Scale Invariant Power Iteration

Cheolmin Kim[1], Youngseok Kim[2], and Diego Klabjan[1]

[1]Department of Industrial Engineering and Management Sciences, Northwestern University
[2]Department of Statistics, University of Chicago

Abstract

Power iteration has been generalized to solve many interesting problems in machine learning and statistics. Despite its striking success, theoretical understanding of when and how such an algorithm enjoys good convergence property is limited. In this work, we introduce a new class of optimization problems called scale invariant problems and prove that they can be efficiently solved by scale invariant power iteration (SCI-PI) with a generalized convergence guarantee of power iteration. By deriving that a stationary point is an eigenvector of the Hessian evaluated at the point, we show that scale invariant problems indeed resemble the leading eigenvector problem near a local optimum. Also, based on a novel reformulation, we geometrically derive SCI-PI which has a general form of power iteration. The convergence analysis shows that SCI-PI attains local linear convergence with a rate being proportional to the top two eigenvalues of the Hessian at the optimum. Moreover, we discuss some extended settings of scale invariant problems and provide similar convergence results for them. In numerical experiments, we introduce applications to independent component analysis, Gaussian mixtures, and non-negative matrix factorization. Experimental results demonstrate that SCI-PI is competitive to state-of-the-art benchmark algorithms and often yield better solutions.

1 Introduction

We consider a generalization of power iteration for finding the leading eigenvector of a matrix A. Power iteration repeats $x_{k+1} \leftarrow Ax_k/\|Ax_k\|$ until some stopping criterion is satisfied. Since no hyperparameter is required, this update rule is practical yet attains global linear convergence with the rate of $|\lambda_2|/|\lambda_1|$ where $|\lambda_i|$ is the i^{th} largest absolute eigenvalue of A. This linear convergence result 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 (VR) [Shamir, 2015, 2016, Kim and Klabjan, 2019b], and stochastic VR with momentum [Xu et al., 2018, Kim and Klabjan, 2019b] power iterations have been developed, drawing a parallel literature to gradient descent for convex optimization.

A general form of power iteration has been used to solve

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

in many applications such as sparse principal component analysis (PCA) [Journée et al., 2010, Luss and Teboulle, 2013], L_1-norm kernel PCA [Kim and Klabjan, 2019a], phase synchronization [Liu et al., 2017], and the Burer-Monteiro factorization of semi-definite programs [Erdogdu et al., 2018]. (All norms are 2-norms unless indicated otherwise.) Nevertheless, theoretical understanding of when such algorithms enjoy the attractive convergence property of power iteration is limited. Only global sublinear convergence has been shown for convex f [Journée et al., 2010], not generalizing the appealing linear convergence property of power iteration.

In view of manifold optimization [Absil et al., 2009], scale invariant problems (1) can be seen as an optimization problem on the real projective plane. Through reformulations, one can obtain an unconstrained optimization problem on the embedding space, which can be solved by general non-convex optimization algorithms such as gradient-based methods with line search or trust region methods. However, these algorithms require hyperparameters such as the step size while power iteration does not.

In this work, we introduce a new class of optimization problems called *scale invariant problems* and show that they can be efficiently solved by a general form of power iteration called *scale invariant power iteration* (SCI-PI) with a generalized convergence guarantee of power iteration. We say that an optimization problem is a scale invariant problem if the objective function f is *scale invariant* in (1). A function f is called scale invariant, which is rigorously defined later, if its geometric surface is invariant under constant multiplication of x. Many important 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), non-negative matrix factorization (NMF), and Gaussian mixture models (GMM) can be formulated as extended settings of scale invariant problems.

Derivatives of scale invariant functions have the interesting relation that $\nabla^2 f(x)x = k\nabla f(x)$ holds for some k. Using the KKT condition, we derive an eigenvector property stating that any stationary point x^* satisfying $\nabla f(x^*) = \lambda^* x^*$ for some λ^* is an eigenvector of $\nabla^2 f(x^*)$. Due to the eigenvector property, scale invariant problems can be locally seen as the leading eigenvector problem. Therefore, we can expect that a simple update rule like power iteration would efficiently solve scale invariant problems near a local optimum x^*. Another interesting property of scale invariant problems is that by swapping the objective function and the constraint, a geometrically interpretable dual problem with the goal of finding the closest point w to the origin from the constraint $f(w) = 1$ is obtained. 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 SCI-PI, which replaces Ax_k with $\nabla f(x_k)$ in power iteration. We show that SCI-PI converges to a local maximum x^* at a linear rate when initialized close to it. The convergence rate is proportional to $\bar{\lambda}_2$ / λ^* where $\bar{\lambda}_2$ is the spectral norm of $\nabla^2 f(x^*)(I - x^*(x^*)^T)$ and λ^* is the Lagrange multiplier corresponding to x^*, generalizing the convergence rate of power iteration. Moreover, under some mild conditions, we provide an explicit expression regarding the initial condition on $\|x_0 - x^*\|$ to ensure convergence.

In the extended settings, we discuss three variants of (1). In the first setting, we consider a sum of scale invariant functions 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 which covers 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 which include general mixture problems such as GMM. For this partially scale invariant problems, we present an alternative 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 KL-NMF, GMM and ICA. 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.

Our work has the following contributions.

1. We introduce scale invariant problems which cover interesting examples in statistics and machine learning yet can be efficiently solved by a general form of power iteration due to the eigenvector property.

2. We present a geometric derivation of SCI-PI and provide a convergence analysis for it. We show that SCI-PI converges to a local maximum x^* at a linear 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.

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 $a \neq 1$.

Proof. 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

$$g(r) = rg(1) \tag{6}$$

for all $r \geq 0$. From the definition of g and (6), we have

$$u(e^r) = e^{g(r)} = (e^r)^{g(1)}. \tag{7}$$

Representing $r > 0$ as $r = e^{\ln(r)}$ and using (7), 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 $r, s \in \mathbb{R}$. 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 $r \in \mathbb{R} \setminus \{0\}$. □

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*

$$c\nabla f(cx) = |c|^p \nabla f(x), \quad \nabla f(x)^T x = pf(x), \quad \nabla^2 f(x)x = (p-1)\nabla f(x). \tag{8}$$

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

$$c\nabla f(cx) = \nabla f(x), \quad \nabla f(x)^T x = \log^{-1}(a), \quad \nabla^2 f(x)x = -\nabla f(x). \tag{9}$$

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

$$f(cx) = u(c)f(x) + v(c) \tag{10}$$

since we can restore a multiplicatively or additively scale-invariant function by setting $v(c) = 0$ or $u(c) = 1$, respectively. By differentiating (10) with respect to x, we have

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

On the other hand, by differentiating (10) with respect to c, we have

$$\nabla f(cx)^T x = u'(c)f(x) + v'(c). \tag{11}$$

By differentiating (11) with respect to x, we obtain

$$c\nabla^2 f(cx)x + \nabla f(cx) = u'(c)\nabla f(x). \tag{12}$$

Plugging $c = 1$ into (11) and (12) completes the proof. □

Proposition 3 states that a scale invariant function satisfies $\nabla^2 f(x) = k\nabla f(x)$ holds for some k. This relation is interesting since using the first-order optimality conditions, we can derive an eigenvector property as follows.

Proposition 4. *Suppose that f is twice differentiable and let (λ^*, x^*) be a stationary point of (1) such that*

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

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. Consider the Lagrangian function

$$L(x, \lambda) = f(x) + \frac{\lambda}{2}\left(1 - \|x\|^2\right)$$

and a stationary point (λ^*, x^*) satisfying

$$\nabla f(x^*) = \lambda^* x^*, \quad \|x^*\| = 1.$$

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 d satisfying $d^T x^* = 0$, we have

$$d^T \nabla^2_{xx} L(x^*, \lambda^*) d = d^T \nabla^2 f(x^*)(I - x^*(x^*)^T)d - \lambda^* \|d\|^2 < 0.$$

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 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 a continuous function f is either multiplicatively scale invariant such that $f(x^*) > 0$ or additively scale invariant with an additive factor $u(c) = \log_a |c|$ with $a > 1$. Then, solving* (1) *is equivalent to solving the following optimization problem*

$$\text{minimize} \quad \|w\| \quad \text{subject to} \quad f(w) = 1. \tag{13}$$

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* (13)*. Conversely, if w^* is an optimal solution to* (13)*, $x^* = w^*/\|w^*\|$ is an optimal solution to* (1)*.*

Proof. First, we consider the case where an objective function f is multiplicative scale invariant with a multiplicative factor $u(c) = |c|^p$ where $p > 0$. Let w^* be an optimal solution to (13). From that $f(w^*) = 1$, we have $w^* \neq 0$, which leads to $\|w^*\| > 0$ and $f\left(w^*/\|w^*\|\right) = 1/\|w^*\|^p > 0$. Suppose an optimal solution to (1) is y with

$$f(y) > f\left(w^*/\|w^*\|\right) > 0. \tag{14}$$

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

$$f(y) = f\left(\frac{\hat{y}}{\|\hat{y}\|}\right) = \frac{1}{\|\hat{y}\|^{1/p}}, \quad f\left(\frac{w^*}{\|w^*\|}\right) = \frac{1}{\|w^*\|^{1/p}}. \tag{15}$$

From (14) and (15), we obtain $\|\hat{y}\| < \|w^*\|$, which contradicts that w^* is an optimal solution to (13).

On the other hand, let x^* be an optimal solution to (1) with $f(x^*) > 0$. Suppose that an optimal solution to (13) is z with

$$\|z\| < \|x^*/f(x^*)^{1/p}\|. \tag{16}$$

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

$$\|z\| = \|\hat{z}/f(\hat{z})^{1/p}\| = 1/f(\hat{z})^{1/p}, \quad \|x^*/f(x^*)^{1/p}\| = 1/f(x^*)^{1/p}. \tag{17}$$

From (16) and (17), we have

$$f(x^*) < f(\hat{z})$$

since $p > 0$, which contradicts the assumption that x^* is an optimal solution to (1).

Next, let f be an additively scale invariant function with an additive factor $v(c) = \log_a|c|$ with $a > 1$. In the same way as above, let w^* be an optimal solution to (13) and suppose that an optimal solution of (1) is y with

$$f(y) > f\left(w^*/\|w^*\|\right). \tag{18}$$

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

$$f(y) = f(\hat{y}) - \log_a\|\hat{y}\| = 1 - \log_a\|\hat{y}\|, \quad f\left(w^*/\|w^*\|\right) = 1 - \log_a\|w^*\|. \tag{19}$$

From (18) and (19), we have

$$\|\hat{y}\| < \|w^*\|$$

due to $a > 1$, contradicting the fact that w^* is an optimal solution to (13).

Conversely, let x^* be an optimal solution to (1) and suppose that an optimal solution to (13) is z with

$$\|z\| < \|a^{1-f(x^*)}x^*\|. \tag{20}$$

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

$$\|z\| = a^{1-f(\hat{z})}, \quad \|a^{1-f(x^*)}x^*\| = a^{1-f(x^*)}. \tag{21}$$

From (20) and (21), we have

$$f(x^*) < f(\hat{z})$$

due to $a > 1$, contradicting the assumption that x^* is an optimal solution to (1). □

Note that a dual reformulation for a multiplicatively scale invariant f with $f(x^*) < 0$ or an additively scale invariant f with $0 < a < 1$ can be obtained by replacing $f(w) = 1$ with $f(w) = -1$ in (13). The dual formulation (13) 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 6 (L_p-norm Kernel PCA)**.** *Given data vectors $a_i \in \mathbb{R}^d$ and a mapping Φ, L_p-norm PCA considers*

$$\text{maximize} \quad \frac{1}{n}\sum_{i=1}^n\|\Phi(a_i)^T x\|_p^p \quad \text{subject to} \quad x \in \partial\mathcal{B}_d \tag{22}$$

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

Example 7 (Estimation of Mixture Proportions)**.** *Given a design matrix $L \in \mathbb{R}^{n\times d}$ satisfying $L_{jk} \geq 0$, the problem of estimating mixture proportions seeks to find a vector π of mixture proportions on the probability simplex $\mathcal{S}^d = \left\{\pi : \sum_{k=1}^d \pi_k = 1,\ \pi \geq 0\right\}$ that maximizes the log-likelihood $\sum_{j=1}^n \log\left(\sum_{k=1}^d L_{jk}\pi_k\right)$. By reparametrizing π_k by x_k^2, we obtain an equivalent optimization problem*

$$\text{maximize} \quad \frac{1}{n}\sum_{j=1}^n \log\left(\sum_{k=1}^d L_{jk}x_k^2\right) \quad \text{subject to} \quad x \in \partial\mathcal{B}_d, \tag{23}$$

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

The reformulation idea in Example 7 implies that any simplex-constrained problem with scale invariant f can be reformulated to a scale invariant problem.

