Convergence Analysis of Gradient Descent for Eigenvector Computation

Zhiqiang Xu\textsuperscript{1}, Xin Cao\textsuperscript{2} and Xin Gao\textsuperscript{1}
\textsuperscript{1} KAUST, Saudi Arabia
\textsuperscript{2} UNSW, Australia
zhiqiang.xu@kaust.edu.sa, xin.cao@unsw.edu.au, xin.gao@kaust.edu.sa

Abstract
We present a novel, simple and systematic convergence analysis of gradient descent for eigenvector computation. As a popular, practical, and provable approach to numerous machine learning problems, gradient descent has found successful applications to eigenvector computation as well. However, surprisingly, it lacks a thorough theoretical analysis for the underlying geodesically non-convex problem. In this work, the convergence of the gradient descent solver for the leading eigenvector computation is shown to be at a global rate $O\left(\min\left\{ \left(\frac{\Delta_p}{\lambda_1}\right)^2 \log \frac{1}{\epsilon}, \frac{1}{\epsilon} \right\} \right)$, where $\Delta_p = \lambda_1 - \lambda_{p+1} > 0$ represents the generalized positive eigengap and always exists without loss of generality with $\lambda_1$ being the $i$-th largest eigenvalue of the given real symmetric matrix and $p$ being the multiplicity of $\lambda_1$. The rate is linear at $O\left(\left(\frac{\Delta_1^2}{\lambda_1}\right) \log \frac{1}{\epsilon} \right)$ if $\left(\frac{\Delta_1^2}{\lambda_1}\right) = O(1)$, otherwise sub-linear at $O\left(\frac{1}{\epsilon} \right)$. We also show that the convergence only logarithmically instead of quadratically depends on the initial iterate. Particularly, this is the first time the linear convergence for the case that the conventionally considered eigengap $\Delta_1 = \lambda_1 - \lambda_2 = 0$ but the generalized eigengap $\Delta_p$ satisfies $\left(\frac{\Delta_p^2}{\lambda_1}\right) = O(1)$, as well as the logarithmic dependence on the initial iterate are established for the gradient descent solver. We are also the first to leverage for analysis the log principal angle between the iterate and the space of globally optimal solutions. Theoretical properties are verified in experiments.

1 Introduction
Eigenvector computation is a ubiquitous problem in data processing nowadays, such as spectral clustering [Ng et al., 2002; Xu and Ke, 2016a], pagerank computation, dimensionality reduction [Fan et al., 2018], and so on. Classic solvers from numerical algebra are power methods and Lanczos algorithms [Golub and Van Loan, 1996], based on which there has been a recent surge of interest in developing varieties of solvers [Hardt and Price, 2014; Musco and Musco, 2015; Garber et al., 2016; Balcan et al., 2016]. However, most of them are purely theoretic. In this work, we focus on a general and practical solver [Wen and Yin, 2013], namely gradient descent from the optimization perspective, for the leading eigenvector computation:

$$\min_{x \in \mathbb{R}^n; \|x\|_2=1} f(x) = -x^TAx, \quad (1)$$

where $A \in \mathbb{R}^{n \times n}$ and $A^T = A$. Gradient descent is the most widely used method in optimization for machine learning [Sra et al., 2011], due to its effectiveness, simplicity, and provable theoretical guarantees. It has also found successful applications to eigenvector computation [Absil et al., 2008; Wen and Yin, 2013; Pitaval et al., 2015; Liu et al., 2016; Zhang et al., 2016; Xu and Ke, 2016b]. In particular, it was demonstrated to have comparable performance and better robustness in comparison to Lanczos algorithms with extensive experimental studies [Wen and Yin, 2013]. Despite being practical, however, its convergence analysis remains unsatisfied thus far. Pitaval et al. proved its global convergence but the rate was unknown [Pitaval et al., 2015]. Under a large positive eigengap assumption a linear but local rate was achieved in [Xu and Ke, 2016b; Liu et al., 2016; Xu et al., 2017], while a gap-free rate $O\left(\frac{1}{\epsilon^2}\right)$ was given by [Arora et al., 2013; Shamir, 2016a]. Very recently, the rate for the case of an arbitrary eigengap has been improved to $O\left(\frac{1}{\epsilon}\right)$ [Xu and Gao, 2018]. As shown in our experimental study, however, certain cases of zero eigengap are significantly underestimated and the quadratic dependence on the initial iterate is not true.

