Convergence Analysis of Gradient Descent for Eigenvector Computation

Zhiqiang Xu[*1], **Xin Cao**[2] and **Xin Gao**[1]
[1] KAUST, Saudi Arabia
[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 $O(\min\{(\frac{\lambda_1}{\Delta_p})^2 \log \frac{1}{\epsilon}, \frac{1}{\epsilon}\})$, where $\Delta_p = \lambda_p - \lambda_{p+1} > 0$ represents the generalized positive eigengap and always exists without loss of generality with λ_i being the i-th largest eigenvalue of the given real symmetric matrix and p being the multiplicity of λ_1. The rate is linear at $O((\frac{\lambda_1}{\Delta_p})^2 \log \frac{1}{\epsilon})$ if $(\frac{\lambda_1}{\Delta_p})^2 = O(1)$, otherwise sub-linear at $O(\frac{1}{\epsilon})$. 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 Δ_p satisfies $(\frac{\lambda_1}{\Delta_p})^2 = O(1)$, 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:

$$\min_{\mathbf{x}\in\mathbb{R}^n:\|\mathbf{x}\|_2=1} f(\mathbf{x}) = -\mathbf{x}^\top \mathbf{A}\mathbf{x}, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{A}^\top = \mathbf{A}$. 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 $O(\frac{1}{\epsilon^2})$ was given by [Arora *et al.*, 2013; Shamir, 2016a]. Very recently, the rate for the case of an arbitrary eigengap has been improved to $O(\frac{1}{\epsilon})$ [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:

$$\mathbf{x}_{t+1} = R(\mathbf{x}_t, -\alpha_{t+1}\tilde{\nabla} f(\mathbf{x}_t)), \tag{2}$$

where $\tilde{\nabla} f(\mathbf{x}_t)$ denotes the Riemannian gradient representing the steepest ascent direction in the equidimensional Euclidean

[*]Corresponding author <zhiqiangxu2001@gmail.com>

space tangent to the manifold at $\mathbf{x}_t$, $\alpha_{t+1} > 0$ represents the step-size, and $R(\mathbf{x}_t, \cdot)$ represents the retraction map from the tangent space at $\mathbf{x}_t$ to the manifold. In particular, in order to measure the progress of the iterate $\mathbf{x}_t$ to one of globally optimal solutions, we use for analysis a novel potential function defined by the log principal angle between $\mathbf{x}_t$ 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 λ_i represents the i-th largest eigenvalue of the given matrix $\mathbf{A}$ and p is the multiplicity of λ_1. Δ_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 Δ_p is as small as the target precision parameter ϵ, the resulting theoretic complexity is high. However, another unified analysis regardless of the value of Δ_p we present subsequently shows that this can be significantly reduced. Specifically, we make the following contributions:

- A global convergence rate $O(\min\{(\frac{\lambda_1}{\Delta_p})^2 \log \frac{1}{\epsilon}, \frac{1}{\epsilon}\})$ of the Riemannian gradient descent solver is achieved for Problem (1). The rate is linear at $O((\frac{\lambda_1}{\Delta_p})^2 \log \frac{1}{\epsilon})$ if $(\frac{\lambda_1}{\Delta_p})^2 = O(1)$, otherwise sub-linear at $O(\frac{1}{\epsilon})$. 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 Δ_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 $O(\frac{1}{\Delta\epsilon})$ or at a global, gap-free, and sub-linear rate $O(\frac{1}{\epsilon^2})$ [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 $O(\frac{1}{\Delta_1^2} \log \frac{1}{\epsilon})$ of the Riemannian SVRG solver [Xu and Ke, 2016b; Xu *et al.*, 2017]. By proving an explicit Łojasiewicz exponent at $\frac{1}{2}$ 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, alecton 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 $O(\frac{1}{\Delta_1^2 \epsilon})$ via the martingale analysis [Sa *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 $\mathbf{A} \in \mathbb{R}^{n \times n}$, its i-th largest eigenvalue is denoted as λ_i, i.e., $\lambda_1 \geq \cdots \geq \lambda_n$, and

