000	0.985
KOS	3,430	6,906	950,000	0.960
WT	287	19,200	5,510,000	0.000

Table 1: A brief summary of data sets used for KL-NMF

*These four data sets are retrieved from `https://www.microsoft.com/en-us/research/project`, `https://archive.ics.uci.edu/ml/datasets/bag+of+words`, and `https://snap.stanford.edu/data/wiki-Vote.html`

Name	# of classes	# of samples	Dimension
Sonar	2	208	60
Ionosphere	2	351	34
HouseVotes84	2	435	16
BrCancer	2	699	10
PIDiabetes	2	768	8
Vehicle	4	846	18
Glass	6	214	9
Zoo	7	101	16
Vowel	11	990	10
Servo	51	167	4

Table 2: A brief summary of data sets used for GMM

Name	# of samples	# of features
Wine	178	14
Soybean	683	35
Vehicel	846	18
Vowel	990	10
Cardio	2,126	22
Satellite	6,435	37
Pendigits	10,992	17
Letter	20,000	16
Shuttle	58,000	9

Table 3: A brief summary of data sets used for ICA

For GMM (Section 5.3), we use ten public real data sets, corresponding to all small and moderate data sets provided by the `mlbench` package in R. We select data sets for multi-class classification problems and run EM and SCI-PI for the given number of classes without class labels. In Table 2, the sample size varies from 101 to 990, the dimension varies from 2 to 60, and the number of classes varies from 2 to 51. In these data sets, only a small portion of entries are missing. If missing entries exist, we impute them with the means.

For ICA, discussed also in Section 5.3, we use nine public data sets (see Table 3) from the UCI Machine Learning repository[†]. The sample size varies from 178 to 58,000 and the dimension varies from 9 to 37.

5.2 KL-divergence Nonnegative Matrix Factorization

We perform experiments on the KL-NMF problem (39) described in Example 5. Let us recall that the original KL-NMF problem can be solved via block SCI-PI where in each iteration the algorithm solves the subproblem of the form (41). Our focus is to compare this algorithm with other well-known alternating minimization algorithms listed below, updating H and W alternatively. We let $z = V \oslash (Wh)$.

- Projected gradient descent (PGD): It iterates $h^{\text{new}} \leftarrow h - \eta \odot W^T(z - \mathbb{1}_n)$ followed by projection onto the simplex, where $\eta \propto h$ is an appropriate learning rate (Lin, 2007).

[†] `https://archive.ics.uci.edu/ml/index.php`

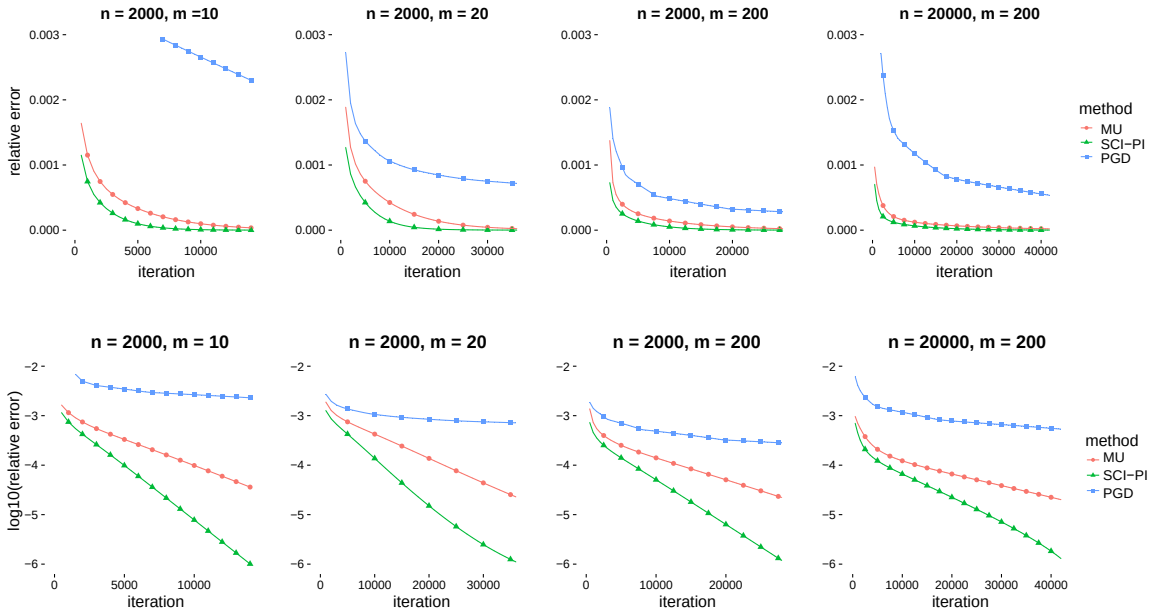

Figure 2: Convergence of the three algorithms for the KL-NMF subproblem; the relative error $|f_k - f^*|/|f_0 - f^*|$ *(Top)* and the log relative error *(Bottom)*; n/m: the number of samples/features of the data matrix

- Multiplicative update (MU): A famous multiplicative update algorithm is originally suggested by (Lee and Seung, 2001), which iterates $h^{\text{new}} \leftarrow h \odot (W^T z) \oslash (W^T \mathbb{1}_n)$ and is learning rate free.

- Our method (SCI-PI): It iterates $h^{\text{new}} \leftarrow h \odot (\sigma + W^T z)^{\odot 2}$ followed by rescaling, where σ is a shift parameter. We simply use $\sigma = 1$ for preconditioning.

- Sequential quadratic programming (MIXSQP): It exactly solves each subproblem via a convex solver `mixsqp` (Kim et al., 2020). This algorithm performs sequential non-negative least squares.

KL-NMF subproblem on synthetic data sets Before presenting experimental results of alternating algorithms on the KL-NMF problem (39), we report small experiments using synthetic data sets on the KL-NMF subproblem (40) where we repeat the above iterations until convergence.

To study the convergence rate for the KL-NMF subproblems, we use the four data sets studied in Kim et al. (2020). We study MU, PGD and SCI-PI since they have the same order of computational complexity per iteration, but omit MIXSQP since it is a second-order method which cannot be directly compared. For PGD, the learning rate is optimized by grid search. The stopping criterion is $\|f_k - f^*\|_2 \leq 10^{-6} f^*$ where f_k is the objective value at iteration k and f^* is the solution obtained by MIXSQP after extensive computation time.

The result is shown in Figure 2[‡]. The average runtime for aforementioned three methods are 33, 33 and 30 seconds for 10,000 iterations, respectively. Although the reformulated scale invariant problem (12) is a non-convex problem, SCI-PI always finds a global optimal solution, regardless of the starting point. Moreover, as shown in the figure, SCI-PI outperforms the other two algorithms for all simulated data sets.

KL-NMF on real world data sets Next, we test the four algorithms on the data sets in Table 1. We estimate $r = 20$ factors. At each iteration, all four algorithms solve m subproblems simultaneously for W and then alternatively for H.

The result is summarized in Figure 3[§]. The convergence plots are based on the average relative errors over ten repeated runs with random initializations. The result shows that SCI-PI is an overall winner, showing faster convergence rates. The stopping criterion is the same as above. To assess the overall performance when initialized differently, we select KOS and WIKI and run MU, PGD, SCI-PI, and MIXSQP ten times[‡]. The three algorithms except MIXSQP have (approximately) the same computational cost per iteration, take runtime of 391, 396, 408 seconds for KOS data and 372, 390, 418 seconds for WIKI data, respectively for 200 iterations. MIXSQP has a larger per iteration cost. After 400 seconds, SCI-PI achieves the lowest objective values in all cases but one for each data set (38 out of 40 in total). Thus it clearly outperforms other methods and also achieves the lowest variance. Unlike the other three algorithms, SCI-PI is not an ascent algorithm but an eigenvalue-based fixed-point algorithm. Admittedly, non-monotone convergence of SCI-PI can hurt reliability of the solution but for the KL-NMF problem its performance turns out to be stable.

5.3 Gaussian Mixture Model and Independent Component Analysis

In this subsection, we study the empirical performance of SCI-PI when it is applied to GMM and ICA.

GMM GMM fits a mixture of Gaussian distributions to the underlying data. Let $L_{ij} = \mathcal{N}(a_i; \mu_j, \Sigma_j)$ where i is the sample index and j the cluster index and let π be the actual mixture proportion vector. GMM fits into our restricted scale invariant setting (Section 4.3) with reparametrization, but the gradient update for μ_j, Σ_j is replaced by the exact coordinate ascent step. The EM and SCI-PI updates for π can be written respectively as

$$r = \mathbb{1}_n \oslash (L\pi), \quad \pi^{\text{new}} \propto \pi \odot (L^T r) \quad \text{(EM)}, \quad \pi^{\text{new}} \propto \pi \odot (\alpha + L^T r)^{\odot 2} \quad \text{(SCI-PI)} \quad (54)$$

where α is a shift parameter set to 1. We compare SCI-PI and EM for different real-world data sets from Table 2. All the algorithms initialize from the same standard Gaussian random variable, repeatedly for ten times. The result is summarized in the left panel in Figure 4. In some cases, SCI-PI achieves much larger objective values even if initialized the same. In many cases the two algorithms exhibit the same performance. This is because estimation of μ_k's and Σ_k's are usually harder than estimation of π, and EM and SCI-PI have the same

[‡]For each evaluation, we randomly draw ten initial points and report the averaged relative errors with respect to f^*. The initial input for the KL-NMF problem is a one-step MU update of a Uniform$(0, 1)$ random matrix.

[§]In all plots we do not show the first few iterations. The initial random solutions have the gap of approximately 50% which drops to a few percent after ten iterations where the plots start.