In this paper, we try to present a systematic convergence analysis of the gradient descent solver for Problem (1) to address all the issues mentioned above, and make it from a view of Riemannian optimization. The sphere constraint in (1) actually represents a Riemannian manifold, called the sphere manifold, which is a special case of the Stiefel manifold defined by the orthogonality constraint [Absil et al., 2008]. In this sense, Problem (1) is geodesically non-convex. A gradient descent step in Riemannian optimization takes the following form:

$$x_{t+1} = R(x_t, -\alpha_{t+1} \nabla f(x_t)), \quad (2)$$

where $\nabla f(x_t)$ denotes the Riemannian gradient representing the steepest ascent direction in the equidimensional Euclidean
space tangent to the manifold at $x_t$, $\alpha_{t+1} > 0$ represents the step-size, and $R(x_{t+1})$ represents the retraction map from the tangent space at $x_t$ to the manifold. In particular, in order to measure the progress of the iterate $x_t$ to one of globally optimal solutions, we use for analysis a novel potential function defined by the log principal angle between $x_t$ and, the space of such solutions rather than only one of such solutions as in previous work [Shamir, 2015]. It turns out that this potential function boasts the advantage of establishing the global convergence and improving the dependence on the initial iterate. We present a novel general analysis that depends on the generalized eigengap, $\Delta_p = \lambda_p - \lambda_{p+1}$, where $\lambda_l$ represents the $l$-th largest eigenvalue of the given matrix $A$ and $p$ is the multiplicity of $\lambda_1$. $\Delta_p$ always remains positive without loss of generality. This unifies all the cases of the conventionally considered eigengap $\Delta_1 = \lambda_1 - \lambda_2$, i.e., $\Delta_1 > 0$ and $\Delta_1 = 0$. When $\Delta_p$ is as small as the target precision parameter $\epsilon$, the resulting theoretic complexity is high. However, another unified analysis regardless of the value of $\Delta_p$ we present subsequently shows that this can be significantly reduced. Specifically, we make the following contributions:

- A global convergence rate $O(\min\{(\frac{\Delta_1}{\Delta_p})^2 \log \frac{1}{\epsilon}, \frac{1}{p}\})$ of the Riemannian gradient descent solver is achieved for Problem (1). The rate is linear at $O((\frac{\Delta_1}{\Delta_p})^2 \log \frac{1}{\epsilon})$ if $(\frac{\Delta_1}{\Delta_p})^2 = O(1)$, otherwise sub-linear at $O(\frac{1}{p})$. This shows that even if $\Delta_1 = 0$, it is possible as well to converge linearly.
- The quadratic dependence of the convergence on the initial iterate is improved to the logarithmic one.
- Theoretical properties, especially those related to $\Delta_1 = 0$, are empirically verified on synthetic or real data.

2 Related Work

Due to the space limit, we only discuss those gradient based solvers and refer readers to the cited references for more literature work. There are two categories of such solvers: projected gradient descent and Riemannian gradient descent. Arora et al. proposed the stochastic power method without theoretical guarantees [Arora et al., 2012], which actually is equivalent to the projected stochastic gradient descent for the principal component analysis (PCA) problem. It was named after the power method because the projected (deterministic) gradient descent for PCA will degenerate to the power method when the step-size goes to infinity. This method was subsequently extended via convex relaxation and proved to have a global, gap-free, and sub-linear convergence rate $O(\frac{1}{\epsilon^2})$ [Arora et al., 2013]. Balsubramani et al. demonstrated that Oja’s algorithm converges at a global, gap-dependent, and sub-linear rate $O(\frac{1}{\epsilon})$ via the martingale analysis [Balsubramani et al., 2013]. Note that the gap dependence refers to the dependence on $\Delta_1$ in most of existing analyses. More recently, stochastic gradient descent (SGD) for PCA was shown to converge either at a global, gap-dependent, and sub-linear rate $O(\frac{1}{\Delta_1^2})$ or at a global, gap-free, and sub-linear rate $O(\frac{1}{\Delta_1})$ [Shamir, 2016a]. Shamir proposed the projected SGD with variance reduction for PCA, called VR-PCA, and proved its global or local, gap-dependent, and linear rate $O(\frac{1}{\Delta_1^2} \log \frac{1}{\epsilon})$ [Shamir, 2015; 2016b].