the corresponding eigenvector denoted as $\mathbf{v}_i$. Define eigengap $\Delta_i = \lambda_i - \lambda_{i+1}$ and let p be the multiplicity of λ_1 with the associated eigenspace denoted as $\mathbf{V}_p = [\mathbf{v}_1, \cdots, \mathbf{v}_p]$. One then should have $\Delta_j = 0$ for $j = 1, \cdots, p-1$ and $\Delta_p > 0$ without loss of generality, i.e., $p < n$, because $f(\mathbf{x})$ 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 Δ_1. To this end, we have to rely on Δ_p and $\mathbf{V}_p$, where we follow [Xu and Gao, 2018] to term Δ_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 $\mathbf{v} \in \mathrm{span}(\mathbf{V}_p)$, instead of a specific solution, e.g., $\mathbf{v}_1$. We thus define the following potential function:

$$\psi(\mathbf{x}_t) = -2\log\|\mathbf{V}_p^\top \mathbf{x}_t\|_2, \tag{3}$$

where actually $\|\mathbf{V}_p^\top \mathbf{x}_t\|_2 = \cos\theta(\mathbf{x}_t, \mathbf{V}_p)$ representing the cosine of the principal angle between the current iterate $\mathbf{x}_t$ and the space of globally optimal solutions. As $\|\mathbf{V}_p^\top \mathbf{x}_t\|_2 \leq 1$, it is easy to see that $\psi(\mathbf{x}_t) \geq 0$ and $\mathbf{x}_t$ converges to a unit vector in $\mathrm{span}(\mathbf{V}_p)$ when $\psi(\mathbf{x}_t)$ goes to 0.

In addition, we take the normalization retraction, i.e., $R(\mathbf{x}, \xi) = \frac{\mathbf{x}+\xi}{\|\mathbf{x}+\xi\|_2}$. Thus, the update in (2) can be explicitly written as

$$\mathbf{x}_{t+1} = \frac{\mathbf{x}_t - \alpha_{t+1}\tilde{\nabla} f(\mathbf{x}_t)}{\|\mathbf{x}_t - \alpha_{t+1}\tilde{\nabla} f(\mathbf{x}_t)\|_2}, \tag{4}$$

where the Riemannian gradient for Problem (1) is $\tilde{\nabla} f(\mathbf{x}) = -(\mathbf{I} - \mathbf{x}\mathbf{x}^\top)\mathbf{A}\mathbf{x}$.

3.2 Main Results

Theorem 1. *If the initial $\mathbf{x}_0 = \frac{\mathbf{y}}{\|\mathbf{y}\|_2}$ where entries of $\mathbf{y}$ are i.i.d. standard normal samples, i.e., $\mathbf{y}_i \sim \mathcal{N}(0,1)$, then the Rimannian gradient descent solver for Problem (1) will converge to certain unit vector $\mathbf{v} \in \mathrm{span}(\mathbf{V}_p)$ at a global rate $O(\min\{(\frac{\lambda_1}{\Delta_p})^2 \log\frac{1}{\epsilon}, \frac{1}{\epsilon}\})$ with high probability. Specifically, for any $\epsilon \in (0,1)$,*

- *I) the solver with constant step-sizes $\alpha < \frac{\Delta_p}{2\lambda_1^2(1+\alpha\Delta_p)}$ will converge, i.e., $\psi(\mathbf{x}_T) < \epsilon$, after $T = O((\frac{\lambda_1}{\Delta_p})^2 \log\frac{\psi(\mathbf{x}_0)}{\epsilon})$ iterations, with probability at least $1 - \nu^p$, where $\nu = 0$ if $p = 1$ otherwise $\nu \in (0,1)$;*
- *II) the solver with diminishing step-sizes $\alpha_t = \frac{c}{\tau+t}$ for sufficiently large constants $c, \tau > 0$ will converge, i.e., $\psi(\mathbf{x}_T) < \epsilon$, after $T = O(\frac{1}{\epsilon})$ 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 [Shamir, 2016a] is $\Omega(\frac{1}{n})$. To prove the theorem, we need a few important lemmas whose proofs are deferred to after that of Theorem 1.

Lemma 2. $\lambda_1 - \mathbf{x}^\top \mathbf{A}\mathbf{x} \geq \Delta_p \sin^2\theta(\mathbf{x}, \mathbf{V}_p)$.