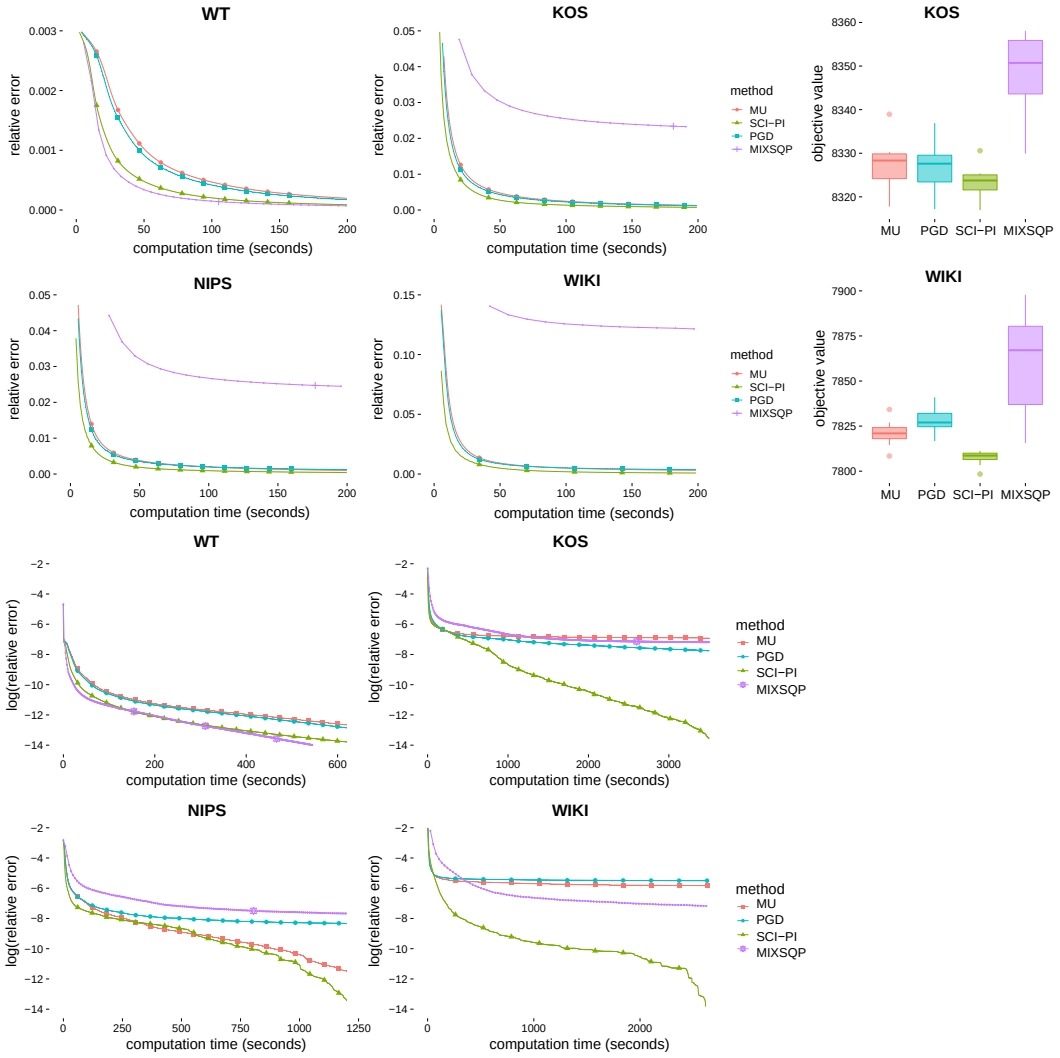

Figure 3: Convergence of the four NMF algorithms; the relative error $|f_k - f^*|/|f_0 - f^*|$ *(Top left)* and the log relative error *(Bottom)*; Boxplots containing ten objective values achieved after 400 seconds *(Top right)*

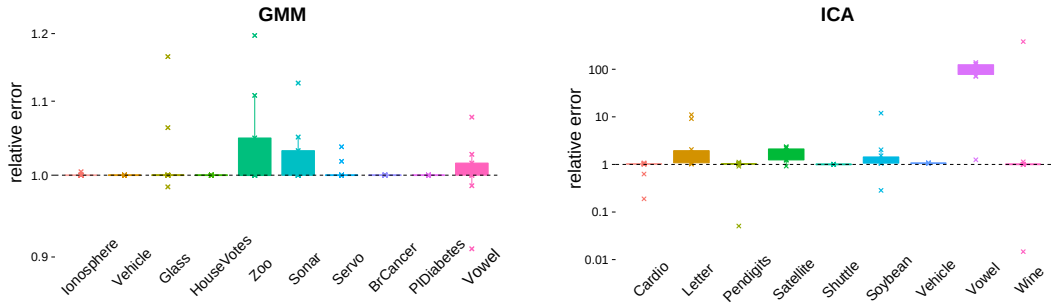

Figure 4: The relative objective $f^*_{\text{SCI-PI}}/f^*_{\text{EM}}$ for GMM *(Left)* and $f^*_{\text{SCI-PI}}/f^*_{\text{FastICA}}$ for ICA *(Right)*

updates for μ and Σ. For a few cases EM outperforms SCI-PI. Let us mention that SCI-PI and EM have the same order of computational complexity and require 591 and 590 seconds of total computation time, respectively.

ICA We implement SCI-PI on the Kurtosis-based ICA problem (Hyvärinen et al., 2004) and compare it with the benchmark algorithm FastICA (Hyvarinen, 1999), which is the most popular algorithm. Given a pre-processed[¶] data matrix $W \in \mathbb{R}^{n\times d}$, we maximize an approximation of negative entropy (Hyvärinen and Oja, 2000), $f(x) = \sum_{i=1}^{n} \left[(w_i^T x)^4 - 3\right]^2$, subject to $x \in \partial\mathcal{B}_d$. This problem fits into the sum of scale invariant setting (Section 4.1). SCI-PI iterates $x_{k+1} \leftarrow W^T[(Wx_k)^{\odot 4} - 3\mathbb{1}_n) \odot (Wx_k)^{\odot 3}]$ and FastICA iterates $x_{k+1} \leftarrow W^T(Wx_k)^{\odot 3} - 3(\mathbb{1}_n^T(Wx_k)^{\odot 2})x_k$, both followed by normalization.

In Figure 4 (right panel), we compare SCI-PI and FastICA on the data sets in Table 3. The majority of data points (81 out of 100 in total) show that SCI-PI tends to find a better solution with a larger objective value, but in a few cases SCI-PI converges to a sub-optimal point. Both algorithms are fixed-point based and thus have no guarantee of global convergence but overall SCI-PI outperforms FastICA. SCI-PI and FastICA have the same order of computational complexity and require 11 and 12 seconds of total computation time, respectively.

6. Final Remarks

In this paper, we propose a new class of optimization problems called the scale invariant problems, together with a generic solver SCI-PI, which is indeed an eigenvalue-based fixed-point iteration. We showed that SCI-PI directly generalizes power iteration and enjoys similar properties such as that SCI-PI has local linear convergence under mild conditions and its convergence rate is determined by eigenvalues of the Hessian matrix at a solution. Also, we extend scale invariant problems to problems with more general settings. We show by experiments that SCI-PI can be a competitive option for numerous important problems such as KL-NMF, GMM and ICA. Moreover, while not studied in this work, SCI-PI can be generalized to solve optimization problems on the Stiefel manifold such as block PCA. Under orthonormality constraints, the problem with a scale invariant function can be locally considered as the top k eigenvector problem. Therefore, we can develop a general form of the QR iteration (Francis, 1961, 1962) and its convergence analysis using similar proof techniques. Finding more examples and extending SCI-PI further to a more general setting is a promising direction for future studies.

[¶]A centered matrix $\widetilde{W} = n^{1/2}UDV^T$ is pre-processed by $W = \widetilde{W}VD^{-1}V^T$ so that $W^TW = nVV^T$.

Appendix A. Proofs of Propositions

A.1 Proof of Proposition 2

From (2) and (3), we infer functional equations of the multiplicative factor u and the additive factor v. Under the continuity assumption on f, these functional equations have the forms of Cauchy equations. Relying on known properties of Cauchy equations, we prove that u and v are homogeneous and log functions, respectively. We next provide details.

We first consider the multiplicative scale invariant case. Let x be a point such that $f(x) \neq 0$. Then, we have

$$f(rsx) = u(rs)f(x) = u(r)u(s)f(x),$$

which results in

$$u(rs) = u(r)u(s)$$

for all $r, s \in \mathbb{R}$. Let $g(r) = \ln(u(e^r))$. Then, we have

$$g(r+s) = \ln(u(e^{r+s})) = \ln(u(e^r e^s)) = \ln(u(e^r)) + \ln(u(e^s)) = g(r) + g(s),$$

which implies that g satisfies the first Cauchy functional equation. Since f is continuous, so is u and thus g. Therefore, by (Sahoo and Kannappan, 2011, pp. 81-82), we have

$$g(r) = rg(1) \tag{55}$$

for all $r \geq 0$. From the definition of g and (55), we have

$$u(e^r) = e^{g(r)} = (e^r)^{g(1)}. \tag{56}$$

Representing $r > 0$ as $r = e^{\ln(r)}$ and using (56), we obtain

$$u(r) = u\left(e^{\ln(r)}\right) = r^{g(1)} = r^{\ln(u(e))} = r^p.$$

Since $f(x) \neq 0$, if $p = \ln(u(e)) < 0$, then we have

$$\lim_{r\to 0_+} f(rx) = \lim_{r\to 0_+} u(r)f(x) = f(x) \cdot \lim_{r\to 0_+} r^p = f(x) \cdot \infty \neq f(0) < \infty,$$

contradicting the fact that f is continuous at 0. Also, if $p = 0$, then we get $u(r) = 1$, which contradicts $u(0) = 0$. Therefore, we must have $p > 0$. From u being an even function, we finally have

$$u(r) = |r|^p$$

for $r \in \mathbb{R}$.

Now, consider the additive scale invariant case. For any $x \in \mathrm{dom}(f)$, we have

$$f(rsx) = f(x) + v(rs) = f(x) + v(r) + v(s),$$

which results in

$$v(rs) = v(r) + v(s)$$

for all $r, s \in \mathbb{R} \setminus \{0\}$. Let $g(r) = v(e^r)$. Then, we have

$$g(r+s) = v(e^{r+s}) = v(e^r e^s) = v(e^r) + v(e^s) = g(r) + g(s).$$

Since g is continuous and satisfies the second Cauchy functional equation, by (Sahoo and Kannappan, 2011, pp. 83-84), we have

$$g(r) = rg(1)$$

for all $r \geq 0$. For $r > 0$, letting $r = e^{\ln(r)}$, we have

$$v(r) = v(e^{\ln(r)}) = g(\ln(r)) = g(1)\ln(r) = v(e)\ln(r) = \log_a(r)$$

where $a = e^{\frac{1}{v(e)}}$. Note that a satisfies $0 < a$ and $a \neq 1$. From the fact that v is an even function, we finally have

$$v(r) = \log_a|r|$$

for $r \in \mathbb{R} \setminus \{0\}$.

A.2 Proof of Proposition 3

Without loss of generality, we can represent a scale-invariant function f as

$$f(cx) = u(c)f(x) + v(c) \tag{57}$$

since we can restore a multiplicatively or additively scale-invariant function by setting $v(c) = 0$ or $u(c) = 1$, respectively.

In order to derive the first-order derivative properties, we differentiate (57) with respect to x and c, respectively. Then, we further differentiate the latter with respect to x to obtain the second-order property.

By differentiating (57) with respect to x, we have

$$\nabla f(cx) = \frac{u(c)}{c}\nabla f(x).$$

On the other hand, by differentiating (57) with respect to c, we have

$$\nabla f(cx)^T x = u'(c)f(x) + v'(c). \tag{58}$$

By differentiating (58) with respect to x, we obtain

$$c\nabla^2 f(cx)x + \nabla f(cx) = u'(c)\nabla f(x). \tag{59}$$

Plugging $c = 1$ into (58) and (59) completes the proof.