More relevant to our work is an increasing body of Riemannian gradient descent methods. It is worth noting that general Riemannian optimization methods are applicable to Problem (1), including Riemannian gradient descent [Edelman et al., 1999; Absil et al., 2008; Wen and Yin, 2013], Riemannian SGD [Bonnabel, 2013] and Riemannian SVRG [Zhang et al., 2016]. Notably, Wen et al. demonstrated that curvilinear search, which actually performs Riemannian gradient descent with Cayley transformation based retraction, achieves a comparable performance and is more robust compared to Lanczos algorithms [Wen and Yin, 2013]. However, the analysis only, at most, achieves either a global, gap-free, and sub-linear convergence to critical points [Absil et al., 2008; Wen and Yin, 2013; Bonnabel, 2013; Zhang et al., 2016] or a local, gap-dependent, and linear convergence to globally optimal solutions [Absil et al., 2008], due to the geodesic non-convexity. As we know, each eigenvector corresponds to a critical point of the objective function in Problem (1) and thus they fail to meet the global optimality in theory. On the other hand, specifically, gradient descent for low-rank approximation was proven to converge globally but without any rate provided in [Pitaval et al., 2015]. Xu et al. established the local, gap-dependent, and linear rate $O(\frac{1}{\Delta_1} \log \frac{1}{\epsilon})$ of the Riemannian SVRG solver [Xu and Ke, 2016b; Xu et al., 2017]. By proving an explicit Łojasiewicz exponent at $\frac{1}{\epsilon}$ in [Liu et al., 2016], Liu et al. demonstrated a local, gap-dependent, and linear convergence of the Riemannian line-search methods for the quadratic problem under the orthogonality constraint, which includes Problem (1) as a special case. In addition, despite the motivation from optimization on Riemannian quotient manifolds, aleatoric as a SGD solver for streaming low-rank matrix approximation ends up with an update like the stochastic power method, and achieves a global, gap-dependent, and sub-linear rate $O(\frac{1}{\Delta_1^2})$ via the martingale analysis [Su et al., 2015].

Although gradient based methods for Problem (1) work well in practice [Arora et al., 2012; Wen and Yin, 2013], related theory is far behind. Despite great interest in developing stochastic solvers, surprisingly, there even does not exist a thorough convergence analysis for the deterministic gradient descent solver as is achieved in this work, which we believe will be an important basis for developing stochastic versions with better theoretical guarantees than the state-of-the-art.

3 Analysis of Gradient Descent

We now conduct the convergence analysis of the Riemannian gradient descent solver for Problem (1), starting from introducing necessary notions and notations. Main results are then stated in theorems and followed by a few important supporting lemmas. The section ends with the proofs of the theorem and lemmas.

3.1 Notions and Notations

Recall that given a real symmetric matrix $A \in \mathbb{R}^{n \times n}$, its $i$-th largest eigenvalue is denoted as $\lambda_i$, i.e., $\lambda_1 \geq \cdots \geq \lambda_n$, and
the corresponding eigenvector denoted as \( v_i \). Define eigen-gap \( \Delta_i = \lambda_i - \lambda_{i+1} \) and let \( p \) be the multiplicity of \( \lambda_i \) with the associated eigenspace denoted as \( V_p = \{ v_1, \ldots, v_p \} \). One then should have \( \Delta_j = 0 \) for \( j = 1, \ldots, p - 1 \) and \( \Delta_p > 0 \) without loss of generality, i.e., \( p < n \), because \( f(x) \) will be a constant function if \( p = n \).

Instead of focusing on a specific case of the eigengap like \( \Delta_1 > 0 \), we target a general framework admitting all the cases of \( \Delta_1 \). To this end, we have to rely on \( \Delta_p \) and \( V_p \), where we follow [Xu and Gao, 2018] to term \( \Delta_p \) as the generalized eigengap. That is, it suffices for us to show the convergence to one of globally optimal solutions, i.e., a unit vector \( v \in \text{span}(V_p) \), instead of a specific solution, e.g., \( v_1 \).