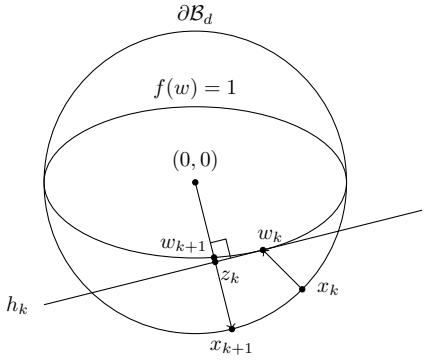
3 Scale Invariant Power Iteration

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 (13) as illustrated in Figure 1. Starting with an iterate $x_k \in \partial\mathcal{B}$, we obtain a dual iterate w_k by projecting x_k to the constraint $f(w) = 1$. Given w_k, we identify the hyperplane h_k which the current iterate w_k lies on and is tangent to $f(w) = 1$. After identifying the equation of h_k, we find the closest point z_k to the origin from h_k and obtain a new dual iterate w_{k+1} by projecting 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 continuous 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 tangent vector of the hyperplane h_k is $\nabla f(w_k)$. Therefore, we can write down the equation of the hyperplane h_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 h_k. Since w_{k+1} is the projection of z_k, it must be a scalar multiple of the tangent 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 $\partial\mathcal{B}$, we obtain

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

where the last equality follows from Proposition 3. Summarizing all the above, we obtain SCI-PI presented in Algorithm 1.



Algorithm 1 SCI-PI

Input: initial point x_0
for $k = 0, 1, \ldots, T-1$ **do**
$x_{k+1} \leftarrow \dfrac{\nabla f(x_k)}{\|\nabla f(x_k)\|}$
end for
Output: x_T

Figure 1: Geometric derivation of SCI-PI

Next, we provide a convergence analysis of SCI-PI.

Global sublinear convergence of SCI-PI for convex f has been addressed in Journée et al. [2010]. We additionally show that SCI-PI yields an ascent step even for quasi-convex f.

Proposition 8. *If f is quasi-convex and differentiable, a sequence of iterates $\{x_k\}_{k=0,1,\ldots}$ generated by SCI-PI satisfies $f(x_{k+1}) \geq f(x_k)$ for $k = 0, 1, \cdots$.*

Proof. If $f(x_{k+1}) < f(x_k)$, by the first-order condition of differentiable quasi-convex functions, we have

$$\nabla f(x_k)^T (x_{k+1} - x_k) = \nabla f(x_k)^T \left(\frac{\nabla f(x_k)}{\|\nabla f(x_k)\|} - x_k \right) = \|\nabla f(x_k)\| - \nabla f(x_k)^T x_k \leq 0. \tag{24}$$

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

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

This contradicts (24). Therefore, we should have $f(x_{k+1}) \geq f(x_k)$. □

If f is quasi-convex, the set $\{w : f(w) \leq 1\}$ is convex, therefore, from Figure 1, we can expect that SCI-PI would yield an ascent step. If f is not quasi-convex, $\{f(x_k)\}_{k=0,1,\ldots}$ is not necessarily increasing, making it hard to analyze global convergence. Assuming that an initial point x_0 is close to a local maximum x^*, we study local convergence of SCI-PI as follows.

Theorem 9. *Let f be a scale invariant, twice continuously differentiable function on an open set containing $\partial\mathcal{B}_d$ and let x^* be a local maximum satisfying $\nabla f(x^*) = \lambda^* x^*$ and $\lambda^* > \bar{\lambda}_2 = \max_{2\le i\le d}|\lambda_i|$ where (λ_i, v_i) is an eigen-pair of $\nabla^2 f(x^*)$ with $x^* = v_1$. Then, there exists some $\delta > 0$ such that under the initial condition $1 - x_0^T x^* < \delta$, the sequence of iterates $\{x_k\}_{k=0,1,\cdots}$ generated by SCI-PI satisfies*

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

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 $\nabla_i f = \partial f/\partial x_i$ has a continuous Hessian H_i on an open set containing $\mathcal{B}_{d,\infty} \triangleq \{x \in \mathbb{R}^d : \|x\|_\infty \le 1\}$, we can explicitly write δ as

$$\delta(\lambda^*, \bar{\lambda}_1, \bar{\lambda}_2, M) = \min\left\{\left(\frac{\lambda^*}{\bar{\lambda}_1 + M}\right)^2, \left(\frac{\lambda^* - \bar{\lambda}_2}{\bar{\lambda}_1 + 2M}\right)^2, 1\right\}$$

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

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

Proof. Since $\nabla^2 f(x^*)$ is real and symmetric, without loss of generality, we assume that $\{v_1, \ldots, v_d\}$ 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 $\nabla f(x)^T v_i$ at x^*, we have

$$\nabla f(x)^T v_i = \nabla f(x^*)^T v_i + (x - x^*)^T \nabla^2 f(x^*) v_i + R_i(x) \tag{25}$$

where

$$R_i(x) = o(\|x - x^*\|). \tag{26}$$

From $\nabla f(x^*) = \lambda^* x^*$ and $x^* = v_1$, we have

$$\begin{aligned}\nabla f(x)^T v_1 &= \nabla f(x^*)^T x^* + (x - x^*)^T \nabla^2 f(x^*) x^* + R_1(x)\\ &= \lambda^* - \lambda_1(1 - x^T x^*) + R_1(x)\\ &= \lambda^* + \alpha(x)\end{aligned} \tag{27}$$

where

$$\alpha(x) = -\lambda_1(1 - x^T x^*) + R_1(x) = o(\|x - x^*\|)$$

due to $R_1(x) = o(\|x - x^*\|)$ and $1 - x^T x^* = o(\|x - x^*\|)$.

On the other hand, for $2 \le i \le d$, due to $\nabla f(x^*) = \lambda^* x^*$, we have

$$\nabla f(x^*)^T v_i = \lambda^* (x^*)^T v_i = 0. \tag{28}$$

From (25), this results in

$$\nabla f(x)^T v_i = \lambda_i x^T v_i + R_i(x). \tag{29}$$

Let $\bar{R}_2(x) = \max_{2\le i\le d}|R_i(x)|$. Note that $\bar{R}_2(x) = o(\|x - x^*\|)$. By (29), we obtain

$$\begin{aligned}\sum_{i=2}^d \left(\nabla f(x)^T v_i\right)^2 &= \sum_{i=2}^d \left[\lambda_i^2 (x^T v_i)^2 + 2\lambda_i (x^T v_i) R_i(x) + (R_i(x))^2\right]\\ &\le \bar{\lambda}_2^2 \sum_{i=2}^d (x^T v_i)^2 + 2\bar{\lambda}_2 \bar{R}_2(x) \sum_{i=2}^d |x^T v_i| + d\left(\bar{R}_2(x)\right)^2.\end{aligned} \tag{30}$$

From $x \in \partial\mathcal{B}_d$, $x^* = v_1$, and the fact that $\{v_1, \ldots, v_d\}$ forms an orthogonal basis in $\mathbb{R}^d$, we have

$$\sum_{i=2}^{d} (x^T v_i)^2 = 1 - (x^T v_1)^2 = 1 - (x^T x^*)^2 \leq 2(1 - x^T x^*) = \|x - x^*\|^2.$$

Also, by the Cauchy Schwartz inequality, we have

$$\sum_{i=2}^{d} |x^T v_i| \leq \sqrt{d} \sqrt{\sum_{i=2}^{d} (x^T v_i)^2} \leq \sqrt{d}\|x - x^*\|.$$

Therefore, we obtain from (30) that

$$\begin{aligned} \sum_{i=2}^{d} \left(\nabla f(x)^T v_i\right)^2 &\leq \bar{\lambda}_2^2 \|x - x^*\|^2 + 2\bar{\lambda}_2 \bar{R}_2(x)\sqrt{d}\|x - x^*\| + d\left(\bar{R}_2(x)\right)^2 \\ &= \left(\bar{\lambda}_2 \|x - x^*\| + \beta(x)\right)^2 \end{aligned} \tag{31}$$

where

$$\beta(x) = \sqrt{d}\bar{R}_2(x) = o(\|x - x^*\|).$$

By (27), (31), and Lemma 18, we obtain the first part of the desired result.

Next, we consider the case where $\nabla_i f$ has a continuous Hessian H_i. From $\nabla_i f(x)$ being twice continuously differentiable in $\mathcal{B}_\infty$, we have

$$\nabla_i f(x_k) = \nabla_i f(x^*) + \nabla\nabla_i f(x^*)(x_k - x^*) + \frac{1}{2}(x_k - x^*)^T H_i(\hat{x}_k^i)(x_k - x^*) \tag{32}$$

where

$$\hat{x}_k^i \in \mathcal{N}(x_k, x^*) \triangleq \{x : x_s = t_s x_s^* + (1 - t_s)x_{k,s},\ 0 \leq t_s \leq 1,\ s = 1, \ldots, d\}.$$

In the above, x_s^* and $x_{k,s}$ denote the s^{th} coordinates of x^* and x_k, respectively.

For each $1 \leq i \leq d$, we have

$$\frac{1}{2}\sum_{j=1}^{d} v_{i,j}(x_k - x^*)^T H_j(\hat{x}_k^j)(x_k - x^*) = \frac{1}{2}(x_k - x^*)^T G_i(\hat{x}_k^1, \cdots, \hat{x}_k^d)(x_k - x^*).$$

From

$$\begin{aligned} &\left|(x_k - x^*)^T G_i(\hat{x}_k^1, \cdots, \hat{x}_k^d)(x_k - x^*)\right| \\ &\qquad = \|x_k - x^*\|^2 \left| \left[\frac{x_k - x^*}{\|x_k - x^*\|}\right]^T G_i(\hat{x}_k^1, \cdots, \hat{x}_k^d) \left[\frac{x_k - x^*}{\|x_k - x^*\|}\right] \right| \end{aligned} \tag{33}$$

and

$$\max_{x \in \partial\mathcal{B}_d} |x^T G_i(\hat{x}_k^1, \cdots, \hat{x}_k^d)x| \leq \max_{\substack{x \in \partial\mathcal{B}_d, \\ y^1, \cdots, y^d \in \mathcal{B}_\infty}} |x^T G_i(y^1, \cdots, y^d)x| \leq M,$$

we have

$$\left|(x_k - x^*)^T G_i(\hat{x}_k^1, \cdots, \hat{x}_k^d)(x_k - x^*)\right| \leq M\|x_k - x^*\|^2,$$

leading to

$$\frac{1}{2}\left|\sum_{j=1}^{d} v_{i,j}(x_k - x^*)^T H_j(\hat{x}_k^j)(x_k - x^*)\right| \leq \frac{1}{2}M\|x_k - x^*\|^2. \tag{34}$$

From (32), (34) and that $x^* = v_1$, we have

$$\nabla f(x_k)^T v_1 \geq \nabla f(x^*)^T x^* + (x_k - x^*)^T \nabla^2 f(x^*)x^* - \frac{M}{2}\|x_k - x^*\|^2,$$

resulting in

$$\nabla f(x_k)^T v_1 \geq \lambda^* - (M + |\lambda_1|)(1 - x_k^T x^*). \tag{35}$$

For $2 \leq i \leq d$, we have

$$\begin{aligned}\nabla f(x_k)^T v_i &= \nabla f(x^*)^T v_i + (x_k - x^*)^T \nabla^2 f(x^*) v_i + \frac{1}{2}(x_k - x^*)^T G_i(\hat{x}_k^1, \cdots, \hat{x}_k^d)(x_k - x^*) \\ &= \lambda_i x_k^T v_i + \frac{1}{2}(x_k - x^*)^T G_i(\hat{x}_k^1, \cdots, \hat{x}_k^d)(x_k - x^*). \end{aligned} \tag{36}$$

Using (33) and

$$\max_{x \in \partial \mathcal{B}_d} \sum_{i=2}^{d} (x^T G_i(\hat{x}_k^1, \cdots, \hat{x}_k^d) x)^2 \leq \max_{\substack{x \in \partial \mathcal{B}_d, \\ y^1, \cdots, y^d \in \mathcal{B}_\infty}} \sum_{i=2}^{d} (x^T G_i(y^1, \cdots, y^d) x)^2 \leq M,$$

we have

$$\sum_{i=2}^{d} \left[(x_k - x^*)^T G_i(\hat{x}_k^1, \cdots, \hat{x}_k^d)(x_k - x^*) \right]^2 \leq M^2 \|x_k - x^*\|^4. \tag{37}$$

Using (36), (37) and the Cauchy-Schwartz inequality, we have

$$\begin{aligned}\sum_{i=2}^{d} (\nabla f(x_k)^T v_i)^2 &\leq \sum_{i=2}^{d} \left(|\lambda_i| |x_k^T v_i| + \frac{1}{2}(x_k - x^*)^T G_i(\hat{x}_k^1, \cdots, \hat{x}_k^d)(x_k - x^*) \right)^2 \\ &\leq \bar{\lambda}_2^2 \sum_{i=2}^{d} (x_k^T v_i)^2 + \bar{\lambda}_2 M \|x_k - x^*\|^2 \sqrt{\sum_{i=2}^{d} (x_k^T v_i)^2} + \frac{M^2}{4} \|x_k - x^*\|^4 \\ &= \left(\bar{\lambda}_2 \sqrt{1 - (x_k^T x^*)^2} + \frac{M}{2} \|x_k - x^*\|^2 \right)^2. \end{aligned} \tag{38}$$

Using (35), (38), and Lemma 19 with

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

we obtain the desired result. □

Theorem 9 presents a local convergence result of SCI-PI with the rate being $\lambda^*/\bar{\lambda}_2$. For the leading eigenvector problem, this rate specializes to λ_1/λ_2, generalizing the convergence rate of power iteration. Note that Theorem 9 requires that a Lagrange multiplier λ^* corresponding to a local maximum x^* satisfies $\lambda^* > \bar{\lambda}_2 = \max_{2 \leq i \leq d} |\lambda_i|$. This assumption is satisfied by all local maxima if f is convex, multiplicatively scale invariant or concave, additively scale invariant. However, in general, not all local maxima satisfy this assumption since it is stronger than the local optimality condition stated as $\lambda^* > \max_{2 \leq i \leq d} \lambda_i$. 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 in shifted power iteration.