A.3 Proof of Proposition 5

In order to prove the equivalence of (1) and (9), we show that a scalar multiple of an optimal solution to one problem is optimal to the other problem. Since $\{x : \|x\|_2 = 1\}$ and $\{w : f(w) = 1\}$ have a one-to-one correspondence, we can uniquely determine such a mapping.

First, we consider the multiplicatively scale invariant case where $f(x^*) > 0$. Suppose that an optimal solution to (9) is z not $x^*/f(x^*)^{1/p}$ such that

$$\|z\|_2 < \|x^*/f(x^*)^{1/p}\|_2. \tag{60}$$

Let $\hat{z} = z/\|z\|_2$. Then, we have $\|\hat{z}\|_2 = 1$ and $z = \hat{z}/f(\hat{z})^{1/p}$. Since $\|\hat{z}\|_2 = \|x^*\|_2 = 1$, we have

$$\|z\|_2 = \|\hat{z}/f(\hat{z})^{1/p}\|_2 = 1/f(\hat{z})^{1/p}, \quad \|x^*/f(x^*)^{1/p}\|_2 = 1/f(x^*)^{1/p}. \tag{61}$$

From (60) and (61), we have $f(x^*) < f(\hat{z})$ since $p > 0$. This contradicts the assumption that x^* is an optimal solution to (1).

For an optimal solution w^* to (9), since $f(w^*) = 1$, we have $w^* \neq 0$ and thus $\|w^*\|_2 > 0$ and $f\left(w^*/\|w^*\|_2\right) = 1/\|w^*\|_2^p > 0$. Suppose an optimal solution to (1) is y with

$$f(y) > f\left(w^*/\|w^*\|_2\right) > 0. \tag{62}$$

Let $\hat{y} = y/f(y)^{1/p}$. Then, $f(\hat{y}) = 1$ and $y = \hat{y}/\|\hat{y}\|_2$. Using $f(\hat{y}) = f(w^*) = 1$, we have

$$f(y) = f\left(\hat{y}/\|\hat{y}\|_2\right) = 1/\|\hat{y}\|_2^{1/p}, \qquad f\left(w^*/\|w^*\|_2\right) = 1/\|w^*\|_2^{1/p}. \tag{63}$$

From (62) and (63), we obtain $\|\hat{y}\|_2 < \|w^*\|_2$, which contradicts that w^* is optimal to (9).

Next, we consider the additively scale invariant case. Suppose that an optimal solution to (9) is z with

$$\|z\|_2 < \|a^{1-f(x^*)}x^*\|_2. \tag{64}$$

Let $\hat{z} = z/\|z\|_2$. Then, we have $\|\hat{z}\|_2 = 1$ and $z = a^{1-f(\hat{z})}\hat{z}$. Using $\|\hat{z}\|_2 = \|x^*\|_2 = 1$, we have

$$\|z\|_2 = a^{1-f(\hat{z})}, \quad \|a^{1-f(x^*)}x^*\|_2 = a^{1-f(x^*)}. \tag{65}$$

From (64) and (65), we obtain $f(x^*) < f(\hat{z})$ due to $a > 1$, which contradicts the assumption that x^* is an optimal solution to (1).

Conversely, suppose that an optimal solution of (1) is y with

$$f(y) > f\left(w^*/\|w^*\|_2\right). \tag{66}$$

Let $\hat{y} = a^{1-f(y)}y$. Then, we have $f(\hat{y}) = 1$ and $y = \hat{y}/\|\hat{y}\|_2$. Since $f(\hat{y}) = f(w^*) = 1$, we have

$$f(y) = f(\hat{y}) - \log_a\|\hat{y}\|_2 = 1 - \log_a\|\hat{y}\|_2, \quad f\left(w^*/\|w^*\|_2\right) = 1 - \log_a\|w^*\|_2. \tag{67}$$

From (66) and (67), we have $\|\hat{y}\|_2 < \|w^*\|_2$ since $a > 1$, contradicting the fact that w^* is an optimal solution to (9).

Appendix B. Sublinear convergence of SCI-PI for quasi-convex f

Proposition 11 *If f is quasi-convex and continuously differentiable, a sequence of iterates $\{x_k\}_{k=0,1,\cdots}$ generated by SCI-PI satisfies $f(x_{k+1}) \geq f(x_k)$ for all $k \geq 0$. Moreover, either Algorithm 1 terminates with a stationary point or every limit point is a stationary point.*

Proof Suppose that $f(x_{k+1}) < f(x_k)$. By the first-order property of differentiable quasi-convex functions, this leads to

$$\nabla f(x_k)^T(x_{k+1} - x_k) = \nabla f(x_k)^T\left(\frac{\nabla f(x_k)}{\|\nabla f(x_k)\|_2} - x_k\right) = \|\nabla f(x_k)\|_2 - \nabla f(x_k)^T x_k \leq 0. \tag{68}$$

However, since $f(x_{k+1}) \neq f(x_k)$, $\nabla f(x_k)$ is not a scalar multiple of x_k, resulting in

$$\|\nabla f(x_k)\|_2 - \nabla f(x_k)^T x_k > 0.$$

This contradicts (68). Therefore, we have $f(x_{k+1}) \geq f(x_k)$.

If Algorithm 1 terminates at iteration k, we have $\nabla f(x_k) = 0$. Therefore, x_k is a stationary point satisfying (8) with the value of the Lagrange multiplier being zero. Otherwise, let x^* be a limit point and suppose that x^* is not a stationary point. Then, there exists some $\epsilon > 0$ such that $\nabla f(x^*)^T x^*/\|\nabla f(x^*)\|_2 < 1 - \epsilon$. Since ∇f is continuous, there exists some $\delta > 0$ such that for $x \in \partial\mathcal{B}_d$ with $\nabla f(x) \neq 0$, if $\|x - x^*\|_2 < \delta$, then

$$\frac{\nabla f(x^*)^T \nabla f(x)}{\|\nabla f(x^*)\|_2 \|\nabla f(x)\|_2} > 1 - \epsilon.$$

Let k' be an index such that $\|x_{k'} - x^*\|_2 < \delta$. Since $\{f(x_k)\}_{k=0,1,\cdots}$ is non-decreasing, we have $f(x_{k'}) \leq f(x_{k'+1}) \leq f(x^*)$. By the first-order derivative property of quasi-convex functions, we obtain

$$\frac{\nabla f(x^*)^T(x_{k'+1} - x^*)}{\|\nabla f(x^*)\|_2} \leq 0.$$

However, since $x_{k'+1} = \nabla f(x_{k'})/\|\nabla f(x_{k'})\|_2$ and $\|x_{k'} - x^*\|_2 < \delta$, we must have

$$\frac{\nabla f(x^*)^T x^*}{\|\nabla f(x^*)\|_2} < 1 - \epsilon < \frac{\nabla f(x^*)^T \nabla f(x_{k'})}{\|\nabla f(x^*)\|_2 \|\nabla f(x_{k'})\|_2}.$$

This leads to a contradiction. Therefore, x^* must be a stationary point. ■

Appendix C. Additional Lemmas

On several occasions, we use if $x \in \partial\mathcal{B}_d$, $y \in \partial\mathcal{B}_d$, then

$$\|x - y\|_2^2 = \|x\|_2^2 + \|y\|_2^2 - 2x^T y = 2(1 - x^T y).$$

Note that if $x^T y \geq 0$, then

$$\sqrt{1 - (x^T y)^2} = \sqrt{(1 - x^T y)(1 + x^T y)} \geq \sqrt{1 - x^T y} = \frac{\|x - y\|_2}{\sqrt{2}}. \tag{69}$$

By Cauchy-Schwarz, we also have

$$\sqrt{1 - (x^T y)^2} = \sqrt{(1 - x^T y)(1 + x^T y)} \leq \sqrt{2}\sqrt{1 - x^T y} = \|x - y\|_2. \tag{70}$$

C.1 In Support of the Proofs of Theorem 6 and Theorem 7

Lemma 12 *Let x^* be a vector in $\mathbb{R}^d$ and $\{\mathbf{v}_1, \cdots, \mathbf{v}_d\}$ be an orthogonal basis in $\mathbb{R}^d$ such that $x^* = \mathbf{v}_1$. If a twice continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies*

$$\nabla f(x)^T \mathbf{v}_1 = \lambda^* + \alpha(x), \quad \sum_{i=2}^{d} (\nabla f(x)^T \mathbf{v}_i)^2 \leq \left(\bar{\lambda}_2 \|x - x^*\|_2 + \beta(x)\right)^2 \tag{71}$$

for every $x \in \partial\mathcal{B}_d$ and some functions $\alpha, \beta : \mathbb{R}^d \to \mathbb{R}$ and scalars $\lambda^*, \bar{\lambda}$ such that

$$\alpha(x) = o(\sqrt{\|x - x^*\|_2}), \quad \beta(x) = o(\|x - x^*\|_2), \quad \lambda^* > \bar{\lambda} \geq 0,$$

then for the sequence of iterates $\{x_k\}_{k=0,1,\cdots}$ generated by SCI-PI, there exists some $\delta > 0$ such that under the initial condition $\|x_0 - x^*\|_2 < \delta$, we have

$$\|x_k - x^*\|_2^2 \leq \prod_{t=0}^{k-1} \left(\frac{\bar{\lambda}_2}{\lambda^*} + \gamma_t\right)^2 \|x_0 - x^*\|_2^2, \ \frac{\bar{\lambda}_2}{\lambda^*} + \gamma_t < 1, \text{ and } \lim_{k\to\infty} \gamma_k = 0.$$

Proof By (71) for every $x \in \partial\mathcal{B}_d$, we have

$$\frac{\sum_{i=2}^{d} (\nabla f(x)^T \mathbf{v}_i)^2}{(\nabla f(x)^T \mathbf{v}_1)^2} \leq \left(\frac{\bar{\lambda}_2 \|x - x^*\|_2 + \beta(x)}{\lambda^* + \alpha(x)}\right)^2.$$

Let

$$\theta(x) = \frac{\bar{\lambda}_2 \|x - x^*\|_2 + \beta(x)}{\lambda^* + \alpha(x)} - \frac{\bar{\lambda}_2}{\lambda^*} \|x - x^*\|_2.$$

Then, we have $\theta(x) = o(\|x - x^*\|_2)$ and

$$\frac{\sum_{i=2}^{d} (\nabla f(x)^T \mathbf{v}_i)^2}{(\nabla f(x)^T \mathbf{v}_1)^2} \leq \left(\frac{\bar{\lambda}_2}{\lambda^*} + \frac{\theta(x)}{\|x - x^*\|_2}\right)^2 \|x - x^*\|_2^2. \tag{72}$$