We thus define the following potential function:

\[
\psi(x_0) = -2 \log \| V_p^T x_0 \|_2, \tag{3}
\]

where actually \( \| V_p^T x_0 \|_2 = \cos \theta(x_0, V_p) \) representing the cosine of the principal angle between the current iterate \( x_t \) and the space of globally optimal solutions. As \( \| V_p^T x_0 \|_2 \leq 1 \), it is easy to see that \( \psi(x_0) \geq 0 \) and \( x_0 \) converges to a unit vector in \( \text{span}(V_p) \) when \( \psi(x_t) \) goes to 0.

In addition, we take the normalization retraction, i.e.,

\[
R(x, \xi) = \frac{x + \xi}{\| x + \xi \|_2}. \tag{4}
\]

Thus, the update in (2) can be explicitly written as

\[
x_{t+1} = x_t - \alpha_{t+1} \nabla \psi(x_t), \tag{5}
\]

where the Riemannian gradient for Problem (1) is \( \nabla \psi(x) = -(I - xx^T)Ax \).

### 3.2 Main Results

**Theorem 1.** If the initial \( x_0 = \frac{y}{\| y \|_2} \) where entries of \( y \) are i.i.d. standard normal samples, i.e., \( y_i \sim \mathcal{N}(0, 1) \), then the Riemannian gradient descent solver for Problem (1) will converge to certain unit vector \( v \in \text{span}(V_p) \) at a global rate \( O(\min\{\frac{\lambda}{2}, 1 + \frac{\alpha}{\Delta_p}\}) \) with high probability. Specifically, for any \( \epsilon \in (0, 1) \),

\( I) \) the solver with constant step-sizes \( \alpha < \frac{\epsilon_1}{\Delta_p} \) will converge, i.e., \( \psi(x_T) < \epsilon \), after \( T = O\left(\frac{\lambda}{\epsilon^2} \log \frac{\| x_0 \|_2}{\epsilon} \right) \) iterations, with probability at least \( 1 - \nu^2 \), where \( \nu = 0 \) if \( p = 1 \) otherwise \( \nu \in (0, 1) \);

\( II) \) the solver with diminishing step-sizes \( \alpha_t = \frac{\epsilon}{\Delta_p} \) will converge, i.e., \( \psi(x_T) < \epsilon \), after \( T = O\left(\frac{1}{\epsilon^2} \right) \) iterations, with probability at least \( 1 - \nu^p \), where \( \nu = 0 \) if \( p = 1 \) otherwise \( \nu \in (0, 1) \).

As the convergence holds with high probability for any random initial iterate, it is global by convention. Contrastingly, the success probability given in [Ishamir, 2016a] is \( \Omega(1) \).

To prove the theorem, we need a few important lemmas whose proofs are deferred to after that of Theorem 1.

**Lemma 2.** \( \lambda_1 - x^T Ax \geq \Delta_p \sin^2 \theta(x, V_p) \).

**Lemma 3.** \( \| x_t - \alpha_{t+1} \nabla \psi(x_t) \|_2^2 \leq 1 + \alpha_{t+1}^2 (\lambda_1^2 + \lambda_2^2) \sin^2 \theta(x_t, V_p) \).

**Lemma 4.** \( -\frac{x}{\log(1-x)} \geq 1 - \frac{1}{1 - \log(1-x)} \) for any \( x \in (0, 1) \), and \( \frac{x}{1+x} \leq \log(1+x) \leq x \) for any \( x > -1 \).

**Lemma 5.** \( \| V_p^T x_0 \|_2 > 0 \) with probability 1 if \( p = 1 \), otherwise with probability at least \( 1 - \nu^p \), where \( 0 < \nu < 1 \) is a constant about \( V_p^T y \).