Lemma 3. $\|\mathbf{x}_t - \alpha_{t+1}\tilde{\nabla} f(\mathbf{x}_t)\|_2^2 \leq 1 + \alpha_{t+1}^2(\lambda_1^2 + \lambda_2^2)\sin^2\theta(\mathbf{x}_t, \mathbf{V}_p)$.

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

Lemma 5. $\|\mathbf{V}_p^\top \mathbf{x}_0\|_2 > 0$ *with probability* 1 *if* $p = 1$, *otherwise with probability at least* $1 - \nu^p$, *where* $0 < \nu < 1$ *is a constant about* $\mathbf{V}_p^\top \mathbf{y}$.

Proof of Theorem 1.

Proof. I) We start from expanding $\psi(\mathbf{x}_{t+1})$ using (3)-(4):

$$\begin{aligned}\psi(\mathbf{x}_{t+1}) &= -2\log\|\mathbf{V}_p^\top \mathbf{x}_{t+1}\|_2 \\ &= -2\log\|\mathbf{V}_p^\top(\mathbf{x}_t - \alpha_{t+1}\tilde{\nabla} f(\mathbf{x}_t))\|_2 \\ &\quad +2\log\|\mathbf{x}_t - \alpha_{t+1}\tilde{\nabla} f(\mathbf{x}_t)\|_2.\end{aligned}$$

To proceed, we need the full eigen-decomposition of $\mathbf{A}$, i.e.,

$$\mathbf{A} = \lambda_1 \mathbf{V}_p \mathbf{V}_p^\top + \mathbf{V}_p^\perp \mathrm{diag}(\lambda_{p+1}, \cdots, \lambda_n)(\mathbf{V}_p^\perp)^\top, \tag{5}$$

where $\mathbf{V}_p^\perp$ represents the orthogonal complement of $\mathbf{V}_p$. Plugging in this decomposition to $\mathbf{V}_p^\top \mathbf{A}$, one then gets

$$\begin{aligned}-\mathbf{V}_p^\top \tilde{\nabla} f(\mathbf{x}_t) &= \mathbf{V}_p^\top(\mathbf{I} - \mathbf{x}_t\mathbf{x}_t^\top)\mathbf{A}\mathbf{x}_t \\ &= \mathbf{V}_p^\top \mathbf{A}\mathbf{x}_t - (\mathbf{x}_t^\top \mathbf{A}\mathbf{x}_t)\mathbf{V}_p^\top \mathbf{x}_t \\ &= (\lambda_1 - \mathbf{x}_t^\top \mathbf{A}\mathbf{x}_t)\mathbf{V}_p^\top \mathbf{x}_t.\end{aligned}$$

We now can write

$$\begin{aligned}\psi(\mathbf{x}_{t+1}) &= \psi(\mathbf{x}_t) - 2\log(1 + \alpha_{t+1}(\lambda_1 - \mathbf{x}_t^\top \mathbf{A}\mathbf{x}_t)) \\ &\quad +2\log\|\mathbf{x}_t - \alpha_{t+1}\tilde{\nabla} f(\mathbf{x}_t)\|_2,\end{aligned}$$

where we have

$$\begin{aligned}&-2\log(1 + \alpha_{t+1}(\lambda_1 - \mathbf{x}_t^\top \mathbf{A}\mathbf{x}_t)) \\ \leq\ &-2\log(1 + \alpha_{t+1}\Delta_p \sin^2\theta(\mathbf{x}_t, \mathbf{V}_p)) \quad \text{(by Lemma 2)} \\ \leq\ &-\frac{2\alpha_{t+1}\Delta_p \sin^2\theta(\mathbf{x}_t, \mathbf{V}_p)}{1 + \alpha_{t+1}\Delta_p \sin^2\theta(\mathbf{x}_t, \mathbf{V}_p)} \quad \text{(by Lemma 4)} \\ \leq\ &-\frac{2\alpha_{t+1}\Delta_p}{1 + \alpha_{t+1}\Delta_p}\sin^2\theta(\mathbf{x}_t, \mathbf{V}_p),\end{aligned}$$