Let

$$\epsilon(x) = \frac{\theta(x)}{\|x - x^*\|_2}. \tag{73}$$

For $x \in \mathbb{R}^d$ such that $x^T x^* > 0$ or $\|x - x^*\|_2 < \sqrt{2}$, we multiply (72) by $2/(1 + x^T x^*)$ to obtain

$$\begin{aligned} \frac{\sum_{i=2}^{d} (\nabla f(x)^T \mathbf{v}_i)^2}{(\nabla f(x)^T \mathbf{v}_1)^2} \left(\frac{2}{1 + x^T x^*}\right) &\leq \left(\frac{\bar{\lambda}_2}{\lambda^*} + \epsilon(x)\right)^2 \left(1 + \frac{1 - x^T x^*}{1 + x^T x^*}\right) \|x - x^*\|_2^2 \\ &= \left(\frac{\bar{\lambda}_2}{\lambda^*} + \gamma(x)\right)^2 \|x - x^*\|_2^2 \end{aligned} \tag{74}$$

where

$$\gamma(x) = \frac{\bar{\lambda}_2}{\lambda^*} \left(\frac{1 - x^T x^*}{1 + x^T x^* + \sqrt{2(1 + x^T x^*)}}\right) + \epsilon(x) \sqrt{1 + \frac{1 - x^T x^*}{1 + x^T x^*}}. \tag{75}$$

By (71), there exists some $\delta_1 > 0$ such that if $\|x - x^*\|_2 < \delta_1$, then

$$\nabla f(x)^T \mathbf{v}_1 > 0. \tag{76}$$

Also, by (73), for any $\bar{\gamma} > 0$ satisfying

$$\frac{\bar{\lambda}_2}{\lambda^*} + \bar{\gamma} < 1, \tag{77}$$

there exists some constant $\delta_2 > 0$ such that if $\|x - x^*\|_2 < \delta_2$, then

$$|\epsilon(x)| \leq \frac{\bar{\gamma}}{4}. \tag{78}$$

Let $\gamma_k = \gamma(x_k), \epsilon_k = \epsilon(x_k)$, and $\delta = \min\{\delta_1, \delta_2, \sqrt{\frac{2\lambda^*}{\bar{\lambda}_2}\bar{\gamma}}, \sqrt{2}\}$. We prove that if $\|x_k - x^*\|_2 < \delta$, then we have

$$\|x_{k+1} - x^*\|_2^2 \leq \left(\frac{\bar{\lambda}_2}{\lambda^*} + \gamma_k\right)^2 \|x_k - x^*\|_2^2 \text{ and } \gamma_k \leq \bar{\gamma}. \tag{79}$$

Since $\delta < \sqrt{2}$, we have $x_k^T x^* > 0$. Also, from $\|x_k - x^*\|_2 < \delta_1$ and $x^* = \mathbf{v}_1$, using the update rule of SCI-PI and (76), we obtain

$$x_{k+1}^T x^* = \frac{\nabla f(x_k)^T x^*}{\|\nabla f(x_k)\|_2} = \frac{\nabla f(x_k)^T \mathbf{v}_1}{\|\nabla f(x_k)\|_2} > 0.$$

On other the hand, since $|x_{k+1}^T \mathbf{v}_1| \leq \|x_{k+1}\|_2 \|\mathbf{v}_1\|_2 = 1$, we have

$$1 - (x_{k+1}^T x^*)^2 \leq \frac{1 - (x_{k+1}^T \mathbf{v}_1)^2}{(x_{k+1}^T \mathbf{v}_1)^2}.$$

Also, from the fact that $\{\mathbf{v}_1, \cdots, \mathbf{v}_d\}$ forms an orthogonal basis in $\mathbb{R}^d$, we have $\nabla f(x_k) = \sum_{i=1}^d (\nabla f(x_k)^T \mathbf{v}_i)\mathbf{v}_i$ and $\|\nabla f(x_k)\|_2^2 = \sum_{i=1}^d (\nabla f(x_k)^T \mathbf{v}_i)^2$. Using the update rule of SCI-PI, we have

$$\frac{1 - (x_{k+1}^T \mathbf{v}_1)^2}{(x_{k+1}^T \mathbf{v}_1)^2} = \frac{\|\nabla f(x_k)\|_2^2 - (\nabla f(x_k)^T \mathbf{v}_1)^2}{(\nabla f(x_k)^T \mathbf{v}_1)^2} = \frac{\sum_{i=2}^d (\nabla f(x_k)^T \mathbf{v}_i)^2}{(\nabla f(x_k)^T \mathbf{v}_1)^2}.$$

By $\|x_k - x^*\|_2 < \sqrt{2}$, using (74), we have

$$\|x_{k+1} - x^*\|_2^2 = (1 - (x_{k+1}^T x^*)^2)\left(\frac{2}{1 + x_{k+1}^T x^*}\right) \leq \left(\frac{\bar{\lambda}_2}{\lambda^*} + \gamma(x_k)\right)^2 \|x_k - x^*\|_2^2.$$

Since $x_k^T x^* > 0$, $\|x_k - x^*\|_2 < \delta_2$, and $1 - x_k^T x^* < \frac{\lambda^*}{\bar{\lambda}_2}\bar{\gamma}$,

$$\gamma_k = \frac{\bar{\lambda}_2}{\lambda^*}\left(\frac{1 - x_k^T x^*}{1 + x_k^T x^* + \sqrt{2(1 + x_k^T x^*)}}\right) + \epsilon_k \sqrt{1 + \frac{1 - x_k^T x^*}{1 + x_k^T x^*}} \leq \frac{\bar{\gamma}}{2} + \frac{\bar{\gamma}}{2} = \bar{\gamma},$$

which proves (79). Therefore, if x_0 satisfies $\|x_0 - x^*\|_2 < \delta$, by repeatedly applying (79), we obtain

$$\|x_k - x^*\|_2^2 \leq \prod_{t=0}^{k-1} \left(\frac{\bar{\lambda}_2}{\lambda^*} + \gamma_t \right)^2 \|x_0 - x^*\|_2^2 \text{ and } \quad \frac{\bar{\lambda}_2}{\lambda^*} + \gamma_k \leq \frac{\bar{\lambda}_2}{\lambda^*} + \bar{\gamma} < 1.$$

From

$$\|x_k - x^*\|_2^2 < \left(\frac{\bar{\lambda}_2}{\lambda^*} + \bar{\gamma} \right)^{2k} \|x_0 - x^*\|_2^2, \tag{80}$$

we have $x_k \to x^*$, and thus $\lim_{k\to\infty} \gamma_k = 0$ by (75). This gives the desired result. ■

Lemma 13 *Let x^* be a vector in $\mathbb{R}^d$ and $\{\mathbf{v}_1, \cdots, \mathbf{v}_d\}$ be an orthogonal basis in $\mathbb{R}^d$ such that $x^* = \mathbf{v}_1$. If a three-times continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies*

$$\nabla f(x)^T \mathbf{v}_1 \geq A - B(1 - x^T x^*) - C\sqrt{1 - x^T x^*} \tag{81}$$

and

$$\sum_{i=2}^{d} (\nabla f(x)^T \mathbf{v}_i)^2 \leq \left(D\sqrt{1 - (x^T x^*)^2} + E\|x - x^*\|_2 + \frac{F}{2}\|x - x^*\|_2^2 \right)^2 \tag{82}$$

for every $x \in \partial\mathcal{B}_d$ and some constants A, B, C, D, E, F such that

$$A > 0, \quad B + C > 0, \quad \frac{D + E}{A} < 1,$$

then for the sequence of iterates $\{x_k\}_{k=0,1,\cdots}$ generated by SCI-PI, under the initial condition that $\|x_0 - x^*\|_2 < \delta$ where

$$\delta = \min \left\{ \frac{\sqrt{2}A}{A + |A - B| + C}, \frac{\sqrt{2}(A - D - E)}{A - D + F + C + |A - B|} \right\}, \tag{83}$$

we have

$$\|x_k - x^*\|_2^2 \leq \prod_{t=0}^{k-1} \left(\frac{D + E}{A} + \gamma_t \right)^2 \|x_0 - x^*\|_2^2, \quad \frac{D + E}{A} + \gamma_t < 1, \text{ and } \lim_{k\to\infty} \gamma_k = 0.$$

Proof Let $\|x_0 - x^*\|_2 = \delta_0 < \delta$. To prove the main result, we show that if $\|x_k - x^*\|_2 < \delta_0$, then we have $x_{k+1}^T x^* > 0$ and

$$\frac{1 - (x_{k+1}^T x^*)^2}{(x_{k+1}^T x^*)^2} \leq \rho_k \frac{1 - (x_k^T x^*)^2}{(x_k^T x^*)^2} \tag{84}$$

where

$$\rho_k = \left(\frac{D + E + (E + F)\|x_k - x^*\|_2/\sqrt{2}}{A - (|A - B| + C)\|x_k - x^*\|_2/(\sqrt{2} - \|x_k - x^*\|_2)} \right)^2.$$

Suppose that $\|x_k - x^*\|_2 < \delta_0$ for $k \geq 0$. Since $\|x_k - x^*\|_2 < \delta$, by (83), we have $x_k^T x^* > 0$ and

$$\begin{aligned} A - B(1 - x_k^T x^*) - C\sqrt{1 - x_k^T x^*} &= Ax_k^T x^* + (A - B)(1 - x_k^T x^*) - C\sqrt{1 - x_k^T x^*} \\ &= Ax_k^T x^* + \frac{A - B}{2}\|x_k - x^*\|_2^2 - \frac{C}{\sqrt{2}}\|x_k - x^*\|_2 \\ &\geq x_k^T x^* \left(A - \frac{|A - B| + C}{\sqrt{2}} \frac{\|x_k - x^*\|_2}{x_k^T x^*} \right) \\ &> 0 \end{aligned} \tag{85}$$

where the first inequality follows from $\|x_k - x^*\|_2 \leq \sqrt{2}$ and the second one follows from

$$\frac{|A - B| + C}{\sqrt{2}} \frac{\|x_k - x^*\|_2}{x_k^T x^*} = \frac{(|A - B| + C)\|x_k - x^*\|_2}{\sqrt{2} - \|x_k - x^*\|_2^2/\sqrt{2}} \leq \frac{(|A - B| + C)\|x_k - x^*\|_2}{\sqrt{2} - \|x_k - x^*\|_2} < A.$$

Inequality (85) implies that

$$x_{k+1}^T x^* = \frac{\nabla f(x_k)^T \mathbf{v}_1}{\|\nabla f(x_k)\|_2} = \frac{A - B(1 - x_k^T x^*) - C\sqrt{1 - x_k^T x^*}}{\|\nabla f(x_k)\|_2} > 0.$$