4 Extended Settings

4.1 Sum of Scale Invariant Functions

Consider a sum of scale invariant functions having the form of $f(x) = \sum_{i=1}^{m} g_i(x) + \sum_{j=1}^{n} h_j(x)$ where g_i is a multiplicatively scale invariant function with $u(c) = |c|^{p_i}$ and h_j is an additively scale invariant function with $v(c) = \log_{a_j} |c|$. Note that this does not imply that f is scale invariant in general. Here is an example that involves a sum of scale invariant functions.

Example 10 (Kurtosis-based ICA)**.** *Given a pre-processed data matrix $W \in \mathbb{R}^{n \times d}$, Kurtosis-based ICA [Hyvärinen and Oja, 2000] solves*

$$\text{maximize} \quad \frac{1}{n} \sum_{i=1}^{n} \left[(w_i^T x)^4 - 3 \right]^2 \quad \text{subject to} \quad x \in \partial \mathcal{B}_d. \tag{39}$$

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

By Proposition 3, the gradient of f has the form of

$$\nabla f(x) = \sum_{i=1}^{m} \nabla g_i(x) + \sum_{j=1}^{n} \nabla h_j(x) = F(x)x,$$

where

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

Note that 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 $F(x)$. We present a local convergence analysis of SCI-PI for a sum of scale invariant functions as follows.

Theorem 11. *Let f be a sum of scale invariant functions and twice continuously differentiable on an open set containing $\partial\mathcal{B}_d$ and let x^* be a local maximum satisfying $\nabla f(x^*) = \lambda^* x^*$ and $\lambda^* > \bar{\lambda}_2 = \|\nabla^2 f(x^*)(I - x^*(x^*)^2)\|$. Then, there exists some $\delta > 0$ such that under the initial condition $1 - x_0^T x^* < \delta$, the sequence of iterates $\{x_k\}_{k=0,1,\cdots}$ generated by SCI-PI satisfies*

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

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 $\nabla_i f = \partial f / \partial x_i$ has a continuous Hessian H_i on an open set containing $\mathcal{B}_{d,\infty}$, we can explicitly write δ as

$$\delta(\lambda^*, \bar{\lambda}_1, \bar{\lambda}_2, M) = \min \left\{ \left(\frac{\lambda^*}{\bar{\lambda}_1 + M} \right)^2, \left(\frac{\lambda^* - \bar{\lambda}_2}{\bar{\lambda}_1 + \bar{\lambda}_2 + 2M} \right)^2, 1 \right\}$$

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

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

Proof. By Proposition 3, the gradient of f has the form of

$$\nabla f(x) = \sum_{i=1}^{m} \nabla g_i(x) + \sum_{j=1}^{n} \nabla h_j(x) = F(x)x,$$

where

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

By the KKT conditions, a local optimal solution x^* is an eigenvector of $F(x^*)$. Let $\{v_1, \ldots, v_d\}$ be a set of eigenvectors of $F(x^*)$ with $x^* = v_1$. Since $F(x^*)$ is real and symmetric, without loss of generality, we assume that $\{v_1, \ldots, v_d\}$ 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 $\nabla f(x)^T v_i$ at x^*, we have

$$\nabla f(x)^T v_i = \nabla f(x^*)^T v_i + (x - x^*)^T \nabla^2 f(x^*) v_i + R_i(x) \tag{40}$$

where $R_i(x) = o(\|x - x^*\|)$. Using (40) with $i = 1$ and $\nabla f(x^*) = \lambda^* x^*$, we obtain

$$\begin{aligned} \nabla f(x)^T v_1 &= \lambda^* (x^*)^T v_1 + (x - x^*)^T \nabla^2 f(x^*) v_1 + R_1(x) \\ &= \lambda^* + \alpha(x) \end{aligned} \tag{41}$$

where

$$\alpha(x) = (x - x^*)^T \nabla^2 f(x^*) v_1 + R_1(x) = o(\sqrt{\|x - x^*\|}).$$

Using (40) and $\nabla f(x^*) = \lambda^* x^*$ for $2 \le i \le d$, we have

$$\begin{aligned}\nabla f(x)^T v_i &= \lambda^* (x^*)^T v_i + (x - x^*)^T \nabla^2 f(x^*) v_i + R_i(x) \\ &= (x - x^*)^T \nabla^2 f(x^*) v_i + R_i(x),\end{aligned}$$

resulting in

$$\sum_{i=2}^{d} (\nabla f(x)^T v_i)^2 = \sum_{i=2}^{d} \left((x - x^*)^T \nabla^2 f(x^*) v_i + R_i(x) \right)^2. \tag{42}$$

Let $\bar{R}_2(x) = \max_{2 \le i \le d} |R_i(x)|$. Note that $\bar{R}_2(x) = o(\|x - x^*\|)$.

From $x^* = v_1$ and the fact that $\{v_1, \ldots, v_d\}$ forms an orthogonal basis in $\mathbb{R}^d$, we have

$$\begin{aligned}\sum_{i=2}^{d} \left((x - x^*)^T \nabla^2 f(x^*) v_i \right)^2 &= \|\nabla^2 f(x^*)(x - x^*)\|_2^2 - \left((x - x^*)^T \nabla^2 f(x^*) v_1 \right)^2 \\ &= (x - x^*)^T \nabla^2 f(x^*) \left(I - x^* (x^*)^T \right) \nabla^2 f(x^*)(x - x^*) \\ &= (x - x^*)^T \nabla^2 f(x^*) \left(I - x^* (x^*)^T \right)^2 \nabla^2 f(x^*)(x - x^*).\end{aligned}$$

Since

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

we have

$$\sum_{i=2}^{d} \left((x - x^*)^T \nabla^2 f(x^*) v_i \right)^2 \le \bar{\lambda}_2^2 \|x - x^*\|^2. \tag{43}$$

Also, from (43) and the Cauchy-Schwartz inequality, we obtain

$$\sum_{i=2}^{d} (x - x^*)^T \nabla^2 f(x^*) v_i \le \sum_{i=2}^{d} |(x - x^*)^T \nabla^2 f(x^*) v_i| \le \bar{\lambda}_2 \sqrt{d} \|x - x^*\|. \tag{44}$$

Using (43) and (44) for (42), we obtain

$$\sum_{i=2}^{d} (\nabla f(x)^T v_i)^2 \le \bar{\lambda}_2^2 \|x - x^*\|^2 + 2\bar{\lambda}_2 \bar{R}_2(x) \sqrt{d} \|x - x^*\| + d(\bar{R}_2(x))^2,$$

resulting in

$$\sum_{i=2}^{d} (\nabla f(x)^T v_i)^2 \le \left(\bar{\lambda}_2 \|x - x^*\|^2 + \beta(x) \right)^2 \tag{45}$$

where

$$\beta(x) = \sqrt{d} \bar{R}_2(x) = o(\|x - x^*\|).$$

By (41), (45), and Lemma 18, we obtain the first part of the desired result.

Next, we assume that $\nabla_i f$ has a continuous Hessian H_i. By the Taylor theorem, we have

$$\nabla_i f(x_k) = \nabla_i f(x^*) + \nabla \nabla_i f(x^*)(x_k - x^*) + \frac{1}{2} (x_k - x^*)^T H_i(\hat{x}_k^i) (x_k - x^*) \tag{46}$$

for some $\hat{x}_k^i \in \mathcal{N}(x_k, x^*)$.

Taking the steps used to derive (34) and (37) in the proof of Theorem 9, we can derive the same inequalities

$$\frac{1}{2} \left| (x_k - x^*)^T G_i(\hat{x}_k^1, \cdots, \hat{x}_k^d)(x_k - x^*) \right| \le \frac{1}{2} M \|x_k - x^*\|^2 \tag{47}$$

and

$$\frac{1}{4}\sum_{i=2}^{d}\left[(x_k-x^*)^T G_i(\hat{x}_k^1,\cdots,\hat{x}_k^d)(x_k-x^*)\right]^2 \leq \frac{M^2}{4}\|x_k-x^*\|^4. \tag{48}$$

Using (46), (48) and that $x^* = v_1$, we have

$$\nabla f(x_k)^T v_1 \geq \nabla f(x^*)^T x^* + (x_k-x^*)^T\nabla^2 f(x^*)x^* - \frac{M}{2}\|x_k-x^*\|^2$$

resulting in

$$\begin{aligned}\nabla f(x_k)^T v_1 &\geq \lambda^* - \|\nabla^2 f(x^*)x^*\|\sqrt{2(1-x_k^T x^*)} - M(1-x_k^T x^*)\\ &= \lambda^* - \bar{\lambda}_1\sqrt{(1-x_k^T x^*)} - M(1-x_k^T x^*)\end{aligned} \tag{49}$$

For $2 \leq i \leq d$, we have

$$\begin{aligned}\nabla f(x_k)^T v_i &\leq \nabla f(x^*)^T v_i + (x_k-x^*)^T\nabla^2 f(x^*)v_i + \frac{1}{2}(x_k-x^*)^T G_i(\hat{x}_k^1,\cdots,\hat{x}_k^d)(x_k-x^*)\\ &= \lambda^*(x^*)^T v_i + (x_k-x^*)^T\nabla^2 f(x^*)v_i + \frac{1}{2}(x_k-x^*)^T G_i(\hat{x}_k^1,\cdots,\hat{x}_k^d)(x_k-x^*)\\ &= (x_k-x^*)^T\nabla^2 f(x^*)v_i + \frac{1}{2}(x_k-x^*)^T G_i(\hat{x}_k^1,\cdots,\hat{x}_k^d)(x_k-x^*).\end{aligned} \tag{50}$$

From (50), (43), (47), (48) and the Cauchy-Shwartz inequality, we have

$$\begin{aligned}\sum_{i=2}^{d}(\nabla f(x_k)^T v_i)^2 &\leq \sum_{i=2}^{d}\left((x_k-x^*)^T\nabla^2 f(x^*)v_i + \frac{1}{2}(x_k-x^*)^T G_i(\hat{x}_k^1,\cdots,\hat{x}_k^d)(x_k-x^*)\right)^2\\ &\leq \left(\bar{\lambda}_2\|x_k-x^*\| + \frac{M}{2}\|x_k-x^*\|^2\right)^2.\end{aligned} \tag{51}$$

Using (49), (51), and Lemma 19 with

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

we obtain the desired result. □

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 11 provides a generalization of the convergence rate in Theorem 9.

4.2 Block Scale Invariant Problems

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

$$\text{maximize} \quad f(x,y) \quad \text{subject to} \quad x \in \partial\mathcal{B}_{d_1},\ y \in \partial\mathcal{B}_{d_2}$$

where $f : \mathbb{R}^{d_1+d_2} \to \mathbb{R}$ is scale invariant in x for fixed y and vice versa. Some examples of block scale invariant problems are given next.

Example 12 (Semidefinite Programming (SDP) [Erdogdu et al., 2018])**.** *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

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

Example 13 (Kullback-Leibler (KL) divergence NMF)**.** *The KL-NMF problem [Févotte and Idier, 2011, Lee and Seung, 2001, Wang and Zhang, 2013] is defined as*

$$\begin{aligned}\text{minimize} \quad & D_{KL}(V\|WH) \triangleq \textstyle\sum_{i,j}\left[V_{ij}\log\frac{V_{ij}}{\sum_k W_{ik}H_{kj}} - V_{ij} + \sum_k W_{ik}H_{kj}\right]\\ \text{subject to} \quad & W_{ik} \geq 0,\ H_{kj} \geq 0,\ i = 1,\cdots,n,\ j = 1,\cdots,m,\ k = 1,\cdots,K.\end{aligned} \tag{52}$$

Many popular algorithms for the KL-NMF problem are based on alternate minimization of W and H. Given $W \geq 0$ and $j \in \{1, \cdots, m\}$, we consider a subproblem such that

$$\text{minimize} \quad f_{KL}(h) = \sum_i \left[v_i \log \frac{v_i}{\sum_k W_{ik} h_k} - v_i + \sum_k W_{ik} h_k \right] \text{ subject to } h_k \geq 0 \tag{53}$$

where we let $v_i = V_{ij}$ and $h_k = H_{kj}$ as the objective is decomposed into m separate subproblems. Note that the KL-NMF problem in the form of (52) 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 14. *The KL-NMF subproblem* (53) *is equivalent to the following scale invariant problem*

$$\text{maximize} \quad -\sum_i v_i \log \sum_k W_{ik} \bar{h}_k \quad \text{subject to} \quad \sum_k \bar{h}_k = 1, \quad \bar{h}_k \geq 0, \tag{54}$$

with the relationship $(\sum_i v_i)\bar{h}_k = (\sum_i W_{ik}) h_k$.