and

$$\begin{aligned}&2\log\|\mathbf{x}_t - \alpha_{t+1}\tilde{\nabla} f(\mathbf{x}_t)\|_2 \\ \leq\ &\log(1 + \alpha_{t+1}^2(\lambda_1^2 + \lambda_2^2)\sin^2\theta(\mathbf{x}_t, \mathbf{V}_p)) \text{(by Lemma 3)} \\ \leq\ &\alpha_{t+1}^2(\lambda_1^2 + \lambda_2^2)\sin^2\theta(\mathbf{x}_t, \mathbf{V}_p) \quad \text{(by Lemma 4)} \\ \leq\ &2\alpha_{t+1}^2\lambda_1^2 \sin^2\theta(\mathbf{x}_t, \mathbf{V}_p).\end{aligned}$$

One thus can arrive at

$$\begin{aligned}\psi(\mathbf{x}_{t+1}) &\leq \psi(\mathbf{x}_t) - \frac{2\alpha_{t+1}\Delta_p}{1 + \alpha_{t+1}\Delta_p}\sin^2\theta(\mathbf{x}_t, \mathbf{V}_p) \\ &\quad +2\alpha_{t+1}^2\lambda_1^2 \sin^2\theta(\mathbf{x}_t, \mathbf{V}_p) \\ &= \psi(\mathbf{x}_t) - (\frac{2\Delta_p}{1 + \alpha_{t+1}\Delta_p} - 2\alpha_{t+1}\lambda_1^2) \\ &\quad \cdot\alpha_{t+1}\frac{\sin^2\theta(\mathbf{x}_t, \mathbf{V}_p)}{\psi(\mathbf{x}_t)} \cdot \psi(\mathbf{x}_t).\end{aligned} \tag{6}$$

Note that by Lemma 4,

$$\begin{aligned}\frac{\sin^2\theta(\mathbf{x}_t,\mathbf{V}_p)}{\psi(\mathbf{x}_t)} &= \frac{\sin^2\theta(\mathbf{x}_t,\mathbf{V}_p)}{-\log(1-\sin^2\theta(\mathbf{x}_t,\mathbf{V}_p))}\\ &\geq \frac{1}{1-\log(1-\sin^2\theta(\mathbf{x}_t,\mathbf{V}_p))}\\ &= \frac{1}{1+\psi(\mathbf{x}_t)}. \qquad (7)\end{aligned}$$

If $\frac{\Delta_p}{1+\alpha_{t+1}\Delta_p} - 2\alpha_{t+1}\lambda_1^2 > 0$, i.e., $\alpha_{t+1} < \frac{\Delta_p}{2\lambda_1^2(1+\alpha_{t+1}\Delta_p)}$, then $\psi(\mathbf{x}_{t+1}) < \psi(\mathbf{x}_t)$ and

$$\begin{aligned}\psi(\mathbf{x}_{t+1}) &\leq \psi(\mathbf{x}_t) - \frac{\Delta_p}{1+\alpha_{t+1}\Delta_p}\frac{\alpha_{t+1}}{1+\psi(\mathbf{x}_t)}\psi(\mathbf{x}_t)\\ &= (1-\frac{\Delta_p}{1+\alpha_{t+1}\Delta_p}\frac{\alpha_{t+1}}{1+\psi(\mathbf{x}_t)})\psi(\mathbf{x}_t) \qquad (8)\\ &\leq (1-\frac{\Delta_p}{1+\alpha\Delta_p}\frac{\alpha}{1+\psi(\mathbf{x}_0)})\psi(\mathbf{x}_t), \qquad (9)\end{aligned}$$

if $\alpha_{t+1} \equiv \alpha$. We thus have

$$\begin{aligned}\psi(\mathbf{x}_t) &\leq (1-\frac{\Delta_p}{1+\alpha\Delta_p}\frac{\alpha}{1+\psi(\mathbf{x}_0)})^t\psi(\mathbf{x}_0)\\ &\leq \exp\{-t\frac{\Delta_p}{1+\alpha\Delta_p}\frac{\alpha}{1+\psi(\mathbf{x}_0)}\}\psi(\mathbf{x}_0) \triangleq \Xi.\end{aligned}$$