On the other hand, since

$$\frac{1 - (x_{k+1}^T \mathbf{v}_1)^2}{(x_{k+1}^T \mathbf{v}_1)^2} = \frac{\|\nabla f(x_k)\|_2^2 - (\nabla f(x_k)^T \mathbf{v}_1)^2}{(\nabla f(x_k)^T \mathbf{v}_1)^2} = \frac{\sum_{i=2}^d (\nabla f(x_k)^T \mathbf{v}_i)^2}{(\nabla f(x_k)^T \mathbf{v}_1)^2},$$

using (81), (82), and (85), we have

$$\begin{aligned} \frac{1 - (x_{k+1}^T x^*)^2}{(x_{k+1}^T x^*)^2} &\leq \left(\frac{D\sqrt{1 - (x_k^T x^*)^2} + E\|x_k - x^*\|_2 + \frac{F}{2}\|x_k - x^*\|_2^2}{A - B(1 - x_k^T x^*) - C\sqrt{1 - x_k^T x^*}} \right)^2 \\ &\leq \left(\frac{D + E + (E + F)\|x_k - x^*\|_2/\sqrt{2}}{A - (|A - B| + C)\|x_k - x^*\|_2/(\sqrt{2} - \|x_k - x^*\|_2)} \right)^2 \frac{1 - (x_k^T x^*)^2}{(x_k^T x^*)^2} \end{aligned}$$

where we use the fact that $\sqrt{1 + x} \leq 1 + \sqrt{x}$ for $x \geq 0$ to derive

$$\begin{aligned} \frac{D\sqrt{1 - (x_k^T x^*)^2} + E\|x_k - x^*\|_2 + \frac{F}{2}\|x_k - x^*\|_2^2}{\sqrt{1 - (x_k^T x^*)^2}} &= D + E\sqrt{1 + \frac{1 - x_k^T x^*}{1 + x_k^T x^*}} + F\sqrt{\frac{1 - x_k^T x^*}{1 + x_k^T x^*}} \\ &\leq D + E + (E + F)\sqrt{\frac{1 - x_k^T x^*}{1 + x_k^T x^*}} \\ &\leq D + E + (E + F)\frac{\|x_k - x^*\|_2}{\sqrt{2}}. \end{aligned}$$

Since $x_k^T x^* > 0$ and $x_{k+1}^T x^* > 0$, we can write (84) as

$$\|x_{k+1} - x^*\|_2^2 \leq \left(\frac{\rho_k(1 + x_k^T x^*)}{\rho_k + (1-\rho_k)(x_k^T x^*)^2 + x_k^T x^* \sqrt{\rho_k + (1-\rho_k)(x_k^T x^*)^2}} \right) \|x_k - x^*\|_2^2.$$

Let

$$\bar{\rho}_k = \frac{\rho_k(1 + x_k^T x^*)}{\rho_k + (1-\rho_k)(x_k^T x^*)^2 + x_k^T x^* \sqrt{\rho_k + (1-\rho_k)(x_k^T x^*)^2}}.$$

Before proving $\bar{\rho}_k \leq \bar{\rho}_0 < 1$, we first show that $\rho_k \leq \rho_0 < 1$. Since $x_k^T x^* \geq x_0^T x^*$, we have $\|x_k - x^*\|_2 \leq \|x_0 - x^*\|_2$, and thus

$$\frac{\|x_k - x^*\|_2}{\sqrt{2} - \|x_k - x^*\|_2} \geq \frac{\|x_0 - x^*\|_2}{\sqrt{2} - \|x_0 - x^*\|_2},$$

which results in $\rho_k \leq \rho_0$. From $\delta_0 < \delta$ and (83), we have

$$\frac{|A - B| + C + E + F}{\sqrt{2} - \|x_0 - x^*\|_2} \|x_0 - x^*\|_2 \leq A - D - E,$$

and thus

$$A - \frac{|A - B| + C}{\sqrt{2} - \|x_0 - x^*\|_2} \|x_0 - x^*\|_2 - \left[D + E + (E + F) \frac{\|x_0 - x^*\|_2}{\sqrt{2}} \right] \geq 0.$$

This leads to $\rho_0 < 1$.

If $\rho_k = 0$, obviously $\rho_k \leq \rho_0$. Otherwise, using $0 < \rho_k \leq \rho_0$ and $x_k^T x^* \geq x_0^T x^*$, we have

$$\begin{aligned}
\bar{\rho}_k &= \frac{1 + x_k^T x^*}{(x_k^T x^*)^2/\rho_k + (1 - (x_k^T x^*)^2) + x_k^T x^* \sqrt{(x_k^T x^*)^2/\rho_k^2 + (1 - (x_k^T x^*)^2)/\rho_k}} \\
&\leq \frac{1 + x_k^T x^*}{(x_k^T x^*)^2/\rho_0 + (1 - (x_k^T x^*)^2) + x_k^T x^* \sqrt{(x_k^T x^*)^2/\rho_0^2 + (1 - (x_k^T x^*)^2)/\rho_0}} \\
&\leq \frac{\rho_0(1 + x_k^T x^*)}{\rho_0 + (1-\rho_0)(x_k^T x^*)^2 + x_k^T x^* \sqrt{\rho_0 + (1-\rho_0)(x_k^T x^*)^2}} \\
&\leq \frac{\rho_0}{\sqrt{\rho_0 + (1-\rho_0)(x_k^T x^*)^2}} \left(1 + \frac{1 - \sqrt{\rho_0 + (1-\rho_0)(x_k^T x^*)^2}}{\sqrt{\rho_0 + (1-\rho_0)(x_k^T x^*)^2} + x_k^T x^*} \right) \\
&\leq \frac{\rho_0}{\sqrt{\rho_0 + (1-\rho_0)(x_0^T x^*)^2}} \left(1 + \frac{1 - \sqrt{\rho_0 + (1-\rho_0)(x_0^T x^*)^2}}{\sqrt{\rho_0 + (1-\rho_0)(x_0^T x^*)^2} + x_0^T x^*} \right) \\
&= \bar{\rho}_0.
\end{aligned}$$

Since

$$\rho_0 + (1-\rho_0)(x_0^T x^*)^2 + x_0^T x^* \sqrt{\rho_k + (1-\rho_k)(x_0^T x^*)^2} - \rho_k(1 + x_0^T x^*) > 0,$$

we finally have $\bar{\rho} < 1$. Therefore, we have

$$\|x_{k+1} - x^*\|_2^2 \leq \left(\frac{D+E}{A} + \gamma_k\right)^2 \|x_k - x^*\|_2^2 \leq \bar{\rho}_0 \|x_k - x^*\|_2^2$$

where

$$\bar{\rho}_k = \left(\frac{D+E}{A} + \gamma_k\right)^2 < 1.$$

Since $\|x_{k+1} - x^*\|_2 < \delta$, by induction, we have

$$\|x_k - x^*\|_2^2 \leq \bar{\rho}_0^k \|x_0 - x^*\|_2^2,$$

which implies the convergence of x^* to x^*. As $x^* \to x^*$, we have $\bar{\rho}_k \to \rho_k$ and $\rho_k \to (\frac{D+E}{A})^2$, which leads to $\gamma_k \to 0$. This completes the proof. ■

Lemma 14 *Let f be three-times continously differentiable on an open set containing $\partial\mathcal{B}_d$ and H_j be the Hessian of $\nabla_j f = \partial f / \partial x_j$. Let $\{\mathbf{v}_1, \cdots, \mathbf{v}_d\}$ be an orthogonal basis in $\mathbb{R}^d$. For $x \in \partial\mathcal{B}_d$ and $y_1, \cdots, y_d \in \mathcal{B}_d$, let*

$$G_i(y^1, \cdots, y^d) = \sum_{j=1}^{d} \mathbf{v}_{i,j} H_j(y^j), \;\; R_i(x) = \nabla f(x)^T \mathbf{v}_i - \nabla f(x^*)^T \mathbf{v}_i - (x - x^*)^T \nabla^2 f(x^*) \mathbf{v}_i.$$

Then, we have

$$\sum_{i=1}^{d} (R_i(x))^2 \leq \frac{1}{4} M^2 \|x - x^*\|_2^4$$

where

$$M = \max_{x \in \partial\mathcal{B}_d, y^1, \cdots, y^d \in \mathcal{B}_d} \sqrt{\textstyle\sum_{i=1}^{d} (x^T G_i(y^1, \cdots, y^d) x)^2}.$$

Proof From $\nabla_j f(x)$ being twice continuously differentiable near $\partial\mathcal{B}_d$, we have

$$\nabla_j f(x) = \nabla_j f(x^*) + \nabla \nabla_j f(x^*)^T (x - x^*) + \frac{1}{2} (x - x^*)^T H_j(y^j) (x - x^*) \tag{86}$$

where $y^j \in \mathcal{N}(x, x^*) := \{y \,|\, y = \lambda x + (1-\lambda)x^*, \, 0 \leq \lambda \leq 1\}$. Since

$$\begin{aligned} R_i(x) &= \frac{1}{2}(x - x^*)^T G_i(y^1, \cdots, y^d)(x - x^*) \\ &= \frac{1}{2}\|x - x^*\|_2^2 \left[\frac{x - x^*}{\|x - x^*\|_2}\right]^T G_i(y^1, \cdots, y^d) \left[\frac{x - x^*}{\|x - x^*\|_2}\right], \end{aligned} \tag{87}$$

using the definition of M, we have the desired result. ■

C.2 In Support of the Proofs of Theorem 9 and Theorem 10

Lemma 15 *Suppose that $f(x,y)$ is scale invariant in $x \in \mathbb{R}^{d_x}$ for each $y \in \mathbb{R}^{d_y}$ and twice continuously differentiable on an open set containing $\partial\mathcal{B}_{d_x} \times \partial\mathcal{B}_{d_y}$. Let (x^*, y^*) be a point satisfying*

$$\nabla_x f(x^*, y^*) = \lambda^* x^*, \quad \lambda^* > \bar{\lambda}_2 = \max_{2 \le i \le d_x} |\lambda_i|, \quad x^* = \mathbf{v}_1$$

where $(\lambda_i, \mathbf{v}_i)$ is an eigen-pair of $\nabla^2_{xx} f(x^*, y^*)$. Then, for any $x \in \partial\mathcal{B}_{d_x}$ and $y \in \partial\mathcal{B}_{d_y}$, we have

$$\nabla_x f(x,y)^T \mathbf{v}_1 = \lambda^* + (y - y^*)^T \nabla^2_{yx} f(x^*, y^*) x^* + \alpha(x,y)$$

and

$$\sum_{i=2}^{d_x} (\nabla_x f(x,y)^T \mathbf{v}_i)^2 \le \left(\bar{\lambda}_2 \|x - x^*\|_2 + \nu \|y - y^*\|_2 + \beta(x,y)\right)^2$$

where

$$\alpha(x,y) = o\left(\left\| \begin{bmatrix} x - x^* \\ y - y^* \end{bmatrix} \right\|_2\right), \ \beta(x,y) = o\left(\left\| \begin{bmatrix} x - x^* \\ y - y^* \end{bmatrix} \right\|_2\right), \ \nu = \|\nabla^2_{xy} f(x^*, y^*)\|_2.$$

Therefore, we have

$$\frac{\sum_{i=2}^{d_x} (\nabla_x f(x,y)^T \mathbf{v}_i)^2}{(\nabla_x f(x,y)^T \mathbf{v}_1)^2} \le \left(\frac{\bar{\lambda}_2}{\lambda^*} \|x - x^*\|_2 + \frac{\nu}{\lambda^*} \|y - y^*\|_2 + \theta(x,y)\right)^2$$

where

$$\theta(x,y) = o\left(\left\| \begin{bmatrix} x - x^* \\ y - y^* \end{bmatrix} \right\|_2\right).$$

Proof Since $\nabla^2_{xx} f(x^*, y^*)$ is real and symmetric, without loss of generality, we assume that $\{\mathbf{v}_1, \cdots, \mathbf{v}_{d_x}\}$ forms an orthogonal basis in $\mathbb{R}^{d_x}$.