Proof. Since a log-linear function is concave, (53) is a convex problem in h. Consider the Lagrangian of the original problem

$$\mathcal{L}(h, \lambda) = f_{KL}(h) - \sum_k \lambda_k h_k \tag{55}$$

where $\lambda \geq 0$. By the first-order KKT conditions, we must have

$$\nabla_k f_{KL}(h^*) = \lambda_k^*, \quad \lambda_k^* h_k^* = 0, \quad \forall k = 1, \cdots, K \tag{56}$$

at an optimal solution (h^*, λ^*). Since (56) implies $\sum_k h_k^* \lambda_k^* = 0$, we have

$$\sum_k h_k^* \lambda_k^* = \sum_k h_k^* \nabla_k f_{KL}(h^*) = -\sum_{i,k} \frac{v_i W_{ik} h_k^*}{\sum_{k'} W_{ik'} h_{k'}^*} + \sum_{i,k} W_{ik} h_k^*,$$

resulting in

$$\sum_i v_i = \sum_{i,k} W_{ik} h_k^*. \tag{57}$$

Next, we show that

$$\text{minimize } f_{SCI}(h) = \sum_i v_i \log \frac{v_i}{\sum_k W_{ik} h_k} \text{ subject to } \sum_i v_i = \sum_{i,k} W_{ik} h_k, \; h_k \geq 0. \tag{58}$$

is equivalent to the original subproblem (53), due to the following:

1. It always satisfies $f_{SCI}^* \geq f_{KL}^*$ since (58) has an additional constraint $\sum_i v_i = \sum_{i,k} W_{ik} h_k$ compared to (53).

2. A solution h^* of (53) is a feasible point of (58) since we have shown that $\sum_i v_i = \sum_{i,k} W_{ik} h_k^*$. This implies $f_{KL}^* \geq f_{SCI}^*$.

Now, we can reparametrize h by $\bar{h}$ so that $\sum_i v_i = \sum_{i,k} W_{ik} h_k$ if and only if $\sum_k \bar{h}_k = 1$, which yields the relationship between two variables $\bar{h}_k = h_k(\sum_i W_{ik})/(\sum_i v_i)$. Note that (54) has the optimization problem as Example 7 and thus a scale invariant problem. □

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

$$x_{k+1} \leftarrow \nabla_x f(x, y_k)/\|\nabla_x f(x, y_k)\|, \quad y_{k+1} \leftarrow \nabla_y f(x_k, y)/\|\nabla_y f(x_k, y)\|. \tag{59}$$

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

Theorem 15. *Suppose that f is twice continuously differentiable on an open set containing $\partial\mathcal{B}_{d_1} \times \partial\mathcal{B}_{d_2}$ and let (x^*, y^*) be a local maximum satisfying*

$$\nabla_x f(x^*, y^*) = \lambda^* x^*, \; \lambda^* > \bar{\lambda}_2 = \max_{2 \leq i \leq d_1} |\lambda_i|, \; \nabla_y f(x^*, y^*) = s^* y^*, \; s^* > \bar{s}_2 = \max_{2 \leq i \leq d_2} |s_i|$$

where (λ_i, v_i) and (s_i, u_i) are eigen-pairs of $\nabla_x^2 f(x^*, y^*)$ and $\nabla_y^2 f(x^*, y^*)$, respectively with $x^* = v_1$ and $y^* = u_1$. If

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

then for the sequence of iterates $\{(x_k, y_k)\}_{k=0,1,\cdots}$ *generated by* (59)*, there exists some* $\delta > 0$ *such that if* $\max\{|1 - x_0^T x^*|, |1 - y_0^T y^*|\} < \delta$*, then we have*

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

where

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

Proof. From Lemma 20 with $w = x_k$, $z = y_k$, we have

$$1 - \frac{(\nabla_x f(x_k, y_k)^T x^*)^2}{\|\nabla_x f(x_k, y_k)\|^2} \leq \left(\frac{\bar{\lambda}_2}{\lambda^*} \sqrt{1 - (x_k^T x^*)^2} + \frac{\nu}{\lambda^*} \|y_k - y^*\| + \theta^x(x_k, y_k) \right)^2.$$

Since

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

we obtain

$$\sqrt{1 - (x_{k+1}^T x^*)^2} \leq \frac{\bar{\lambda}_2}{\lambda^*} \sqrt{1 - (x_k^T x^*)^2} + \frac{\nu}{\lambda^*} \|y_k - y^*\| + \theta^x(x_k, y_k).$$

Using

$$\|y_k - y^*\| = \sqrt{2(1 - y_k^T y^*)} = \left(1 + \frac{1 - y_k^T y^*}{1 + y_k^T y^* + \sqrt{2(1 + y_k^T y^*)}} \right) \sqrt{1 - (y_k^T y^*)^2},$$

we have

$$\sqrt{1 - (x_{k+1}^T x^*)^2} \leq \frac{\bar{\lambda}_2}{\lambda^*} \sqrt{1 - (x_k^T x^*)^2} + \frac{\nu}{\lambda^*} \sqrt{1 - (y_k^T y^*)^2} + \bar{\theta}^x(x_k, y_k) \tag{60}$$

where

$$\bar{\theta}^x(x_k, y_k) = \theta^x(x_k, y_k) + \left[\frac{1 - y_k^T y^*}{1 + y_k^T y^* + \sqrt{2(1 + y_k^T y^*)}} \right] \sqrt{1 - (y_k^T y^*)^2} = o\left(\left\| \begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix} \right\| \right).$$

Using Lemma 20 for $w = y_k$, $z = x_k$ and the definition of y_{k+1}, we have

$$\sqrt{1 - (y_{k+1}^T y^*)^2} \leq \frac{\nu}{s^*} \sqrt{1 - (x_k^T x^*)^2} + \frac{\bar{s}_2}{s^*} \sqrt{1 - (y_k^T y^*)^2} + \bar{\theta}^y(x_k, y_k) \tag{61}$$

where

$$\bar{\theta}^y(x_k, y_k) = \theta^y(x_k, y_k) + \left[\frac{1 - x_k^T x^*}{1 + x_k^T x^* + \sqrt{2(1 + x_k^T x^*)}} \right] \sqrt{1 - (x_k^T x^*)^2} = o\left(\left\| \begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix} \right\| \right).$$

Combining (60) and (61), we obtain

$$\begin{bmatrix} \sqrt{1 - (x_{k+1}^T x^*)^2} \\ \sqrt{1 - (y_{k+1}^T y^*)^2} \end{bmatrix} \leq \begin{bmatrix} \frac{\bar{\lambda}_2}{\lambda^*} & \frac{\nu}{\lambda^*} \\ \frac{\nu}{s^*} & \frac{\bar{s}_2}{s^*} \end{bmatrix} \begin{bmatrix} \sqrt{1 - (x_k^T x^*)^2} \\ \sqrt{1 - (y_k^T y^*)^2} \end{bmatrix} + \begin{bmatrix} \bar{\theta}^x(x_k, y_k) \\ \bar{\theta}^y(x_k, y_k) \end{bmatrix} \tag{62}$$

$$\leq (M + N(x_k, y_k)) \begin{bmatrix} \sqrt{1 - (x_k^T x^*)^2} \\ \sqrt{1 - (y_k^T y^*)^2} \end{bmatrix} \tag{63}$$

where

$$M = \begin{bmatrix} \frac{\bar{\lambda}_2}{\lambda^*} & \frac{\nu}{\lambda^*} \\ \frac{\nu}{s^*} & \frac{\bar{s}_2}{s^*} \end{bmatrix}, \quad \epsilon(x, y) = \frac{\max\{\bar{\theta}^x(x, y), \bar{\theta}^y(x, y)\}}{\sqrt{2 - x^T x^* - y^T y^*}},$$

and

$$N(x,y) = \frac{\epsilon(x,y)}{\sqrt{2 - x^T x^* - y^T y^*}} \begin{bmatrix} \sqrt{\frac{1-x^T x^*}{1+x^T x^*}} & \sqrt{\frac{1-y^T y^*}{1+y^T y^*}} \\ \sqrt{\frac{1-x^T x^*}{1+x^T x^*}} & \sqrt{\frac{1-y^T y^*}{1+y^T y^*}} \end{bmatrix}.$$

Note that the spectral radius ρ of M satisfies

$$\rho = \frac{1}{2}\left(\frac{\bar{\lambda}_2}{\lambda^*} + \frac{\bar{s}_2}{s^*} + \sqrt{\left(\frac{\bar{\lambda}_2}{\lambda^*} - \frac{\bar{s}_2}{s^*}\right)^2 + \frac{4\nu^2}{\lambda^* s^*}} \right) < 1$$

due to $\nu^2 < (\lambda^* - \bar{\lambda}_2)(s^* - \bar{s}_2)$. Also, for $i, j = 1, 2$, we have

$$\lim_{(x,y)\to(x^*,y^*)} N_{ij}(x,y) = 0.$$

By Lemma 22, there exists a sequence ω_t such that

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

Let

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

By Lemma 20, we have

$$\nabla_x f(x,y)^T v_1 = \lambda^* + (y - y^*)^T \nabla^2_{yx} f(x^*, y^*) x^* + \alpha^x(x,y)$$
$$\nabla_y f(x,y)^T u_1 = s^* + (x - x^*)^T \nabla^2_{xy} f(x^*, y^*) y^* + \alpha^y(x,y)$$

where

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

Therefore, there exists some $\delta_1 > 0$ such that if

$$x^T x^* > 0, \quad y^T y^* > 0, \quad \left\| \begin{bmatrix} \sqrt{1 - (x^T x^*)^2} \\ \sqrt{1 - (y^T y^*)^2} \end{bmatrix} \right\| < \delta_1,$$

then

$$\nabla_x f(x,y)^T v_1 > 0, \quad \nabla_y f(x,y)^T u_1 > 0. \tag{64}$$

Also, since $N_{ij}(x,y) \to 0$ as $(x,y) \to (x^*, y^*)$ for $i, j = 1, 2$, there exists some $\delta_2 > 0$ such that if

$$x^T x^* > 0, \quad y^T y^* > 0, \quad \left\| \begin{bmatrix} \sqrt{1 - (x^T x^*)^2} \\ \sqrt{1 - (y^T y^*)^2} \end{bmatrix} \right\| < \delta_2,$$

then we have

$$\left\| \prod_{l=0}^{\tau-1} \big(M + N(\phi(x,y,l))\big) \right\| < \bar{\rho}, \quad \max_{0<m\le\tau} \left\| \prod_{l=0}^{m-1} \big(M + N(\phi(x,y,l))\big) \right\| < 1 + \rho_{\max} \tag{65}$$

where $\phi(x,y,l)$ denotes the vector after l iterations of the algorithm starting with (x,y). To see this, let us define

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

By (63) and (64), 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

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

Therefore, there exists some $\delta_{2,\tau} > 0$ such that $g(x, y, \tau) < \bar{\rho}$. Also, for each $1 \leq m < \tau$, there exists some $\delta_{2,m} > 0$ such that $g(x, y, m) < 1 + \rho_{\max}$. Taking the minimum of $\delta_{2,m}$ for $1 \leq m \leq \tau$, we obtain δ_2 satisfying (65).

Let

$$\delta = \frac{\bar{\delta}}{\sqrt{2}}, \quad \bar{\delta} = \min\left\{\delta_1, \frac{\delta_1}{1+\rho_{\max}}, \delta_2, 1\right\}, \quad N_k = N(x_k, y_k).$$

By mathematical induction, we show that for any $n \geq 0$, if

$$x_{n\tau}^T x^* > 0, \quad y_{n\tau}^T y^* > 0, \quad \Delta_{n\tau} < \bar{\delta}, \tag{66}$$

then for $0 \leq m \leq \tau$, we have

$$x_{n\tau+m}^T x^* > 0, \quad y_{n\tau+m}^T y^* > 0, \quad \Delta_{n\tau+m} \leq (1+\rho_{\max})\Delta_{n\tau} < \delta_1. \tag{67}$$

By (66), it is obvious that we have (67) for $m = 0$. This proves the base case. Next, suppose that we have (67) for $0 \leq m < \tau$. Then, by the definition of δ_1, we have

$$x_{n\tau+m+1}^T x^* = x_{n\tau+m+1}^T v_1 = \frac{\nabla_x f(x_{n\tau+m}, y_{n\tau+m})^T v_1}{\|\nabla_x f(x_{n\tau+m}, y_{n\tau+m})\|} > 0$$

and

$$y_{n\tau+m+1}^T y^* = y_{n\tau+m+1}^T u_1 = \frac{\nabla_y f(x_{n\tau+m}, y_{n\tau+m})^T u_1}{\|\nabla_y f(x_{n\tau+m}, y_{n\tau+m})\|} > 0.$$

Also, by (63), (66) and (65), we have

$$\Delta_{n\tau+m+1} \leq \left\|\prod_{l=0}^{m}(M + N_{n\tau+l})\right\| \Delta_{n\tau} \leq (1+\rho_{\max})\Delta_{n\tau} < \delta_1.$$

This completes the induction proof.