**Proof of Theorem 1.**

Proof. 1) We start from expanding \( \psi(x_{t+1}) \) using (3)-(4):

\[
\psi(x_{t+1}) = -2 \log \| V_p^T x_{t+1} \|_2,
\]

\[
= -2 \log \| V_p^T (x_t - \alpha_{t+1} \nabla \psi(x_t)) \|_2 + 2 \log \| x_t - \alpha_{t+1} \nabla \psi(x_t) \|_2.
\]

To proceed, we need the full eigen-decomposition of \( A \), i.e.,

\[
A = V_p^T V_p + V_p^\perp \text{diag}(\lambda_{p+1}, \ldots, \lambda_n)(V_p^\perp)^T,
\]

where \( V_p^\perp \) represents the orthogonal complement of \( V_p \). Plugging in this decomposition to \( V_p^T A \), one then gets

\[
-V_p^T \nabla \psi(x_t) = V_p^T (I - x_t x_t^T) A x_t
\]

\[
= V_p^T A x_t - (x_t^T A x_t) V_p^T x_t
\]

\[
= (\lambda_1 - x_t^T A x_t) V_p^T x_t.
\]

We now can write

\[
\psi(x_{t+1}) = \psi(x_t) - 2 \log(1 + \alpha_{t+1} (\lambda_1 - x_t^T A x_t)) + 2 \log \| x_t - \alpha_{t+1} \nabla \psi(x_t) \|_2,
\]

where we have

\[
-2 \log(1 + \alpha_{t+1} (\lambda_1 - x_t^T A x_t))
\]

\[
\leq -2 \log(1 + \alpha_{t+1} \Delta_p \sin^2 \theta(x_t, V_p)) \quad \text{(by Lemma 2)}
\]

\[
\leq -2 \alpha_{t+1} \Delta_p \sin^2 \theta(x_t, V_p) \quad \text{(by Lemma 4)}
\]

\[
\leq -2 \alpha_{t+1} \Delta_p \sin^2 \theta(x_t, V_p),
\]

and

\[
2 \log \| x_t - \alpha_{t+1} \nabla \psi(x_t) \|_2
\]

\[
\leq \log(1 + \alpha_{t+1} \Delta_p \sin^2 \theta(x_t, V_p)) \quad \text{(by Lemma 3)}
\]

\[
\leq \alpha_{t+1} \Delta_p \sin^2 \theta(x_t, V_p) \quad \text{(by Lemma 4)}
\]

\[
\leq 2 \alpha_{t+1} \Delta_p \sin^2 \theta(x_t, V_p).
\]

One thus can arrive at

\[
\psi(x_{t+1}) \leq \psi(x_t) - 2 \alpha_{t+1} \Delta_p \sin^2 \theta(x_t, V_p)
\]

\[
+ 2 \alpha_{t+1} \lambda_1^2 \sin^2 \theta(x_t, V_p) \quad \text{(6)}
\]

\[
= \psi(x_t) - \left( \frac{2 \Delta_p}{1 + \alpha_{t+1} \Delta_p} - 2 \alpha_{t+1} \lambda_1^2 \right)
\]

\[
\cdot \frac{\sin^2 \theta(x_t, V_p)}{\psi(x_t)} \cdot \psi(x_t).
\]
Note that by Lemma 4, 
\[
\frac{\sin^2 \theta(x_t, V_p)}{\psi(x_t)} = \frac{\sin^2 \theta(x_t, V_p)}{-\log(1 - \sin^2 \theta(x_t, V_p))} \geq \frac{1}{1 - \log(1 - \sin^2 \theta(x_t, V_p))} = \frac{1}{1 + \psi(x_t)}.
\]

If \( \frac{\Delta_p}{1 + \alpha + \lambda_1^2} > 2\alpha + \lambda_1^2 > 0 \), i.e., \( \alpha_{t+1} < \frac{\Delta_p}{2\alpha + \lambda_1^2} \), then \( \psi(x_{t+1}) < \psi(x_t) \) and
\[
\psi(x_{t+1}) \leq \psi(x_t) - \frac{\Delta_p}{1 + \alpha + \lambda_1^2} \frac{\alpha_{t+1}}{1 + \psi(x_t)} \psi(x_t)
\]
(8)
\[
\leq (1 - \frac{\Delta_p}{1 + \alpha + \lambda_1^2} \frac{\alpha_{t+1}}{1 + \psi(x_t)}) \psi(x_t)
\]
(9)