If $\Xi < \epsilon$, i.e.,

$$\psi(\mathbf{x}_0) < +\infty \qquad (10)$$

and

$$T > \frac{(1+\alpha\Delta_p)(1+\psi(\mathbf{x}_0))}{\alpha\Delta_p}\log\frac{\psi(\mathbf{x}_0)}{\epsilon},$$

we must have $\psi(\mathbf{x}_T) < \epsilon$. Plugging in $\alpha < \frac{\Delta_p}{2\lambda_1^2(1+\alpha\Delta_p)}$ to T, we can write

$$\begin{aligned}T &= O(\frac{1}{\alpha\Delta_p}\log\frac{\psi(\mathbf{x}_0)}{\epsilon}) \qquad (11)\\ &= O((\frac{\lambda_1}{\Delta_p})^2\log\frac{\psi(\mathbf{x}_0)}{\epsilon}). \qquad (12)\end{aligned}$$

Finally note that (10) is equivalent to $\|\mathbf{V}_p^\top\mathbf{x}_0\|_2 > 0$, 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

$$\begin{aligned}\psi(\mathbf{x}_{t+1}) &\leq \psi(\mathbf{x}_t) - \frac{2\alpha_{t+1}\Delta_p}{1+\alpha_{t+1}\Delta_p}\sin^2\theta(\mathbf{x}_t,\mathbf{V}_p)\\ &\quad +2\alpha_{t+1}^2\lambda_1^2\sin^2\theta(\mathbf{x}_t,\mathbf{V}_p)\\ &\leq \psi(\mathbf{x}_t) - \frac{2\alpha_{t+1}\Delta_p}{1+\alpha_{t+1}\Delta_p}\frac{\psi(\mathbf{x}_t)}{1+\psi(\mathbf{x}_t)} + 2\lambda_1^2\alpha_{t+1}^2.\end{aligned}$$

Let $\alpha_{t+1} = \frac{c}{\tau+t}$ where $c, \tau > 0$ are constants. It is easy to see that for sufficiently small α_{t+1}, equivalently, sufficiently large τ, we have $\psi(\mathbf{x}_{t+1}) < \psi(\mathbf{x}_t)$. Thus, one gets

$$\psi(\mathbf{x}_{t+1}) \leq (1-\frac{2\alpha_{t+1}\Delta_p}{1+\frac{c}{\tau}\Delta_p}\frac{1}{1+\psi(\mathbf{x}_0)})\psi(\mathbf{x}_t) + 2\lambda_1^2\alpha_{t+1}^2.$$

Denoting $a = \frac{2c\Delta_p}{1+\frac{c}{\tau}\Delta_p}\frac{1}{1+\psi(\mathbf{x}_0)}$ and $b = 2c^2\lambda_1^2$, one can write

$$\psi(\mathbf{x}_{t+1}) \leq (1-\frac{a}{\tau+t})\psi(\mathbf{x}_t) + \frac{b}{(\tau+t)^2}.$$

As long as $\psi(\mathbf{x}_0) < +\infty$ and $a > 1$, by Lemma D.1 in [Balsubramani *et al.*, 2013], the recursion of the above inequality will yield

$$\psi(\mathbf{x}_t) \leq (\frac{\tau+1}{t+\tau+1})^a\psi(\mathbf{x}_0) + \frac{2^{a+1}b}{a-1}\frac{1}{t+\tau+1}.$$

We thus have $\psi(\mathbf{x}_T) = O(\frac{1}{T})$, i.e., $T = O(\frac{1}{\epsilon})$. The proof completes by noting that $\psi(\mathbf{x}_0) < +\infty$ occurs with high probability similarly and $a > 1$ can be guaranteed by choosing sufficiently large constants c, τ. □

Remark We make a few remarks on our proof. Equation (8) shows that the convergence is at least linear because $\psi(\mathbf{x}_{t+1}) < \psi(\mathbf{x}_t)$ and thus the contraction factor keep decreasing for constant step-sizes. In Equation (9), α_{t+1} is not necessarily constant. In fact, the results will hold for any step-size scheme as long as $\alpha_{t+1} < \frac{\Delta_p}{2\lambda_1^2(1+\alpha_{t+1}\Delta_p)}$. 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 $\frac{\Delta_p}{\lambda_1}$. 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 Δ_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 $\psi(\mathbf{x}_0) = -2\log\|\mathbf{V}_p^\top\mathbf{x}_0\|_2$.