Suppose that (x_0, y_0) satisfies $\max\{|1 - x_0^T x^*|, |1 - y_0^T y^*|\} < \delta$. Then, we have

$$x_0^T x^* > 0, \quad y_0^T y^* > 0, \quad \Delta_0 < \bar{\delta}. \tag{68}$$

Now, we show

$$x_{n\tau}^T x^* > 0, \quad y_{n\tau}^T y^* > 0, \quad \Delta_{n\tau} \leq \bar{\rho}^n \Delta_0. \tag{69}$$

For $n = 0$, we have (69) by (68). This proves the base case. Next, suppose that we have (69) for n. Then, since (69) implies that $\Delta_{n\tau} \leq \bar{\rho}^n \Delta_0 < \bar{\delta}$, by (67), we have

$$x_{(n+1)\tau}^T x^* > 0, \quad y_{(n+1)\tau}^T y^* > 0.$$

Moreover, using (63) and (65), we have

$$\Delta_{(n+1)\tau} \leq \left\|\prod_{l=0}^{\tau-1}(M + N_{n\tau+l})\right\| \Delta_{n\tau} \leq \bar{\rho}\Delta_{n\tau} < \bar{\rho}^{n+1}\Delta_0,$$

which completes the induction proof. By repeatedly applying (69), we have

$$(x_{n\tau}, y_{n\tau}) \to (x^*, y^*) \text{ as } n \to \infty.$$

Furthermore, due to (67), we have

$$(x_{n\tau+m}, y_{n\tau+m}) \to (x^*, y^*) \text{ for every } 0 < m \leq \tau,$$

indicating that

$$(x_k, y_k) \to (x^*, y^*).$$

This in turn implies that $N_k \to 0$. Letting

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

we have

$$\left\|\prod_{t=0}^{k-1}(M+N_t)\right\| = \prod_{t=0}^{k-1}(\rho+\omega_t+\eta_t) = \prod_{t=0}^{k-1}(\rho+\gamma_t). \tag{70}$$

Since $\eta_k \to 0$ as $N_k \to 0$, we have $\lim \gamma_k = 0$. This concludes the proof. □

If x and y are independent ($\nu = 0$), we have $\rho = \max\{\bar{\lambda}_2/\lambda^*, \bar{s}_2/s^*\}$. Otherwise, ρ increases as ν increases. Note that the result of Theorem 9 can be restored by dropping x or y in Theorem 15. While we consider the two-block case, 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

$$\text{maximize} \quad f(x,y) \quad \text{subject to} \quad x \in \partial\mathcal{B}_{d_1}$$

where $f(x,y) : \mathbb{R}^{d_1+d_2} \to \mathbb{R}$ is a scale invariant function in x for each $y \in \mathbb{R}^{d_2}$. 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.

Example 16 (Gaussian Mixture Model (GMM))**.** *The GMM problem is defined as*

$$\text{maximize} \quad \sum_{i=1}^{n} \log \sum_{k=1}^{d} x_k^2 \mathcal{N}(x_i; \mu_k, \Sigma_k) \quad \text{subject to} \quad x \in \partial\mathcal{B}_d.$$

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

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

$$x_{k+1} \leftarrow \nabla_x f(x_k,y_k)/\|\nabla_x f(x_k,y_k)\|, \quad y_{k+1} \leftarrow y_k + \alpha\nabla_y f(x_k,y_k). \tag{71}$$

While the gradient method is used in (71), any method for unconstrained optimization can replace it. We present a convergence analysis of (71) below.

Theorem 17. *Suppose that $f(x,y)$ is scale invariant in x for each $y \in \mathbb{R}^{d_2}$, μ-strongly concave in y with an L-Lipschitz continuous $\nabla_y f(x,y)$ for each $x \in \partial\mathcal{B}_{d_1}$, and three-times continuously differentiable on an open set containing $\partial\mathcal{B}_{d_1} \times \mathbb{R}^{d_2}$. Let (x^*,y^*) be a local maximum satisfying*

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

where (λ_i, v_i) is an eigen-pair of $\nabla^2 f(x^*)$ with $x^* = v_1$. If

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

then for the sequence of iterates $\{(x_k,y_k)\}_{k=0,1,\cdots}$ generated by (71) with $\alpha = 2/(L+\mu)$, there exists some $\delta > 0$ such that if $\max\{|1 - x_0^T x^*|, \|y - y^*\|\} < \delta$, then we have

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

where

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

Proof. Using Lemma 20 for $w = x_k$, $z = y_k$ and the definition of x_{k+1}, we have

$$\sqrt{1-(x_{k+1}^T x^*)^2} \leq \frac{\bar{\lambda}_2}{\lambda^*}\sqrt{1-(x_k^T x^*)^2} + \frac{\nu}{\lambda^*}\|y_k - y^*\| + \theta^x(x_k, y_k). \tag{72}$$

where

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

By Lemma 21 with $w = x_k, z = y_k$, we also have

$$\|y_{k+1} - y^*\| \leq \left(\frac{2\nu}{L+\mu}\right)\|x_k - x^*\| + \left(\frac{L-\mu}{L+\mu}\right)\|y_k - y^*\| + \theta^y(x_k, y_k). \tag{73}$$

Using

$$\bar{\theta}^y(x_k, y_k) = \theta^y(x_k, y_k) + \left[\frac{1 - x_k^T x^*}{1 + x_k^T x^* + \sqrt{2(1+x_k^T x^*)}}\right]\sqrt{1-(x_k^T x^*)^2} = o\left(\left\|\begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix}\right\|\right),$$

we can write (73) as

$$\|y_{k+1} - y^*\| \leq \left(\frac{2\nu}{L+\mu}\right)\sqrt{1-(x_k^T x^*)^2} + \left(\frac{L-\mu}{L+\mu}\right)\|y_k - y^*\| + \bar{\theta}^y(x_k, y_k). \tag{74}$$

Combining (72) and (74), we obtain

$$\begin{bmatrix} \sqrt{1-(x_{k+1}^T x^*)^2} \\ \|y_{k+1} - y^*\| \end{bmatrix} \leq \begin{bmatrix} \frac{\bar{\lambda}_2}{\lambda^*} & \frac{\nu}{\lambda^*} \\ \frac{2\nu}{L+\mu} & \frac{L-\mu}{L+\mu} \end{bmatrix}\begin{bmatrix} \sqrt{1-(x_k^T x^*)^2} \\ \|y_k - y^*\| \end{bmatrix} + \begin{bmatrix} \theta^x(x_k, y_k) \\ \bar{\theta}^y(x_k, y_k) \end{bmatrix} \tag{75}$$

$$\leq (M + N(x_k, y_k))\begin{bmatrix} \sqrt{1-(x_k^T x^*)^2} \\ \|y_k - y^*\| \end{bmatrix} \tag{76}$$

where

$$M = \begin{bmatrix} \frac{\bar{\lambda}_2}{\lambda^*} & \frac{\nu}{\lambda^*} \\ \frac{2\nu}{L+\mu} & \frac{L-\mu}{L+\mu} \end{bmatrix}, \quad \epsilon(x,y) = \frac{\max\{\theta^x(x,y), \bar{\theta}^y(x,y)\}}{\sqrt{1 - x^T x^* + \|y - y^*\|^2}}$$

and

$$N(x,y) = \frac{\epsilon(x,y)}{\sqrt{1 - x^T x^* + \|y - y^*\|^2}}\begin{bmatrix} \sqrt{\frac{1-x^T x^*}{1+x^T x^*}} & \|y - y^*\| \\ \sqrt{\frac{1-x^T x^*}{1+x^T x^*}} & \|y - y^*\| \end{bmatrix}.$$

Since $\nu^2 < \mu(\lambda^* - \bar{\lambda}_2)$, the spectral radius ρ of M satisfies

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

The rest of the proof is the same as the steps taken in the proof of Theorem 15. □

As in the result of Theorem 15, 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 9.

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 3 important applications, KL-NMF, GMM and ICA. A description of the data sets is provided below.

5.1 Description of Data Sets

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

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

For KL divergence nonnegative matrix factorization (Section 5.2), we use 4 public real data sets available online[1] 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.

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

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

For GMM (Section 5.3), we use 10 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. Only a small portion of entries are missing, if missing data exists, and we simply impute by mean.

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

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

For ICA, discussed also in Section 5.3, we use 9 public data sets (see Table 3) from the UCI Machine Learning repository[2]. 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-divergence NMF (KL-NMF) problem (52) described in Example 13. 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 (54). Our focus is to compare this algorithm with other well-known alternating minimization algorithms listed below, updating H and

[1] These 4 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`

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

W alternatively. To lighten the notation, let $\odot$, $\oslash$ and $(\cdot)^{\odot 2}$ denote element-wise product, division and square, respectively. We let $z = V \oslash (Wh)$ and $\mathbb{1}_n$ denote a vector of ones.

- 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].
- 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 $h^{\text{new}} \leftarrow h \odot (\sigma + W^T z)^{\odot 2}$ and rescales h, 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., 2018]. This algorithm performs sequential non-negative least squares.

KL-NMF Subproblem Note that the KL-NMF subproblem (54) has exactly the same form of the estimation of mixture proportions (23) described in the Example 7.

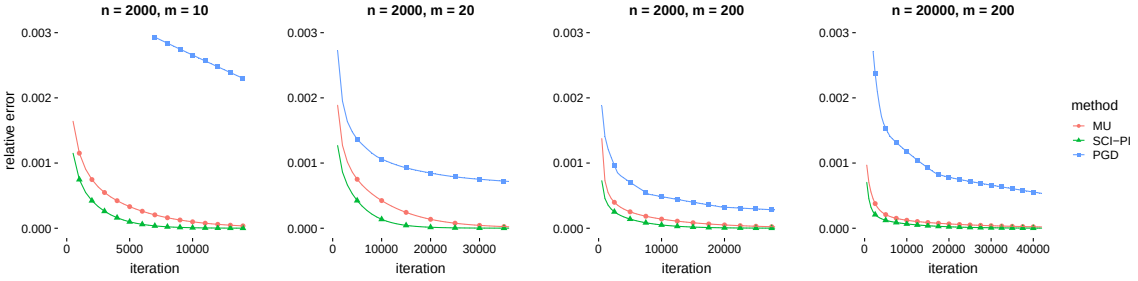


Figure 2: Convergence of 3 algorithms for the KL-NMF subproblem. n, m : the number of samples/features of the data matrix.

To study the convergence rate for the KL-NMF subproblems, we use the 4 data sets studied in Kim et al. [2018]. 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 $\|f(x_k) - f^*\| \leq 10^{-6} f^*$ where f^* is the solution obtained by MIXSQP after extensive computation time. The average runtime for aforementioned 3 methods are 33, 33 and 30 seconds for 10,000 iterations, respectively. The result is shown in Figure 2[3]. It shows that SCI-PI outperforms the other 2 for all simulated data sets. Also, all methods seems to exhibit linear convergence.

KL-NMF: Synthetic data sets We design a simple simulation study to evaluate the performance of block SCI-PI on KL-NMF problems. To this end, we sample a data matrix V independently from a single "zero-inflated" Poisson distribution (ZIP):

$$V_{ij} \sim \pi_0 \delta_0 + (1 - \pi_0)\text{Poisson}(l) \tag{77}$$

where π_0 is the proportion of zero inflation and l is the mean parameter of the Poisson distribution. Although this data generating distribution does not always reflect empirical distributions of real-world data sets, our focus here is to understand the behavior of SCI-PI compared to the other two methods, MU and PGD. We let n, m, l, K and π_0 vary to understand how the algorithms work for different settings. We also report the proportion s of zero entries of V.

Figure 3 summarizes the result. SCI-PI outperforms for two difference choices of $(n, m) = (500, 300)$ and $(1500, 100)$. We also conclude that SCI-PI tends to perform better in comparison to MU and PGD when V is denser ((1,1) vs. (1,3) and (2,1) vs. (2,2) in Figure 3), when K is larger ((1,1) vs. (2,3) in Figure 3) and when V is more uniformly distributed ((1,1) vs (1,2) and (1,1) vs (2,1) in Figure 3).

[3]For each evaluation, we randomly draw 10 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 Unif$(0, 1)$ random matrix.