if \( \alpha_{t+1} \equiv \alpha \). We thus have
\[
\psi(x_t) \leq (1 - \frac{\Delta_p}{1 + \alpha \lambda_1^2} \frac{\alpha}{1 + \psi(x_0)})^t \psi(x_0)
\]
(10)
and
\[
T \geq (1 + \alpha \lambda_1^2)(1 + \psi(x_0)) \log \frac{\psi(x_0)}{\epsilon},
\]
we must have \( \psi(x_T) < \epsilon \). Plugging in \( \alpha < \frac{\Delta_p}{2\alpha + \lambda_1^2} \) to \( T \), we can write
\[
T = O\left(\frac{1}{\alpha \lambda_1^2} \log \frac{\psi(x_0)}{\epsilon}\right)
\]
(11)
Finally note that (10) is equivalent to \( \|V_p^T x_0\|_2 > 0 \), which occurs with probability 1 if \( p = 1 \), otherwise at least \( 1 - \nu^p \) for certain constant \( \nu \in (0, 1) \), by Lemma 5.

II) We now restart from (6). Plugging in (7) to (6), we can write
\[
\psi(x_{t+1}) \leq \psi(x_t) - \frac{2\alpha_{t+1} \Delta_p}{1 + \alpha_{t+1} \lambda_1^2} \sin^2 \theta(x_t, V_p)
\]
(11)
\[
+ 2\alpha_{t+1} \lambda_1^2 \sin^2 \theta(x_t, V_p) \leq \psi(x_t) - \frac{2\alpha_{t+1} \Delta_p}{1 + \alpha_{t+1} \lambda_1^2} \psi(x_t) + 2\lambda_1^2 \alpha_{t+1}^2.
\]
Let \( \alpha_{t+1} = \frac{\alpha_{t+1}}{\lambda_1^2} \) where \( c, \tau > 0 \) are constants. It is easy to see that for sufficiently small \( \alpha_{t+1} \), equivalently, sufficiently large \( \tau \), we have \( \psi(x_{t+1}) < \psi(x_t) \). Thus, one gets
\[
\psi(x_{t+1}) \leq (1 - \frac{2\alpha_{t+1} \Delta_p}{1 + \frac{\tau}{\alpha \lambda_1^2}} \frac{1}{1 + \psi(x_0)}) \psi(x_t) + 2\lambda_1^2 \alpha_{t+1}^2.
\]

Denoting \( a = \frac{2\alpha \Delta_p}{1 + \frac{\tau}{\alpha \lambda_1^2}} \frac{1}{1 + \psi(x_0)} \) and \( b = 2\lambda_1^2 \), one can write
\[
\psi(x_{t+1}) \leq (1 - \frac{a}{\tau + t}) \psi(x_t) + \frac{b}{(\tau + t)^2}.
\]

As long as \( \psi(x_0) < +\infty \) and \( a > 1 \), by Lemma D.1 in [Balsubramani et al., 2013], the recursion of the above inequality will yield
\[
\psi(x_t) \leq \frac{t}{\tau + t + 1})^n \psi(x_0) + \frac{\alpha_{t+1}^2}{1 - a} \frac{1}{\tau + t + 1}.
\]

We thus have \( \psi(x_T) = O\left(\frac{1}{(\tau + t + 1)^n}\right) \psi(x_0) \), i.e., \( T = O\left(\frac{1}{\tau + t + 1}\right) \). The proof completes by noting that \( \psi(x_0) < +\infty \) occurs with high probability similarly and \( a > 1 \) can be guaranteed by choosing sufficiently large constants \( c, \tau \). \( \square \)

Remark We make a few remarks on our proof. Equation (8) shows that the convergence is at least linear because \( \psi(x_{t+1}) < \psi(x_t) \) and thus the contraction factor keep decreasing for constant step-sizes. Equation (9), \( \alpha_{t+1} \) is not necessarily constant. In fact, the results will hold for any step-size scheme as long as \( \alpha_{t+1} < \frac{\Delta_p}{2\alpha + \lambda_1^2} \). This provides theoretical support for us to flexibly choose step-size schemes. Equation (11) shows that larger step-sizes in a safe range will lead to faster convergence. Moreover, we can see from Equation (12) that the convergence actually depends on the relative generalized eigengap \( \frac{\Delta_p}{\lambda_1^2} \). This seems to be explicitly shown for the first time for the considered solver. As mentioned in Section 1, the proof provides two general analysis frameworks. The generality lies in that both rates hold for any value of the positive \( \Delta_p \). The combination of the results from two kinds of analysis comprehensively portrays the convergence behaviors of the solver. For both kinds of analysis, the convergence has only logarithmic dependence on the initial iterate via \( \psi(x_0) = -2\log \|V_p^T x_0\|_2 \).