By Taylor expansion of $\nabla_x f(x,y)^T \mathbf{v}_i$ at (x^*, y^*), we have

$$\nabla_x f(x,y)^T \mathbf{v}_i = \nabla_x f(x^*, y^*)^T \mathbf{v}_i + \begin{bmatrix} x - x^* \\ y - y^* \end{bmatrix}^T \begin{bmatrix} \nabla^2_{xx} f(x^*, y^*) \\ \nabla^2_{yx} f(x^*, y^*) \end{bmatrix} \mathbf{v}_i + R_i(x,y)$$

where

$$R_i(x,y) = o\left(\left\| \begin{bmatrix} x - x^* \\ y - y^* \end{bmatrix} \right\|_2\right).$$

Using $\nabla_x f(x^*, y^*) = \lambda^* x^*$ and $x^* = \mathbf{v}_1$, we have

$$\nabla_x f(x^*, y^*)^T \mathbf{v}_1 = \lambda^*, \quad (x - x^*)^T \nabla^2_{xx} f(x^*, y^*) \mathbf{v}_1 = -\lambda_1 (1 - x_k^T x^*).$$

Therefore, we obtain

$$\nabla_x f(x,y)^T \mathbf{v}_1 = \lambda^* + (x - x^*)^T \nabla^2_{yx} f(x^*, y^*) x^* + \alpha(x,y) \tag{88}$$

where

$$\alpha(x,y) = R_1(x,y) - \lambda_1(1 - x^T x^*) = o\left(\left\|\begin{bmatrix} x - x^* \\ y - y^* \end{bmatrix}\right\|_2\right).$$

In the same way, for $2 \le i \le d_x$, we have

$$\nabla_x f(x^*, y^*)^T \mathbf{v}_i = \lambda^* (x^*)^T \mathbf{v}_i = 0, \quad (x - x^*)^T \nabla^2_{xx} f(x^*, y^*) \mathbf{v}_i = \lambda_i x^T \mathbf{v}_i,$$

resulting in

$$\nabla_x f(x,y)^T \mathbf{v}_i = \lambda_i x^T \mathbf{v}_i + (y - y^*)^T \nabla^2_{yx} f(x^*, y^*) \mathbf{v}_i + R_i(x,y). \tag{89}$$

From (89), we obtain

$$\begin{aligned}
\sum_{i=2}^{d_x} (\nabla_x f(x,y)^T \mathbf{v}_i)^2 = &\sum_{i=2}^{d_x} (\lambda_i)^2 (x^T \mathbf{v}_i)^2 + \sum_{i=2}^{d_x} \left((y - y^*)^T \nabla^2_{yx} f(x^*, y^*) \mathbf{v}_i\right)^2 \\
&+ \sum_{i=2}^{d_x} (R_i(x,y))^2 + 2 \sum_{i=2}^{d_x} \lambda_i (x^T \mathbf{v}_i)(y - y^*)^T \nabla^2_{yx} f(x^*, y^*) \mathbf{v}_i \\
&+ 2 \sum_{i=2}^{d_x} \lambda_i (x^T \mathbf{v}_i) R_i(x,y) \\
&+ 2 \sum_{i=2}^{d_x} (y - y^*)^T \nabla^2_{yx} f(x^*, y^*) \mathbf{v}_i R_i(x,y).
\end{aligned}$$

Since $\{\mathbf{v}_1, \cdots, \mathbf{v}_{d_x}\}$ forms an orthogonal basis in $\mathbb{R}^{d_x}$, with $x^* = \mathbf{v}_1$ and $\|x\|_2^2 = 1$, we have

$$\sum_{i=2}^{d_x} (\lambda_i)^2 (x^T \mathbf{v}_i)^2 \le (\bar{\lambda}_2)^2 \left(1 - (x^T x^*)^2\right) \le (\bar{\lambda}_2)^2 \|x - x^*\|_2^2$$

and

$$\sum_{i=2}^{d_x} \left((y - y^*)^T \nabla^2_{yx} f(x^*, y^*) \mathbf{v}_i\right)^2 \le \|(y - y^*)^T \nabla^2_{yx} f(x^*, y^*)\|_2^2 \le \nu^2 \|y - y^*\|_2^2.$$

Let $\bar{R}_2(x,y) = \max_{2 \le i \le d_x} |R_i(x,y)|$. Note that

$$\bar{R}_2(x,y) = o\left(\left\|\begin{bmatrix} x - x^* \\ y - y^* \end{bmatrix}\right\|_2\right).$$

Using the Cauchy-Shwartz inequality, we have

$$\sum_{i=2}^{d_x} \lambda_i (x^T \mathbf{v}_i)(y - y^*)^T \nabla^2_{yx} f(x^*, y^*) \mathbf{v}_i \le \bar{\lambda}_2 \nu \|y - y^*\|_2 \|x - x^*\|_2.$$

Also, we have

$$\sum_{i=2}^{d_x} \lambda_i (x^T \mathbf{v}_i) R_i(x,y) \le \bar{\lambda}_2 \bar{R}_2(x,y) \sqrt{d_x} \|x - x^*\|_2$$

and

$$\sum_{i=2}^{d_x} R_i(x,y)(y-y^*)^T\nabla^2_{yx}f(x^*,y^*)\mathbf{v}_i \leq \nu\bar{R}_2(x,y)\sqrt{d_x}\|y-y^*\|_2.$$

Therefore, we obtain

$$\sum_{i=2}^{d_x}(\nabla_x f(x,y)^T\mathbf{v}_i)^2 \leq \left(\bar{\lambda}_2\|x-x^*\|_2+\nu\|y-y^*\|_2+\beta(x,y)\right)^2 \tag{90}$$

where

$$\beta(x,y)=\bar{R}_2(x,y)\sqrt{d_x}=o\left(\left\|\begin{bmatrix}x-x^*\\ y-y^*\end{bmatrix}\right\|_2\right).$$

Since $\{\mathbf{v}_1,\cdots,\mathbf{v}_{d_x}\}$ forms an orthogonal basis in $\mathbb{R}^{d_x}$ and $|x^Tx^*|\leq\|x\|_2\|x^*\|_2=1$, we have

$$1-\frac{(\nabla_x f(x,y)^Tx^*)^2}{\|\nabla_x f(x,y)\|_2^2}\leq\frac{\sum_{i=2}^{d_x}(\nabla_x f(x,y)^T\mathbf{v}_i)^2}{(\nabla_x f(x,y)^T\mathbf{v}_1)^2}.$$

Using (88) and (90), we have

$$\frac{\sum_{i=2}^{d_x}(\nabla_x f(x,y)^T\mathbf{v}_i)^2}{(\nabla_x f(x,y)^T\mathbf{v}_1)^2}\leq\left(\frac{\bar{\lambda}_2}{\lambda^*}\|x-x^*\|_2+\frac{\nu}{\lambda^*}\|y-y^*\|_2+\theta(x,y)\right)^2$$

where

$$\theta(x,y)=o\left(\left\|\begin{bmatrix}x-x^*\\ y-y^*\end{bmatrix}\right\|_2\right).$$

This completes the proof. ■

Lemma 16 *Suppose that $f(x,y)$ is μ-strongly concave in $y\in\mathbb{R}^{d_y}$ with an L-Lipschitz continuous $\nabla_y f(x,y)$ for each $x\in\partial\mathcal{B}_{d_x}$ and three-times continously differentiable with respect to x and y on an open set containing $\partial\mathcal{B}_{d_x}\times\mathbb{R}^{d_y}$. Let (x^*,y^*) be a point such that $\nabla_y f(x^*,y^*)=0$. Then, for any $x\in\partial\mathcal{B}_{d_x}$ and $y\in\mathbb{R}^{d_y}$, with $\alpha=2/(L+\mu)$, we have*

$$\|y+\alpha\nabla_y f(x,y)-y^*\|_2\leq\left(\frac{2\nu}{L+\mu}\right)\|x-x^*\|_2+\left(\frac{L-\mu}{L+\mu}\right)\|y-y^*\|_2+\theta(x,y) \tag{91}$$

where

$$\nu=\|\nabla^2_{yx}f(x^*,y^*)\|_2,\quad \theta(x,y)=o\left(\left\|\begin{bmatrix}x-x^*\\ y-y^*\end{bmatrix}\right\|_2\right).$$

Proof Let $\nabla_{y,i}f$ be the i^{th} coordinate of $\nabla_y f$ and

$$H_{y,i}=\begin{bmatrix}H^{xx}_{y,i} & H^{xy}_{y,i}\\ H^{yx}_{y,i} & H^{yy}_{y,i}\end{bmatrix}$$

be the Hessian of $\nabla_{y,i} f$. By Taylor expansion of $\nabla_{y,i} f(x,y)$ at (x^*, y), we have

$$\nabla_{y,i} f(x,y) = \nabla_{y,i} f(x^*,y) + \nabla_x \nabla_{y,i} f(x^*,y)^T (x - x^*) + R_i(x,y) \tag{92}$$

where $\nabla_x \nabla_{y,i} f(x^*,y)$ denotes the i^{th} column of $\nabla^2_{yx} f(x^*,y)$ and

$$R_i(x,y) = \frac{1}{2}(x-x^*)^T H^{xx}_{y,i}(\hat{x}^i, y)(x-x^*), \quad \hat{x}^i \in \mathcal{N}(x,x^*). \tag{93}$$

Also, from f being three-times continuously differentiable, we have

$$\nabla_x \nabla_{y,i} f(x^*,y) = \nabla_x \nabla_{y,i} f(x^*,y^*) + H^{xy}_{y,i}(x^*, \hat{y}^i)(y-y^*), \quad \hat{y}^i \in \mathcal{N}(y,y^*). \tag{94}$$

Since

$$\begin{aligned} |(y-y^*)^T H^{yx}_{y,i}(x^*,\hat{y}^i)(x-x^*)| &\leq \|H^{yx}_{y,i}(x^*,\hat{y}^i)\|_2 \|x-x^*\|_2 \|y-y^*\|_2 \\ &\leq \frac{1}{2}\|H^{yx}_{y,i}(x^*,\hat{y}^i)\|_2 \left(\|x-x^*\|_2^2 + \|y-y^*\|_2^2\right), \end{aligned}$$

we have

$$(y-y^*)^T H^{yx}_{y,i}(x^*,\hat{y}^i)(x-x^*) = o\left(\left\| \begin{bmatrix} x-x^* \\ y-y^* \end{bmatrix} \right\|_2\right). \tag{95}$$