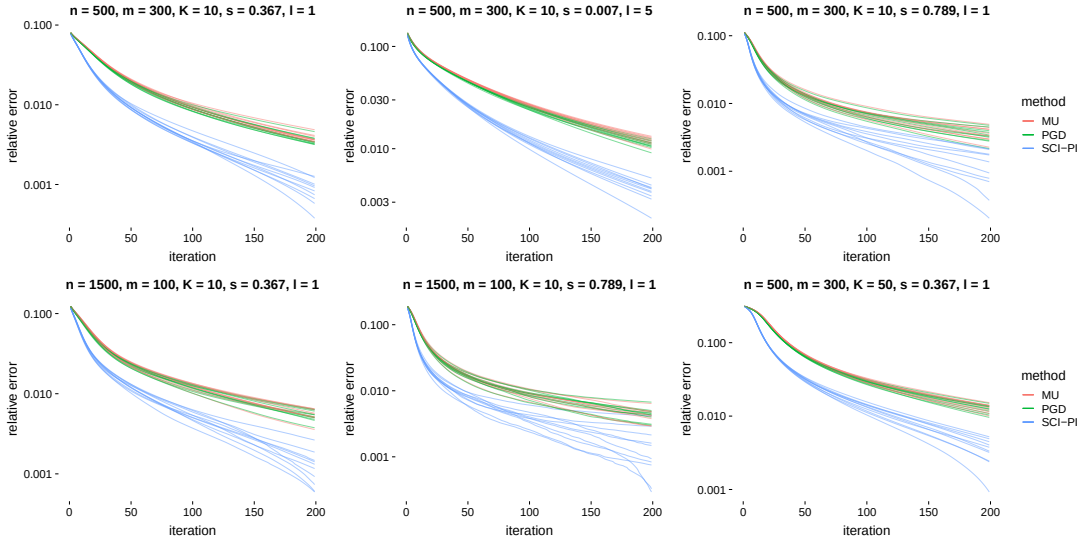


Figure 3: Convergence plots of 3 methods for KL-NMF on 6 synthetic data sets. We draw 10 convergence plots for each method differently initialized at random.

KL-NMF on Real world data sets Next, we test the 4 algorithms on the data sets in Table 1. We estimate $k = 20$ factors. At each iteration, all 4 algorithms solve m subproblems simultaneously for W and then alternatively for H.

The result is summarized in Figure 4[4]. The convergence plots are based on the average relative errors over 10 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 10 times[3]. The 3 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 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 3 algorithms, SCI-PI is not an ascent algorithm but an eigenvalue-based fixed-point algorithm. We observe that sometimes SCI-PI converges to a better solution due to this fact. 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 $L_{ik} = \mathcal{N}(x_i; \mu_k, \Sigma_k)$ where i is the sample index and k 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_k, Σ_k is replaced by the exact coordinate ascent step. The EM and SCI-PI updates for π can be written respectively as

$$r = \mathbb{1} \oslash (L\pi), \quad \pi_k^{\text{new}} \propto \pi \odot (L^T r) \quad \text{(EM)}, \quad \pi_k^{\text{new}} \propto \pi \odot (\alpha + L^T r)^{\odot 2} \quad \text{(SCI-PI)}. \tag{78}$$

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 10 times. The result is summarized in the left panel in Figure 5. The stopping criterion is $\|x_{k+1} - x_k\| < 10^{-8}$. In some cases, SCI-PI achieves much larger objective values even if initialized the same. In many cases the 2 algorithms exhibit the same performance. This is because estimation of μ_k's and Σ_k's are

[4]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 10 iterations where the plots start.

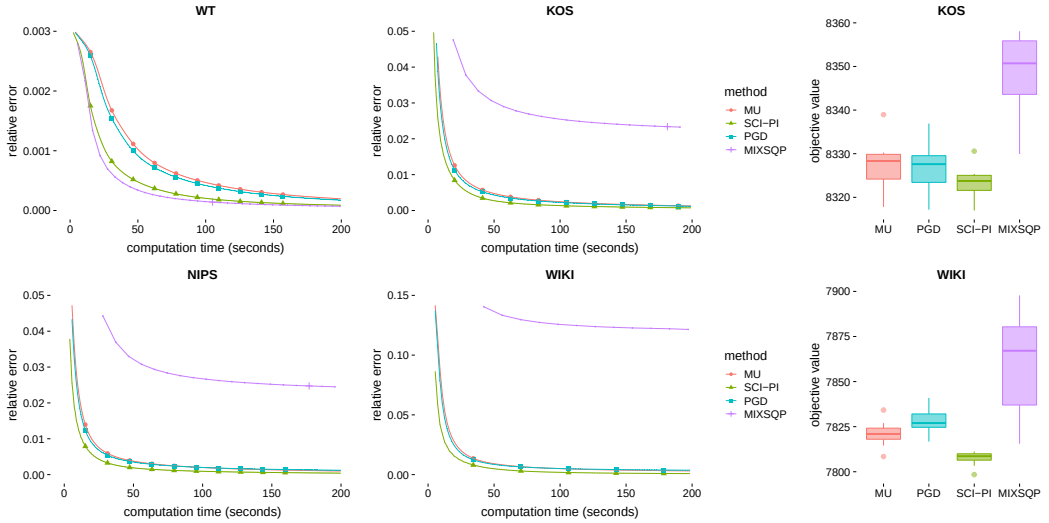


Figure 4: *(Left and center)* Convergence of the 4 NMF algorithms. *(Right)* Boxplots containing 10 objective values achieved after 400 seconds.

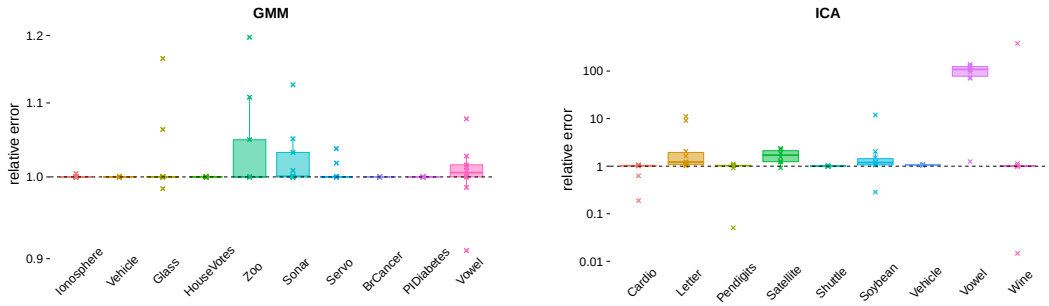


Figure 5: The relative error $f^*_{\text{SCI-PI}}/f^*_{\text{EM}}$ for GMM *(Left)* and $f^*_{\text{SCI-PI}}/f^*_{\text{FastICA}}$ for ICA *(Right)*.

usually harder than estimation of π, and EM and SCI-PI have the same 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[5] data matrix $W \in \mathbb{R}^{n\times d}$, we seek to maximize an approximated negative entropy $f(x) = \sum_{i=1}^{n} \left[(w_i^T x)^4 - 3\right]^2$ subject to $x \in \partial\mathcal{B}_d$, for maximizing Kurtosis-based non-Gaussianity [Hyvärinen and Oja, 2000]. 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 $x_{k+1} \leftarrow W^T(Wx_k)^{\odot 3} - 3(\mathbb{1}^T(Wx_k)^{\odot 2})x_k$, both followed by normalization.

In Figure 5 (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.

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

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. Finding more examples and extending SCI-PI further to a more general setting is a promising direction for future studies.

A Additional Lemmas

On several occasions, we use if $x \in \partial B_d$, $y \in \partial B_d$, then

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

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

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

By Cauchy-Schwarz, we also have

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

A.1 For the Proofs of Theorem 9 and Theorem 11

Lemma 18. *Let $\{v_1, \ldots, v_d\}$ be an orthogonal basis in $\mathbb{R}^d$ with $x^* = v_1$ and $\{x_k\}_{k=0,1,\cdots}$ be the sequence of iterates generated by SCI-PI. If for every $x \in \partial\mathcal{B}_d$ we have*

$$\nabla f(x)^T v_1 = \lambda^* + \alpha(x), \quad \sum_{i=2}^{d} (\nabla f(x)^T v_i)^2 \leq \left(\bar{\lambda}_2 \|x-x^*\| + \beta(x)\right)^2 \tag{79}$$

where

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

then there exists some $\delta > 0$ such that under the initial condition $1 - x_0^T x^* < \delta$, we have

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

Proof. By (79) for every $x \in \partial\mathcal{B}_d$, we have

$$\frac{\sum_{i=2}^{d} (\nabla f(x)^T v_i)^2}{(\nabla f(x)^T v_1)^2} \leq \left(\frac{\bar{\lambda}_2 \|x-x^*\| + \beta(x)}{\lambda^* + \alpha(x)}\right)^2.$$

Let

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

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

$$\frac{\sum_{i=2}^{d} (\nabla f(x)^T v_i)^2}{(\nabla f(x)^T v_1)^2} \leq \left(\frac{\bar{\lambda}_2}{\lambda^*} + \frac{\theta(x)}{\|x-x^*\|}\right)^2 \|x-x^*\|^2. \tag{80}$$

Letting

$$\epsilon(x) = \frac{\theta(x)}{\|x-x^*\|}, \tag{81}$$

we can further represent (80) as

$$\begin{aligned}\frac{\sum_{i=2}^{d} (\nabla f(x)^T v_i)^2}{(\nabla f(x)^T v_1)^2} &\leq \left(\frac{\bar{\lambda}_2}{\lambda^*} + \epsilon(x)\right)^2 \left(1 + \frac{1-x^T x^*}{1+x^T x^*}\right)\left(1-(x^T x^*)^2\right) \\ &= \left(\frac{\bar{\lambda}_2}{\lambda^*} + \gamma(x)\right)^2 \left(1-(x^T x^*)^2\right)\end{aligned} \tag{82}$$

where

$$\gamma(x) = \frac{\bar{\lambda}_2}{\lambda^*}\left(\frac{1-x^T x^*}{1+x^T x^* + \sqrt{2(1+x^T x^*)}}\right) + \epsilon(x)\sqrt{1+\frac{1-x^T x^*}{1+x^T x^*}}. \tag{83}$$

From (79), there exists some $\delta_1 > 0$ such that if $1 - x^T x^* < \delta_1$, then

$$\nabla f(x)^T v_1 > 0. \tag{84}$$

Also, by (81), for any $\bar{\gamma} > 0$ satisfying

$$\frac{\bar{\lambda}_2}{\lambda^*} + \bar{\gamma} < 1, \tag{85}$$

there exists some constant $\delta_2 > 0$ such that if $1 - x^T x^* < \delta_2$, then

$$|\epsilon(x)| \leq \frac{\bar{\gamma}}{4}. \tag{86}$$

Let $\delta = \min\{\delta_1, \delta_2, \frac{\lambda^*}{\bar{\lambda}_2}\bar{\gamma}, 1\}$. Before proving the main result, we first show the following two statements:

1. If $1 - x_k^T x^* < \delta$, then we have

$$x_{k+1}^T x^* > 0,\ 1 - (x_{k+1}^T x^*)^2 \leq \left(\frac{\bar{\lambda}_2}{\lambda^*} + \gamma_k\right)^2 \left(1 - (x_k^T x^*)^2\right),\ \text{and}\ \gamma_k \leq \bar{\gamma}. \tag{87}$$

Since $\delta < 1$, we have $x_k^T x^* > 0$. Also, from $1 - x_k^T x^* < \delta_1$ and $x^* = v_1$, using the update rule of SCI-PI and (84), we obtain

$$x_{k+1}^T x^* = \frac{\nabla f(x_k)^T x^*}{\|\nabla f(x_k)\|} = \frac{\nabla f(x_k)^T v_1}{\|\nabla f(x_k)\|} > 0.$$

On other the hand, since $|x_{k+1}^T v_1| \leq \|x_{k+1}\|\|v_1\| = 1$, we have

$$1 - (x_{k+1}^T x^*)^2 \leq \frac{1 - (x_{k+1}^T v_1)^2}{(x_{k+1}^T v_1)^2}.$$

Also, from the fact that $\{v_1, \ldots, v_d\}$ forms an orthogonal basis in $\mathbb{R}^d$, we have $\nabla f(x_k) = \sum_{i=1}^d (\nabla f(x_k)^T v_i) v_i$ and $\|\nabla f(x_k)\|^2 = \sum_{i=1}^d (\nabla f(x_k)^T v_i)^2$. Using the update rule of SCI-PI, we have

$$\frac{1 - (x_{k+1}^T v_1)^2}{(x_{k+1}^T v_1)^2} = \frac{\|\nabla f(x_k)\|^2 - (\nabla f(x_k)^T v_1)^2}{(\nabla f(x_k)^T v_1)^2} = \frac{\sum_{i=2}^d (\nabla f(x_k)^T v_i)^2}{(\nabla f(x_k)^T v_1)^2},$$

resulting in

$$1 - (x_{k+1}^T x^*)^2 \leq \frac{\sum_{i=2}^d (\nabla f(x_k)^T v_i)^2}{(\nabla f(x_k)^T v_1)^2}.$$

Let $\gamma_k = \gamma(x_k)$ and $\epsilon_k = \epsilon(x_k)$. Since $x_k^T x^* > 0$ and $1 - x_k^T x^* < \min\{\delta_2, \frac{\lambda^*}{\bar{\lambda}_2}\bar{\gamma}\}$, from (83), we have

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

2. Using mathematical induction, we show that if

$$1 - x_0^T x^* < \delta, \tag{88}$$

then, for all $k \geq 0$. we have

$$1 - x_k^T x^* < \delta. \tag{89}$$

By (88), we have $1 - x_0^T x^* < \delta$, which shows the base case. Next, suppose that $1 - x_k^T x^* < \delta$ holds. Then, we have (87). Also, from $\delta < 1$, we have $x_k^T x^* > 0$. Since

$$x_{k+1}^T x^* > 0,\ x_k^T x^* > 0,\ 1 - (x_{k+1}^T x^*)^2 \leq \left(\frac{\bar{\lambda}_2}{\lambda^*} + \bar{\gamma}\right)^2 \left(1 - (x_k^T x^*)^2\right) < 1 - (x_k^T x^*)^2$$

we have

$$1 - x_{k+1}^T x^* < 1 - x_k^T x^* < \delta,$$

which completes the induction proof.