Proof of Lemma 2

Proof. Plugging in (5) to \( x^T Ax \), we get
\[
x^T Ax = \lambda_1 \|V_p^T x\|_2^2 + x^T V_p^T \text{diag} \lambda_{p+1}, \ldots, \lambda_n \rangle V_p^T \rangle x 
\leq \lambda_1 \|V_p^T x\|_2^2 + \lambda_{p+1} \|V_p^T \rangle x \|
\leq \lambda_1 \|V_p^T x\|_2^2 + \lambda_{p+1} \langle x - V_p^T \rangle x \|
\leq \lambda_1 \cos \theta(x, V_p) + \lambda_{p+1} \sin ^2 \theta(x, V_p).
\]

Thus, one arrives at
\[
\lambda_1 - x^T Ax \geq \lambda_1 - \lambda_1 \cos \theta(x, V_p) - \lambda_{p+1} \sin ^2 \theta(x, V_p) = (\lambda_1 - \lambda_{p+1}) \sin ^2 \theta(x, V_p).
\]
\( \square \)
Proof of Lemma 3

Proof.

\[ \|x_t - x_{t+1}\|^2 = (x_t - x_{t+1})^T (x_t - x_{t+1}) \]

where, plugging in the following form of A’s full eigen-decomposition,

\[ A = \lambda_1 v v^T + \cdots + \lambda_n v^T, \]

for any unit vector \( v \in \text{span}(V_p) \), we get

\[ \|\nabla f(x_t)\|^2 = (x_t - x_{t+1})^T (x_t - x_{t+1}) \]

where, \( \lambda_i \) and \( v_i \) are the eigenvalues and eigenvectors of matrix \( A \), respectively.

By the definition of principal angles [Golub and Van Loan, 1996],

\[ 1 - \|V_p^T x\|^2 = \min_{v \in \text{span}(V_p)} (1 - (v^T x)^2). \]

One thus has

\[ \|\nabla f(x_t)\|^2 \leq (\lambda_1^2 + \lambda_2^2) \sin^2 \theta(x, V_p). \]

Proof of Lemma 5

Proof. Let \( \sigma_{\min}(\cdot) \) and \( \sigma_{\max}(\cdot) \) be the smallest and largest singular value of a matrix, respectively. We then have

\[ \|V_p^T x_0\|_2 = \frac{\|V_p^T y\|_2}{\|y\|_2} = \frac{\sigma_{\min}(V_p^T y)}{\sigma_{\max}(y)}, \]

where \( \sigma_{\max}(y) > 0 \) almost surely. Since entries of \( y \) are i.i.d. samples from \( \mathcal{N}(0, 1) \) and \( V_p \) is orthonormal, entries of \( V_p^T y \) are i.i.d. standard normal samples as well. By Equation (3.2) in [Rudelson and Vershynin, 2010], \( \sigma_{\min}(V_p^T y) > 0 \) almost surely (i.e., with probability 1). By Theorem 3.3 in [Rudelson and Vershynin, 2010], we have \( \sigma_{\min}(V_p^T y) > 0 \) with probability at least \( 1 - \nu^p \), where \( 0 < \nu < 1 \) is a constant about \( V_p^T y \).

4 Experiments

Theoretical properties of the considered solver were already observed and verified empirically in, e.g., [Wen and Yin, 2013; Liu et al., 2016], albeit without thorough theoretical support. For the sake of completeness, we reproduce a few experiments on both synthetic and real data here, with new testing on matrices of zero eigengap. All the ground truth information, e.g., \( v_1 \) or \( V_p \), is obtained by MATLAB’s eigs function for the purpose of benchmarking.