By (92), (93), (94), and (95), we have

$$\nabla_y f(x,y) = \nabla_y f(x^*,y) + \nabla^2_{yx} f(x^*,y^*)(x-x^*) + \bar{R}(x,y) \tag{96}$$

where

$$\bar{R}_i(x,y) = R_i(x,y) + (y-y^*)^T H^{yx}_{y,i}(x^*,\hat{y}^i)(x-x^*) = o\left(\left\| \begin{bmatrix} x-x^* \\ y-y^* \end{bmatrix} \right\|_2\right).$$

Using (96), we have

$$y + \alpha \nabla_y f(x,y) - y^* = y - y^* + \alpha \nabla_y f(x^*,y) + \alpha \nabla^2_{yx} f(x^*,y^*)(x-x^*) + \bar{R}(x,y),$$

resulting in

$$\begin{aligned} \|y + \alpha \nabla_y f(x,y) - y^*\|_2 \leq \|y - y^* + \alpha \nabla_y f(x^*,y)\|_2 \\ + \alpha \|\nabla^2_{yx} f(x^*,y^*)(x-x^*)\|_2 + \alpha \|\bar{R}(x,y)\|_2. \end{aligned} \tag{97}$$

Since $-f(x^*,y)$ is μ-strongly convex in y with an L-Lipschitz continuous gradient $-\nabla_y f(x^*,y)$, by theory of convex optimization (Bubeck, 2015, p. 279), we have

$$\|y - y^* + \alpha \nabla_y f(x^*,z)\|_2 \leq \left(\frac{L-\mu}{L+\mu}\right) \|y-y^*\|_2 \tag{98}$$

due to $\alpha = 2/(L+\mu)$. Also, we have

$$\alpha \|\nabla^2_{yx} f(x^*,y^*)(x-x^*)\|_2 \leq \left(\frac{2\nu}{L+\mu}\right) \|x-x^*\|_2. \tag{99}$$

Plugging (98), (99) into (97), we finally obtain

$$\|y - y^* + \alpha \nabla_y f(x^*, y)\|_2 \leq \left(\frac{L-\mu}{L+\mu}\right) \|y - y^*\|_2 + \left(\frac{2\nu}{L+\mu}\right) \|x - x^*\|_2 + \theta(x, y)$$

where

$$\theta(x, y) = \|\bar{R}(x, y)\|_2 = o\left(\left\| \begin{bmatrix} x - x^* \\ y - y^* \end{bmatrix} \right\|_2\right).$$

■

Lemma 17 *Suppose that a sequence of iterates* $\{(x_k, y_k)\}_{k=0,1,\cdots}$ *satisfies*

$$\begin{bmatrix} \|x_{k+1} - x^*\|_2 \\ \|y_{k+1} - y^*\|_2 \end{bmatrix} \leq \begin{bmatrix} a & e/b \\ e/c & d \end{bmatrix} \begin{bmatrix} \|x_k - x^*\|_2 \\ \|y_k - y^*\|_2 \end{bmatrix} + \begin{bmatrix} \theta^x(x_k, y_k) \\ \theta^y(x_k, y_k) \end{bmatrix} \tag{100}$$

for some functions θ^x *and* θ^y *such that*

$$\theta^x(x_k, y_k) = o\left(\left\| \begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix} \right\|_2\right), \quad \theta^y(x_k, y_k) = o\left(\left\| \begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix} \right\|_2\right). \tag{101}$$

If $a, d, e \geq 0$ *and* $b, c > 0$ *satisfy*

$$\rho = \frac{1}{2}\left(a + b + \sqrt{(a-b)^2 + \frac{4e^2}{bc}}\right) < 1,$$

then there exists some $\delta > 0$ *such that if* $\Delta_0 < \delta$, *then we have*

$$\Delta_k^2 \leq \prod_{t=1}^{k-1} (\rho + \gamma_t)^2 \Delta_0^2 \text{ and } \lim_{t \to \infty} \gamma_t = 0$$

where

$$\Delta_k = \left\| \begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix} \right\|_2.$$

Proof From (100), we have

$$\begin{bmatrix} \|x_{k+1} - x^*\|_2 \\ \|y_{k+1} - y^*\|_2 \end{bmatrix} \leq \begin{bmatrix} a & e/b \\ e/c & d \end{bmatrix} \begin{bmatrix} \|x_k - x^*\|_2 \\ \|y_k - y^*\|_2 \end{bmatrix} + \begin{bmatrix} \theta^x(x_k, y_k) \\ \theta^y(x_k, y_k) \end{bmatrix} \tag{102}$$

$$\leq (M + N(x_k, y_k)) \begin{bmatrix} \|x_k - x^*\|_2 \\ \|y_k - y^*\|_2 \end{bmatrix} \tag{103}$$

where

$$M = \begin{bmatrix} a & e/b \\ e/c & d \end{bmatrix}, \quad \epsilon(x, y) = \frac{\max\{\theta^x(x, y), \theta^y(x, y)\}}{\sqrt{\|x_k - x^*\|_2^2 + \|y_k - y^*\|_2^2}},$$

and

$$N(x,y) = \frac{\epsilon(x,y)}{\sqrt{\|x_k - x^*\|_2^2 + \|y_k - y^*\|_2^2}} \begin{bmatrix} \|x_k - x^*\|_2 & \|y_k - y^*\|_2 \\ \|x_k - x^*\|_2 & \|y_k - y^*\|_2 \end{bmatrix}.$$

Note that we have

$$\lim_{(x,y)\to(x^*,y^*)} N_{ij}(x,y) = 0, \quad i,j = 1,2.$$

By Lemma 18, there exists a sequence $\{\omega_t\}_{t=0,1,\dots}$ such that

$$\|M^k\|_2 = \prod_{t=0}^{k-1}(\rho + \omega_t) \quad \text{and} \quad \lim_{t\to\infty}\omega_t = 0.$$

Since $\rho < 1$, this implies that $\|M^k\|_2$ converge to 0. Let

$$\tau = \min\{k : \|M^k\|_2 < 1\}, \quad \bar{\rho} = \frac{\|M^\tau\|_2 + 1}{2}, \quad \rho_{\max} = \max_{1\le k\le\tau} \|M^k\|_2.$$

Due to $N_{ij}(x,y) \to 0$ as $(x,y) \to (x^*,y^*)$ for $i,j = 1,2$, there exists some $\delta > 0$ such that if

$$\left\| \begin{bmatrix} x - x^* \\ y - y^* \end{bmatrix} \right\|_2 < \delta,$$

then we have

$$\left\| \prod_{l=0}^{\tau-1} \big(M + N(\phi(x,y,l))\big) \right\|_2 < \bar{\rho}, \quad \max_{0<m\le\tau} \left\| \prod_{l=0}^{m-1} \big(M + N(\phi(x,y,l))\big) \right\|_2 < 1 + \rho_{\max} \tag{104}$$

where $\phi(x,y,l)$ denotes the l^{th} iterate (x_l, y_l) of the underlying algorithm starting with $(x_0, y_0) = (x, y)$.

To see this, let us define

$$g(x,y,m) = \left\| \prod_{l=0}^{m-1} \big(M + N(\phi(x,y,l))\big) \right\|_2.$$

By (100) and (101), if $x \to x^*$ and $y \to y^*$, then for any $0 \le l \le \tau$, we have

$$\phi(x,y,l) \to (x^*, y^*),$$

resulting in

$$g(x,y,m) \to \|M^m\|_2.$$

Therefore, there exists some $\delta_\tau > 0$ such that $g(x,y,\tau) < \bar{\rho}$. Also, for each $1 \le m < \tau$, there exists some $\delta_m > 0$ such that $g(x,y,m) < 1 + \rho_{\max}$. Taking the minimum of δ_m for $1 \le m \le \tau$, we obtain δ satisfying (104).

For any $n \ge 0$, if $\Delta_{n\tau} < \delta$, using (104) for (103), we have

$$\Delta_{n\tau+m} \le (1 + \rho_{\max})\Delta_{n\tau}, \quad 1 \le m < \tau, \quad \Delta_{(n+1)\tau} \le \bar{\rho}\Delta_{n\tau}. \tag{105}$$

Suppose that $\Delta_0 \leq \delta$. Then, by repeatedly applying (105), for any $n \geq 0$ and $0 \leq m \leq \tau$, we have

$$\Delta_{n\tau+m} \leq (\bar{\rho})^n (1+\rho_{\max})\Delta_0,$$

which implies that $\Delta_k \to 0$ as $k \to \infty$. Let

$$N_t = N(x_t, y_t), \quad \eta_k = \frac{\|\prod_{t=0}^{k}(M+N_t)\|_2}{\|\prod_{t=0}^{k-1}(M+N_t)\|_2} - \frac{\|M^{k+1}\|_2}{\|M^k\|_2}, \quad \gamma_k = \omega_k + \eta_k.$$

Then, we have

$$\left\|\prod_{t=0}^{k-1}(M+N_t)\right\|_2 = \prod_{t=0}^{k-1}(\rho+\omega_t+\eta_t) = \prod_{t=0}^{k-1}(\rho+\gamma_t). \tag{106}$$

Since $N_t \to 0$, we have $\eta_t \to 0$, and thus $\lim_{t\to\infty}\gamma_t = 0$. This concludes the proof. ■

Lemma 18 *Let M be a 2×2 matrix such that*

$$M = \begin{bmatrix} a & e/b \\ e/c & d \end{bmatrix}$$

for some $a>0, b>0, c>0, d\geq 0, e\geq 0$ and let ρ be the largest absolute eigenvalue of M. Then, there exists a sequence $\{\omega_t\}_{t=0,1,\ldots}$ such that

$$\|M^k\|_2 = \prod_{t=0}^{k-1}(\rho+\omega_t) \quad \textit{and} \quad \lim_{t\to\infty}\omega_t = 0.$$

Proof The characteristic equation reads

$$\det(M-\lambda I) = \lambda^2 - \lambda(a+d) + ad - \frac{e^2}{bc} = 0$$

with the discriminant of

$$(a-d)^2 + \frac{4e^2}{bc} \geq 0.$$

Thus, all eigenvalues are real.