Now, we prove the main statement. Since (89) holds for all $k \geq 0$, we can repeatedly apply (87) to obtain

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

Since

$$1-(x_k^T x^*)^2 < \left(\frac{\bar{\lambda}_2}{\lambda^*}+\bar{\gamma}\right)^{2k}\left(1-(x_0^T x^*)^2\right), \tag{90}$$

we have $(x_k^T x^*)^2 \to 1$. Moreover, from that $x_k^T x^* > 0$ for all $k \geq 0$ by (89), we have $x_k \to x^*$, and thus $\lim_{k\to\infty}\gamma_k = 0$ by (83). With (90), this gives the desired result. □

Lemma 19. *Let $\{v_1, \ldots, v_d\}$ be an orthogonal basis in $\mathbb{R}^d$. If $x^* = v_1$ and a sequence of iterates $\{x_k\}_{k=0,1,\cdots}$ generated by SCI-PI satisfies*

$$\nabla f(x_k)^T v_1 \geq A - B(1-x_k^T x^*) - C\sqrt{1-x_k^T x^*} \tag{91}$$

and

$$\sum_{i=2}^{d}(\nabla f(x_k)^T v_i)^2 \leq \left(D\sqrt{1-(x_k^T x^*)^2}+E\sqrt{2(1-x_k^T x^*)}+\frac{F}{2}\|x_k-x^*\|^2\right)^2 \tag{92}$$

where $A > 0$ and B, C, D, E, F are non-negative real numbers such that

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

Then, under the initial condition that $1 - x_0^T x^* < \delta$ where

$$\delta = \min\left\{\left(\frac{A}{B+C}\right)^2, \left(\frac{A-D-E}{B+C+E+F}\right)^2, 1\right\}, \tag{93}$$

we have

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

Proof. Before proving the main result, we first show the following two statements:

1. If $1 - x_k^T x^* < \delta$, then we have

$$x_{k+1}^T x^* > 0, 1-(x_{k+1}^T x^*)^2 < \left(\frac{D+E}{A}+\gamma_k\right)^2\left(1-(x_k^T x^*)^2\right), \frac{D+E}{A}+\gamma_k < 1 \tag{94}$$

for all $k \geq 0$ where

$$\gamma_k = \frac{\left(A(E+F)+(B+C)(D+E)\right)\sqrt{1-x_k^T x^*}}{A\left(A-(B+C)\sqrt{1-x_k^T x^*}\right)}. \tag{95}$$

Since $0 < x_k^T x^* \leq 1$, we have $\sqrt{1-x_k^T x^*} \geq 1 - x_k^T x^*$. Using $x^* = v_1$, the update rule of SCI-PI, (91), and the fact that $\delta \leq (A/(B+C))^2$, we have

$$x_{k+1}^T x^* = \frac{\nabla f(x_k)^T v_1}{\|\nabla f(x_k)\|} \geq \frac{A - B(1-x_k^T x^*) - C\sqrt{1-x_k^T x^*}}{\|\nabla f(x_k)\|} > 0 \tag{96}$$

since

$$\frac{A - B(1-x_k^T x^*) - C\sqrt{1-x_k^T x^*}}{\|\nabla f(x_k)\|} \geq \frac{A-(B+C)\sqrt{1-x_k^T x^*}}{\|\nabla f(x_k)\|} > 0.$$

Using the same arguments in Lemma 18, we have

$$1-(x_{k+1}^T x^*)^2 \le \frac{\sum_{i=2}^d (\nabla f(x_k)^T v_i)^2}{(\nabla f(x_k)^T v_1)^2}. \tag{97}$$

By (96), we have

$$A - B(1-x_k^T x^*) - C\sqrt{1-x_k^T x^*} > 0.$$

Therefore, by plugging (91) and (92) into (97) and using that $x_k^T x^* > 0$, we have

$$\begin{aligned}
1-(x_{k+1}^T x^*)^2 &\le \left(\frac{D\sqrt{1-(x_k^T x^*)^2} + E\sqrt{2(1-x_k^T x^*)} + \frac{F}{2}\|x_k - x^*\|^2}{A - B(1-x_k^T x^*) - C\sqrt{1-x_k^T x^*}} \right)^2 \\
&= \left(\frac{D + E\sqrt{1+\frac{1-x_k^T x^*}{1+x_k^T x^*}} + F\sqrt{\frac{1-x_k^T x^*}{1+x_k^T x^*}}}{A - B(1-x_k^T x^*) - C\sqrt{1-x_k^T x^*}} \right)^2 \left(1-(x_k^T x^*)^2\right) \\
&\le \left(\frac{D + E\left(1+\sqrt{1-x_k^T x^*}\right) + F\sqrt{1-x_k^T x^*}}{A - (B+C)\sqrt{1-x_k^T x^*}} \right)^2 \left(1-(x_k^T x^*)^2\right) \\
&= \left(\frac{D+E}{A} + \gamma_k \right)^2 \left(1-(x_k^T x^*)^2\right)
\end{aligned} \tag{98}$$

where we use the fact that $\sqrt{1+x} \le 1 + \sqrt{x}$ for $x \ge 0$ to derive the second inequality. Lastly, from

$$\sqrt{1-x_k^T x^*} < \sqrt{\delta} \le \frac{A-D-E}{B+C+E+F},$$

we have

$$\gamma_k < 1 - \frac{D+E}{A}.$$

2. Using mathematical induction, we show that if

$$1 - x_0^T x^* < \delta, \tag{99}$$

then, for all $k \ge 0$, we have

$$1 - x_k^T x^* < \delta. \tag{100}$$

By (99), we have $1-x_0^T x^* < \delta$, which proves the base case. Next, suppose that we have $1-x_k^T x^* < \delta$. Then, we have (94). Also, from $\delta < 1$, we have $x_k^T x^* > 0$. Since

$$x_{k+1}^T x^* > 0,\ x_k^T x^* > 0,\ 1-(x_{k+1}^T x^*)^2 < 1-(x_k^T x^*)^2,$$

we have

$$1 - x_{k+1}^T x^* < 1 - x_k^T x^* < \delta.$$

This completes the induction proof.

Now, we prove the main statement. Since (100) holds for all $k \ge 0$, by repeatedly applying (94), we obtain

$$1-(x_k^T x^*)^2 \le \prod_{t=0}^{k-1} \left(\frac{D+E}{A} + \gamma_t \right)^2 \left(1-(x_0^T x^*)^2\right), \text{ and } \frac{D+E}{A} + \gamma_k < 1. \tag{101}$$

Since $(D+E)/A+\gamma_k<1$ for all $k\geq 0$, $1-(x_k^Tx^*)^2$ is monotone decreasing, and so is $1-x_k^Tx^*$ by non-negativity. Moreover, from that γ_k is a monotone increasing function of $1-x_k^Tx^*$, we have $\gamma_{k+1}\leq\gamma_k$ for all $k\geq 0$, resulting in

$$\prod_{t=0}^{k-1}\left(\frac{D+E}{A}+\gamma_t\right)^2\leq\left(\frac{D+E}{A}+\gamma_0\right)^{2k}.$$

Since $(D+E)/A+\gamma_0<1$ by (94), we have $(x_k^Tx^*)^2\to 1$. Due to $x_k^Tx^*>0$ for all $k\geq 0$, this implies $x_k\to x^*$, and thus $\lim_{k\to\infty}\gamma_k=0$ due to (95). With (101), this gives the desired result. □

A.2 For the Proofs of Theorem 15 and Theorem 17

Lemma 20. *Suppose that $f(w,z)$ is scale invariant in $w\in\mathbb{R}^{d_w}$ for each $z\in\mathbb{R}^{d_z}$ and twice continuously differentiable on an open set containing $\partial\mathcal{B}_{d_w}\times\partial\mathcal{B}_{d_z}$. Let (w^*,z^*) be a point satisfying*

$$\nabla_w f(w^*,z^*)=\lambda_w^*w^*,\quad \lambda_w^*>\bar{\lambda}_2^w=\max_{2\leq i\leq d_w}|\lambda_i^w|,\quad w^*=v_1^w$$

where (λ_i^w,v_i^w) is an eigen-pair of $\nabla_{ww}^2f(w^*,z^*)$. Then, for any $w\in\partial\mathcal{B}_{d_w}$ and $z\in\partial\mathcal{B}_{d_z}$, we have

$$\nabla_w f(w,z)^Tv_1^w=\lambda_w^*+(z-z^*)^T\nabla_{zw}^2f(w^*,z^*)w^*+\alpha^w(w,z)$$

and

$$\sum_{i=2}^{d_w}(\nabla_w f(w,z)^Tv_i^w)^2\leq\left(\bar{\lambda}_2^w\sqrt{1-(w^Tw^*)^2}+\nu^{wz}\|z-z^*\|+\beta^w(w,z)\right)^2$$

where

$$\alpha^w(w,z)=o\left(\left\|\begin{bmatrix}w-w^*\\z-z^*\end{bmatrix}\right\|\right),\quad \beta^w(w,z)=o\left(\left\|\begin{bmatrix}w-w^*\\z-z^*\end{bmatrix}\right\|\right).$$

Therefore, we have

$$1-\frac{(\nabla_w f(w,z)^Tw^*)^2}{\|\nabla_w f(w,z)\|^2}\leq\left(\frac{\bar{\lambda}_2^w}{\lambda_w^*}\sqrt{1-(w^Tw^*)^2}+\frac{\nu^{wz}}{\lambda_w^*}\|z-z^*\|+\theta^w(w,z)\right)^2$$

where

$$\nu^{wz}=\|\nabla_{wz}^2f(w^*,z^*)\|,\quad \theta^w(w,z)=o\left(\left\|\begin{bmatrix}w-w^*\\z-z^*\end{bmatrix}\right\|\right).$$

Proof. Since $\nabla_{ww}^2f(w^*,z^*)$ is real and symmetric, without loss of generality, we assume that $\{v_1^w,\ldots,v_{d_w}^w\}$ forms an orthogonal basis in $\mathbb{R}^{d_w}$.