The advantage of using synthetic data mainly lies in the control of the eigengap. We set \( n = 1000 \) and follow [Shamir, 2015] to generate data using the full eigenvalue decomposition \( A = U \Sigma U^T \). \( U \) is an orthogonal matrix and is set the same way as \( x_0 \). \( \Sigma = \text{diag}(\Sigma_1, \Sigma_2) \), where \( \Sigma_1 = \text{diag}(\lambda_1, \cdots, \lambda_{n-1}) \) with \( g_1 \sim \mathcal{N}(0, 1) \) and \( r \) being the order of \( \Sigma_1 \). In addition, the following two settings are considered:

- \( p = 1: \Sigma_1 = \text{diag}(1, 1 - \eta, 1 - 1.1 \eta, 1 - 1.2 \eta, 1 - 1.3 \eta, 1 - 1.4 \eta) \), and then \( \frac{\Delta_1}{\lambda_1} = \eta > 0 \), where \( \eta \in \{0.2, 0.5, 0.8\} \).
- \( p = 3: \Sigma_1 = \text{diag}(1, 1, 1) \), and then \( \frac{\Delta_1}{\lambda_1} = 1 - \frac{\log n}{n} > 0 \).

For the conventionally considered case that \( \Delta_1 > 0 \), two types of convergence curves, in terms of the relative function error \( \frac{f(x_0) - f(v_1)}{f(v_1)} \) and the commonly used potential function \( \sin^2 \theta(x_i, v_1) \), respectively, are shown in Figure 2 for different eigengap values. Global linear convergence of the solver is observed, and a larger eigengap yields faster convergence. It agrees with the theory. The convergence trends remain greatly consistent for the two types, which holds in the rest of experiments as well.

For the case that \( \Delta_1 = 0 \) and \( \alpha_{t+1} = \frac{\gamma}{1 + T} \) (see Remark), different pairs of step-size parameters are tested in Figure 1. As described by Theorem 1, the convergence reported in Figures 1(a)-(b) is global linear, rather than sub-linear by [Xu and Gao, 2018], and large step-sizes result in faster convergence as well. As it is unknown which leading eigenvector \( x_i \) will converge to finally, it is necessary for us to change the potential from the commonly used \( \sin^2 \theta(x_i, v_1) \) to \( \sin^2 \theta(x_i, V_p) \) in this case. As demonstrated in Figure 1(c),
$\sin^2 \theta(x_t, v_1)$ is unable to converge, which implies that $x_t$ has converged to another leading eigenvector in $\text{span}(V_p)$.

We run two implementations of the solver on a real symmetric matrix $A$, named Schenk, which is of size $10,728 \times 10,728$ with 85,000 nonzero entries. The implementation is characterized by the combination of the retraction and step-size scheme used. Curvilinear search was proposed in [Wen and Yin, 2013] which uses “Cayley transform + Barzilai-Borwein step-size”. The other implementation uses “normalization + constant step-size”. They were fed with the same random initial iterate. The results are reported in Figure 3, which reflect the global linear convergence as well and meanwhile shows that different step-size schemes matter in practice.

![Figure 1: Synthetic data: $\Delta_1 = 0$](image1)

![Figure 2: Synthetic data: $\Delta_1 > 0$](image2)

![Figure 3: Real data.](image3)

### 5 Conclusion

We presented a simple yet comprehensive convergence analysis of the Riemannian gradient descent solver for the leading eigenvector computation. Two kinds of general analysis jointly established the true global rate of convergence to one of the leading eigenvectors. The generalized eigengap $\Delta_p$ eliminates the limitation of the commonly considered eigengap $\Delta_1$. When $\Delta_p$ is large, the convergence is linear, which is described by the first general analysis. If it is as small as $\epsilon$, the second general analysis shows that it is sub-linear $O(1/t)$ rather than $O(1/t^2)$. It was also shown that the convergence only logarithmically depends on the initial iterate. The key to these breakthroughs made for the considered solver is to use for analysis the log principal angle between iterates and the space of the leading eigenvectors. Along this work, there are a few interesting future research directions. For example, it is unknown if the results can be extended to $k > 1$ for the top-
$k$ eigenspace computation without deflation. In particular, it would be more attractive to practitioners if the analysis can be translated to the case of the stochastic gradient descent solver, which is well worth further investigation. Also, although the rate is optimal for the considered solver, it would be of great interest to develop faster solvers of this type to match the optimal rate for the problem, e.g., those of Lanczos algorithms.

Acknowledgements

This research is supported in part by the funding from King Abdullah University of Science and Technology (KAUST).

References