First, we consider the case when $\det(M-\lambda I)=0$ has a double root. We obtain the condition for a double root as

$$(a-d)^2 + \frac{4e^2}{bc} = 0.$$

Since $b>0$ and $c>0$, this implies $a=d$ and $e=0$. Therefore, $M=aI$ and $\rho=a$. From $M^k = a^k I$, we have $\|M^k\|_2 = \sqrt{a^{2k}} = \rho^k$, resulting in

$$\omega_k = \frac{\|M^{k+1}\|_2}{\|M^k\|_2} - \rho = \rho - \rho = 0, \quad k \geq 0.$$

Next, we consider the case when M has two distinct eigenvalues λ_1 and λ_2. Since $a + d > 0$, we have $\lambda_1 + \lambda_2 > 0$. Without loss of generality, assume $\lambda_1 > \lambda_2$. Then, $\rho = \lambda_1$. Let v_1 and v_2 be corresponding eigenvectors of λ_1 and λ_2, respectively. Since v_1 and v_2 are linearly independent we can represent each column of M as a linear combination of v_1 and v_2 as

$$M = [\alpha_1 v_1 + \beta_1 v_2 \quad \alpha_2 v_1 + \beta_2 v_2].$$

By repeatedly multiplying M, we obtain

$$M^k = [\alpha_1 \lambda_1^{k-1} v_1 + \beta_1 \lambda_2^{k-1} v_2 \quad \alpha_2 \lambda_1^{k-1} v_1 + \beta_2 \lambda_2^{k-1} v_2].$$

Let $C^k = (M^k)^T M^k$. Then, we have

$$C_{11}^k = \alpha_1^2 \lambda_1^{2(k-1)} + \beta_1^2 \lambda_2^{2(k-1)} + 2\alpha_1\beta_1(\lambda_1\lambda_2)^{k-1} v_1^T v_2$$
$$C_{22}^k = \alpha_2^2 \lambda_1^{2(k-1)} + \beta_2^2 \lambda_2^{2(k-1)} + 2\alpha_2\beta_2(\lambda_1\lambda_2)^{k-1} v_1^T v_2$$

and

$$C_{12}^k = \alpha_1\alpha_2 \lambda_1^{2(k-1)} + \beta_1\beta_2 \lambda_2^{2(k-1)} + (\alpha_1\beta_2 + \alpha_2\beta_1)(\lambda_1\lambda_2)^{k-1} v_1^T v_2, \quad C_{21}^k = C_{12}^k.$$

Since

$$C_{11}^k \geq \alpha_1^2 \lambda_1^{2(k-1)} + \beta_1^2 \lambda_2^{2(k-1)} - 2\alpha_1\beta_1(\lambda_1\lambda_2)^{k-1} = \left(\alpha_1 \lambda_1^{k-1} - \beta_1 \lambda_2^{k-1}\right)^2 \geq 0$$
$$C_{22}^k \geq \alpha_2^2 \lambda_1^{2(k-1)} + \beta_2^2 \lambda_2^{2(k-1)} - 2\alpha_2\beta_2(\lambda_1\lambda_2)^{k-1} = \left(\alpha_2 \lambda_1^{k-1} - \beta_2 \lambda_2^{k-1}\right)^2 \geq 0,$$

we have

$$\|M^k\|_2 = \sqrt{\frac{1}{2}\left[C_{11}^k + C_{22}^k + \sqrt{\left(C_{11}^k - C_{22}^k\right)^2 + 4(C_{12}^k)^2}\right]},$$

leading to

$$\frac{\|M^{k+1}\|_2}{\|M^k\|_2} = \sqrt{\frac{C_{11}^{k+1} + C_{22}^{k+1} + \sqrt{\left(C_{11}^{k+1} - C_{22}^{k+1}\right)^2 + 4(C_{12}^{k+1})^2}}{C_{11}^k + C_{22}^k + \sqrt{\left(C_{11}^k - C_{22}^k\right)^2 + 4(C_{12}^k)^2}}}.$$

From

$$\lim_{k\to\infty} \frac{C_{11}^k}{\lambda_1^{2(k-1)}} = \alpha_1^2, \quad \lim_{k\to\infty} \frac{C_{22}^k}{\lambda_1^{2(k-1)}} = \alpha_2^2, \quad \lim_{k\to\infty} \frac{C_{12}^k}{\lambda_1^{2(k-1)}} = \lim_{k\to\infty} \frac{C_{21}^k}{\lambda_1^{2(k-1)}} = \alpha_1\alpha_2,$$

we obtain

$$\lim_{k\to\infty} \frac{\|M^{k+1}\|_2}{\|M^k\|_2} = \sqrt{\lambda_1^2} = \rho.$$

From

$$\lim_{k\to\infty} \omega_k = \lim_{k\to\infty} \frac{\|M^{k+1}\|_2}{\|M^k\|_2} - \rho = \rho - \rho = 0,$$

we obtain the desired result. ■

References

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

Michal Aharon, Michael Elad, and Alfred Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006.

Amir Beck. On the convergence of alternating minimization for convex programming with applications to iteratively reweighted least squares and decomposition schemes. *SIAM Journal on Optimization*, 25(1):185–209, 2015.

Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Nicolas Boumal. Nonconvex phase synchronization. *SIAM Journal on Optimization*, 26(4): 2355–2377, 2016.

Christos Boutsidis, Dan Garber, Zohar Karnin, and Edo Liberty. Online principal components analysis. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 887–901, 2015.

Stephen Boyd, Neal Parikh, and Eric Chu. *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*. Now Publishers Inc, 2011.

Sébastien Bubeck. *Convex Optimization: Algorithms and Complexity*. Now Publishers, Inc., 2015.

Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2): 329–357, 2003.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717, 2009.

Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176:5–37, 2019.

Murat A Erdogdu, Asuman Ozdaglar, Pablo A Parrilo, and Nuri Denizcan Vanli. Convergence rate of block-coordinate maximization Burer-Monteiro method for solving large SDPs. *Mathematical Programming*, 195(1):243–281, 2022.

Cédric Févotte and Jérôme Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural Computation*, 23(9):2421–2456, 2011.

John GF Francis. The QR transformation a unitary analogue to the LR transformation - part 1. *The Computer Journal*, 4(3):265–271, 1961.

John GF Francis. The QR transformation - part 2. *The Computer Journal*, 4(4):332–345, 1962.

Robert M Freund. Dual gauge programs, with applications to quadratic programming and the minimum-norm problem. *Mathematical Programming*, 38(1):47–67, 1987.

Michael P Friedlander, Ives Macedo, and Ting Kei Pong. Gauge optimization and duality. *SIAM Journal on Optimization*, 24(4):1999–2022, 2014.

Dan Garber, Elad Hazan, and Tengyu Ma. Online learning of eigenvectors. In *International Conference on Machine Learning*, pages 560–568, 2015.

Gene H Golub and Henk A Van der Vorst. Eigenvalue computation in the 20th century. *Journal of Computational and Applied Mathematics*, 123(1-2):35–65, 2000.

Gene H Golub and Charles F Van Loan. *Matrix Computations*. JHU Press, 2012.

Fang Han and Han Liu. Scale-invariant sparse PCA on high-dimensional meta-elliptical data. *Journal of the American Statistical Association*, 109(505):275–287, 2014.

Mingyi Hong and Zhi-Quan Luo. On the linear convergence of the alternating direction method of multipliers. *Mathematical Programming*, 162(1-2):165–199, 2017.

Meng Huang and Zhiqiang Xu. Strong convexity of affine phase retrieval. *arXiv preprint arXiv:2204.09412*, 2022.

Ruitong Huang, Tor Lattimore, András György, and Csaba Szepesvári. Following the leader and fast rates in online linear prediction: Curved constraint sets and other regularities. *Journal of Machine Learning Research*, 18(145):1–31, 2017.

Aapo Hyvarinen. Fast ICA for noisy data using Gaussian moments. In *IEEE International Symposium on Circuits and Systems VLSI*, volume 5, pages 57–61, 1999.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000.

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. John Wiley & Sons, 2004.

Michel Journée, Yurii Nesterov, Peter Richtárik, and Rodolphe Sepulchre. Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11(Feb):517–553, 2010.

Cheolmin Kim and Diego Klabjan. A simple and fast algorithm for L1-norm kernel PCA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8):1842–1855, 2020a.

Cheolmin Kim and Diego Klabjan. Stochastic variance-reduced algorithms for PCA with arbitrary mini-batch sizes. In *International Conference on Artificial Intelligence and Statistics*, volume 108, pages 4302–4312, 2020b.

Youngseok Kim, Peter Carbonetto, Matthew Stephens, and Mihai Anitescu. A fast algorithm for maximum likelihood estimation of mixture proportions using sequential quadratic programming. *Journal of Computational and Graphical Statistics*, 29(2):261–273, 2020.

Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, pages 556–562, 2001.

Qi Lei, Kai Zhong, and Inderjit S Dhillon. Coordinate-wise power method. In *Advances in Neural Information Processing Systems*, pages 2064–2072, 2016.

Adrian S Lewis, D Russell Luke, and Jérôme Malick. Local linear convergence for alternating and averaged nonconvex projections. *Foundations of Computational Mathematics*, 9(4): 485–513, 2009.

Chih-Jen Lin. Projected gradient methods for non-negative matrix factorization. *Neural Computation*, 19(10):2756–2779, 2007.

Huikang Liu, Man-Chung Yue, and Anthony Man-Cho So. On the estimation performance and convergence rate of the generalized power method for phase synchronization. *SIAM Journal on Optimization*, 27(4):2426–2446, 2017.

Ronny Luss and Marc Teboulle. Conditional gradient algorithms for rank-one matrix approximations with a sparsity constraint. *SIAM Review*, 55(1):65–98, 2013.

James MacQueen. Some methods for classification and analysis of multivariate observations. In *Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, 1967.

RV Mises and Hilda Pollaczek-Geiringer. Praktische verfahren der gleichungsau-flösung. *Zeitschrift für Angewandte Mathematik und Mechanik*, 9(1):58–77, 1929.

CL Muntz. Solution direct de l'equation seculaire et de quelques problems analogues. *Comptus Rendus de l'Academie des Sciences*, 156:43–46, 1913.

Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15(3):267–273, 1982.

Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23):3311–3325, 1997.

Prasanna K Sahoo and Palaniappan Kannappan. *Introduction to Functional Equations*. Chapman and Hall/CRC, 2011.

Ohad Shamir. A stochastic PCA and SVD algorithm with an exponential convergence rate. In *International Conference on Machine Learning*, pages 144–152, 2015.

Ohad Shamir. Fast stochastic algorithms for SVD and PCA: Convergence properties and convexity. In *International Conference on Machine Learning*, pages 248–256, 2016.

Richard A Tapia, John E Dennis Jr, and Jan P Schäfermeyer. Inverse, shifted inverse, and Rayleigh quotient iteration as Newton's method. *SIAM Review*, 60(1):3–55, 2018.

Lieven Vandenberghe and Stephen Boyd. Semidefinite programming. *SIAM Review*, 38(1): 49–95, 1996.

Yu-Xiong Wang and Yu-Jin Zhang. Nonnegative matrix factorization: A comprehensive review. *IEEE Transactions on Knowledge and Data Engineering*, 25(6):1336–1353, 2013.

James H Wilkinson. *The Algebraic Eigenvalue Problem*. Clarendon Press, Oxford, 1965.

Peng Xu, Bryan He, Christopher De Sa, Ioannis Mitliagkas, and Chris Re. Accelerated stochastic power iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 58–67, 2018.

Xiao-Tong Yuan and Tong Zhang. Truncated power method for sparse eigenvalue problems. *Journal of Machine Learning Research*, 14(Apr):899–925, 2013.