By Taylor expansion of $\nabla_w f(w,z)^Tv_i^w$ at (w^*,z^*), we have

$$\nabla_w f(w,z)^Tv_i^w=\nabla_x f(w^*,z^*)^Tv_i^w+\begin{bmatrix}w-w^*\\z-z^*\end{bmatrix}^T\begin{bmatrix}\nabla_{ww}^2f(w^*,z^*)\\\nabla_{zw}^2f(w^*,z^*)\end{bmatrix}v_i^w+R_i^w(w,z)$$

where

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

Using $\nabla_w f(w^*,z^*)=\lambda_w^*w^*$ and $w^*=v_1^w$, we have

$$\nabla_w f(w^*,z^*)^Tv_1^w=\lambda_w^*,\quad (w-w^*)^T\nabla_{ww}^2f(w^*,z^*)v_1^w=-\lambda_1^w(1-w_k^Tw^*).$$

Therefore, we obtain

$$\nabla_w f(w,z)^Tv_1^w=\lambda_w^*+(w-w^*)^T\nabla_{zw}^2f(w^*,z^*)w^*+\alpha^w(w,z)\tag{102}$$

where

$$\alpha^w(w,z)=R_1^w(w,z)-\lambda_1^w(1-w^Tw^*)=o\left(\left\|\begin{bmatrix}w-w^*\\z-z^*\end{bmatrix}\right\|\right).$$

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

$$\nabla_w f(w^*, z^*)^T v_i^w = \lambda_w^* (w^*)^T v_i^w = 0, \quad (w - w^*)^T \nabla_{ww}^2 f(w^*, z^*) v_i^w = \lambda_i^w w^T v_i^w,$$

resulting in

$$\nabla_w f(w, z)^T v_i^w = \lambda_i^w w^T v_i^w + (z - z^*)^T \nabla_{zw}^2 f(w^*, z^*) v_i^w + R_i^w(w, z). \tag{103}$$

From (103), we obtain

$$\begin{aligned}\sum_{i=2}^{d_w} (\nabla_w f(w, z)^T v_i^w)^2 = &\sum_{i=2}^{d_w} (\lambda_i^w)^2 (w^T v_i^w)^2 + \sum_{i=2}^{d_w} \left((z - z^*)^T \nabla_{zw}^2 f(w^*, z^*) v_i^w\right)^2 \\ &+ \sum_{i=2}^{d_w} (R_i^w(w, z))^2 + 2 \sum_{i=2}^{d_w} \lambda_i^w (w^T v_i^w)(z - z^*)^T \nabla_{zw}^2 f(w^*, z^*) v_i^w \\ &+ 2 \sum_{i=2}^{d_w} \lambda_i^w (w^T v_i^w) R_i^w(w, z) \\ &+ 2 \sum_{i=2}^{d_w} (z - z^*)^T \nabla_{zw}^2 f(w^*, z^*) v_i^w R_i^w(w, z).\end{aligned}$$

Since $\{v_1^w, \ldots, v_{d_w}^w\}$ forms an orthogonal basis in $\mathbb{R}^{d_w}$, with $w^* = v_1^w$ and $\|w\|^2 = 1$, we have

$$\sum_{i=2}^{d_w} (\lambda_i^w)^2 (w^T v_i^w)^2 \le (\bar{\lambda}_2^w)^2 \left(1 - (w^T w^*)^2\right)$$

and

$$\sum_{i=2}^{d_w} \left((z - z^*)^T \nabla_{zw}^2 f(w^*, z^*) v_i^w\right)^2 \le \|(z - z^*)^T \nabla_{zw}^2 f(w^*, z^*)\|^2 \le (\nu^{wz})^2 \|z - z^*\|^2.$$

Let $\bar{R}_2^w(w, z) = \max_{2 \le i \le d_w} |R_i^w(w, z)|$. Note that

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

Using the Cauchy-Shwartz inequality, we have

$$\sum_{i=2}^{d_w} \lambda_i^w (w^T v_i^w)(z - z^*)^T \nabla_{zw}^2 f(w^*, z^*) v_i^w \le \bar{\lambda}_2^w \nu^{wz} \|z - z^*\| \sqrt{1 - (w^T w^*)^2}.$$

Also, we have

$$\sum_{i=2}^{d_w} \lambda_i^w (w^T v_i^w) R_i^w(w, z) \le \bar{\lambda}_2^w \bar{R}_2^w(w, z) \sqrt{d_w} \sqrt{1 - (w^T w^*)^2}$$

and

$$\sum_{i=2}^{d_w} R_i^w(w, z)(z - z^*)^T \nabla_{zw}^2 f(w^*, z^*) v_i^w \le \nu^{wz} \bar{R}_2^w(w, z) \sqrt{d_w} \|z - z^*\|.$$

Therefore, we obtain

$$\sum_{i=2}^{d_w} (\nabla_w f(w, z)^T v_i^w)^2 \le \left(\bar{\lambda}_2^w \sqrt{1 - (w^T w^*)^2} + \nu^{wz} \|z - z^*\| + \beta^w(w, z)\right)^2 \tag{104}$$

where

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

Since $\{v_1^w, \ldots, v_{d_w}^w\}$ forms an orthogonal basis in $\mathbb{R}^{d_w}$ and $|w^T w^*| \le \|w\|\|w^*\| = 1$, we have

$$1 - \frac{(\nabla_w f(w,z)^T w^*)^2}{\|\nabla_w f(w,z)\|^2} \le \frac{\sum_{i=2}^{d_w} (\nabla_w f(w,z)^T v_i^w)^2}{(\nabla_w f(w,z)^T v_1^w)^2}.$$

Using (102) and (104), we have

$$\frac{\sum_{i=2}^{d_w} (\nabla_w f(w,z)^T v_i^w)^2}{(\nabla_w f(w,z)^T v_1^w)^2} \le \left(\frac{\bar{\lambda}_2^w}{\lambda_w^*} \sqrt{1 - (w^T w^*)^2} + \frac{\nu^{wz}}{\lambda_w^*} \|z - z^*\| + \theta^w(w,z) \right)^2$$

where

$$\theta^w(w,z) = \frac{\beta^w(w,z)}{\lambda_w^*} - \left(\frac{\bar{\lambda}_2^w \sqrt{1 - (w^T w^*)^2} + \nu^{wz} \|z - z^*\| + \sqrt{d_w} \beta^w(w,z)}{\lambda_w^*} \right) \cdot \left(\frac{(z - z^*)^T \nabla_{zw}^2 f(w^*, z^*) w^* + \beta^w(w,z)}{\lambda_w^* + (z - z^*)^T \nabla_{zw}^2 f(w^*, z^*) w^* + \beta^w(w,z)} \right).$$

Since

$$|(z - z^*)^T \nabla_{zw}^2 f(w^*, z^*) w^*| \le \nu^{wz} \|z - z^*\|,$$

we have

$$|(z - z^*)^T \nabla_{zw}^2 f(w^*, z^*) w^*| \sqrt{1 - (w^T w^*)^2} \le \frac{1}{2} \left(1 - (w^T w^*)^2\right) + \frac{1}{2} (\nu^{wz})^2 \|z - z^*\|^2$$

and

$$\nu^{wz} |(z - z^*)^T \nabla_{zw}^2 f(w^*, z^*) w^*| \|z - z^*\| \le (\nu^{wz})^2 \|z - z^*\|^2.$$

From

$$1 - (w^T w^*)^2 = o\left(\left\| \begin{bmatrix} w - w^* \\ z - z^* \end{bmatrix} \right\| \right), \quad \|z - z^*\|^2 = o\left(\left\| \begin{bmatrix} w - w^* \\ z - z^* \end{bmatrix} \right\| \right),$$

we finally obtain

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

This completes the proof. □

Lemma 21. *Suppose that $f(w,z)$ is μ-strongly concave in $z \in \mathbb{R}^{d_z}$ with an L-Lipschitz continuous $\nabla_z f(w,z)$ for each $w \in \partial\mathcal{B}_{d_w}$ and three-times continously differentiable with respect to x and y on an open set containing $\partial\mathcal{B}_{d_w}$ and $\mathbb{R}^{d_z}$, respectively. Let (w^*, z^*) be a point such that $\nabla_z f(w^*, z^*) = 0$. Then, for any $w \in \partial\mathcal{B}_{d_w}$ and $z \in \partial\mathcal{B}_{d_z}$, with $\alpha = 2/(L + \mu)$, we have*

$$\|z + \alpha \nabla_z f(w,z) - z^*\| \le \left(\frac{2\nu^{zw}}{L + \mu} \right) \|w - w^*\| + \left(\frac{L - \mu}{L + \mu} \right) \|z - z^*\| + \theta^z(w,z) \tag{105}$$

where

$$\nu^{zw} = \|\nabla_{zw}^2 f(w^*, z^*)\|, \quad \theta^z(w,z) = o\left(\left\| \begin{bmatrix} w - w^* \\ z - z^* \end{bmatrix} \right\| \right).$$

Proof. Let $\nabla_{z,i} f$ be the i^{th} coordinate of $\nabla_z f$ and

$$H_{z,i} = \begin{bmatrix} H_{z,i}^{ww} & H_{z,i}^{wz} \\ H_{z,i}^{zw} & H_{z,i}^{zz} \end{bmatrix}$$

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

$$\nabla_{z,i} f(w,z) = \nabla_{z,i} f(w^*, z) + \nabla_{zw,\cdot i}^2 f(w^*, z)^T (w - w^*) + R_i^z(w,z) \tag{106}$$

where $\nabla_{zw,\cdot i}^2 f(w^*, z) = \nabla_w \nabla_{z,i} f(w^*, z)$ denotes the i^{th} column of $\nabla_{zw}^2 f(w^*, z)$ and

$$R_i^z(w,z) = \frac{1}{2} (w - w^*)^T H_{z,i}^{ww}(\hat{w}^i, z)(w - w^*), \quad \hat{w}^i \in \mathcal{N}(w, w^*). \tag{107}$$

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

$$\nabla^2_{zw,\cdot i} f(w^*, z) = \nabla^2_{zw,\cdot i} f(w^*, z^*) + H^{wz}_{z,i}(w^*, \hat{z}^i)(z - z^*), \quad \hat{z}^i \in \mathcal{N}(z, z^*). \tag{108}$$

Since

$$\begin{aligned} |(z - z^*)^T H^{zw}_{z,i}(w^*, \hat{z}^i)(w - w^*)| &\leq \|H^{zw}_{z,i}(w^*, \hat{z}^i)\| \|w - w^*\| \|z - z^*\| \\ &\leq \frac{1}{2} \|H^{zw}_{z,i}(w^*, \hat{z}^i)\| \left(\|w - w^*\|^2 + \|z - z^*\|^2 \right), \end{aligned}$$

we have

$$(z - z^*)^T H^{wz}_{z,i}(w^*, \hat{z}^i)(w - w^*) = o\left(\left\| \begin{bmatrix} w - w^* \\ z - z^* \end{bmatrix} \right\| \right). \tag{109}$$

By (106), (107), (108), and (109), we have

$$\nabla_z f(w, z) = \nabla_z f(w^*, z) + \nabla^2_{zw} f(w^*, z^*)(w - w^*) + \bar{R}^z(w, z) \tag{110}$$

where

$$\bar{R}^z_i(w, z) = R^z_i(w, z) + (z - z^*)^T H^{zw}_{z,i}(w^*, \hat{z}^i)(w - w^*) = o\left(\left\| \begin{bmatrix} w - w^* \\ z - z^* \end{bmatrix} \right\| \right).$$

Using (110), we have

$$z + \alpha \nabla_z f(w, z) - z^* = z - z^* + \alpha \nabla_z f(w^*, z) + \alpha \nabla^2_{zw} f(w^*, z^*)(w - w^*) + \bar{R}^z(w, z),$$

resulting in

$$\begin{aligned} \|z + \alpha \nabla_z f(w, z) - z^*\| \leq \|z - z^* + \alpha \nabla_z f(w^*, z)\| \\ + \alpha \|\nabla^2_{zw} f(w^*, z^*)(w - w^*)\| + \|\bar{R}^z(w, z)\|. \end{aligned} \tag{111}$$

Since $-f(w^*, z)$ is μ-strongly convex in z with an L-Lipschitz continuous gradient $-\nabla_z f(w^*, z)$, by theory of convex optimization [Bubeck, 2015, p. 270], we have

$$\|z - z^* + \alpha \nabla_z f(w^*, z)\| \leq \left(\frac{L - \mu}{L + \mu} \right) \|z - z^*\| \tag{112}$$

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

$$\alpha \|\nabla^2_{zw} f(w^*, z^*)(w - w^*)\| \leq \left(\frac{2\nu^{zw}}{L + \mu} \right) \|w - w^*\|. \tag{113}$$

Plugging (112), (113) into (111), we finally obtain

$$\|z - z^* + \alpha \nabla_z f(w^*, z)\| \leq \left(\frac{L - \mu}{L + \mu} \right) \|z - z^*\| + \left(\frac{2\nu^{zw}}{L + \mu} \right) \|w - w^*\| + \theta^z(w, z)$$

where

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

□

Lemma 22. *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 ω_t such that

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

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

$$a = d, \quad e = 0.$$

Therefore, $M = aI$ and $\rho = a$. From $M^k = a^k I$, we have

$$\|M^k\| = \sqrt{a^{2k}} = \rho^k,$$

resulting in

$$\omega_k = \frac{\|M^{k+1}\|}{\|M^k\|} - \rho = \rho - \rho = 0$$

for all $k \geq 0$.

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

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

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

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

and

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

we have

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

leading to

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

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

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

From

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

we obtain the desired result. □

References

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

Christos Boutsidis, Dan Garber, Zohar Karnin, and Edo Liberty. Online Principal Components Analysis. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 887–901. Society for Industrial and Applied Mathematics, 2015.

Sébastien Bubeck. Convex Optimization: Algorithms and Complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 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.

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. *arXiv preprint arXiv:1807.04428*, 2018.

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

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

Aapo Hyvarinen. Fast ICA for Noisy Data using Gaussian Moments. In *Proceedings of the 1999 IEEE International Symposium on Circuits and Systems VLSI*, volume 5, pages 57–61. IEEE, 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*, volume 46. 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*, 2019a.

Cheolmin Kim and Diego Klabjan. Stochastic Variance-reduced Heavy Ball Power Iteration. *arXiv preprint arXiv:1901.08179*, 2019b.

Youngseok Kim, Peter Carbonetto, Matthew Stephens, and Mihai Anitescu. A Fast Algorithm for Maximum Likelihood Estimation of Mixture Proportions Using Sequential Quadratic Programming. *arXiv preprint arXiv:1806.01412*, 2018.

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.

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.

Erkki Oja. Simplified Neuron Model as a Principal Component Analyzer. *Journal of Mathematical Biology*, 15(3):267–273, 1982.

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.

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.

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.