Proof of Lemma 2

Proof. Plugging in (5) to $\mathbf{x}^\top\mathbf{A}\mathbf{x}$, we get

$$\begin{aligned}&\mathbf{x}^\top\mathbf{A}\mathbf{x}\\ =\ &\lambda_1\|\mathbf{V}_p^\top\mathbf{x}\|_2^2 + \mathbf{x}^\top\mathbf{V}_p^\perp\mathrm{diag}(\lambda_{p+1},\cdots,\lambda_n)(\mathbf{V}_p^\perp)^\top\mathbf{x}\\ \leq\ &\lambda_1\|\mathbf{V}_p^\top\mathbf{x}\|_2^2 + \lambda_{p+1}\mathbf{x}^\top\mathbf{V}_p^\perp(\mathbf{V}_p^\perp)^\top\mathbf{x}\\ =\ &\lambda_1\|\mathbf{V}_p^\top\mathbf{x}\|_2^2 + \lambda_{p+1}\mathbf{x}^\top(\mathbf{I}-\mathbf{V}_p\mathbf{V}_p^\top)\mathbf{x}\\ =\ &\lambda_1\|\mathbf{V}_p^\top\mathbf{x}\|_2^2 + \lambda_{p+1}(1-\|\mathbf{V}_p^\top\mathbf{x}\|_2^2)\\ =\ &\lambda_1\cos^2\theta(\mathbf{x},\mathbf{V}_p) + \lambda_{p+1}\sin^2\theta(\mathbf{x},\mathbf{V}_p).\end{aligned}$$

Thus, one arrives at

$$\begin{aligned}&\lambda_1 - \mathbf{x}^\top\mathbf{A}\mathbf{x}\\ \geq\ &\lambda_1 - \lambda_1\cos^2\theta(\mathbf{x},\mathbf{V}_p) - \lambda_{p+1}\sin^2\theta(\mathbf{x},\mathbf{V}_p)\\ =\ &(\lambda_1-\lambda_{p+1})\sin^2\theta(\mathbf{x},\mathbf{V}_p).\end{aligned}$$

□

Proof of Lemma 3

Proof.

$$\begin{aligned} & \|\mathbf{x}_t - \alpha_{t+1}\tilde{\nabla} f(\mathbf{x}_t)\|_2^2 \\ = \; & (\mathbf{x}_t - \alpha_{t+1}\tilde{\nabla} f(\mathbf{x}_t))^\top (\mathbf{x}_t - \alpha_{t+1}\tilde{\nabla} f(\mathbf{x}_t)) \\ = \; & 1 + \alpha_{t+1}^2 \|\tilde{\nabla} f(\mathbf{x}_t)\|_2^2, \end{aligned}$$

where, plugging in the following form of $\mathbf{A}$'s full eigen-decomposition,

$$\mathbf{A} = \lambda_1 \mathbf{v}\mathbf{v}^\top + \mathbf{v}_\perp \mathrm{diag}(\lambda_2, \cdots, \lambda_n)\mathbf{v}_\perp^\top,$$

for any unit vector $\mathbf{v} \in \mathrm{span}(\mathbf{V}_p)$, we get

$$\begin{aligned} & \|\tilde{\nabla} f(\mathbf{x}_t)\|_2^2 = \|\mathbf{x}_\perp^\top \mathbf{A}\mathbf{x}\|_2^2 \\ = \; & \|\lambda_1 \mathbf{x}_\perp^\top \mathbf{v}\mathbf{v}^\top \mathbf{x} + \mathbf{x}_\perp^\top \mathbf{v}_\perp \mathrm{diag}(\lambda_2, \cdots, \lambda_n)\mathbf{v}_\perp^\top \mathbf{x}\|_2^2 \\ \leq \; & 2\|\lambda_1 \mathbf{x}_\perp^\top \mathbf{v}\mathbf{v}^\top \mathbf{x}\|_2^2 + 2\|\mathbf{x}_\perp^\top \mathbf{v}_\perp \mathrm{diag}(\lambda_2, \cdots, \lambda_n)\mathbf{v}_\perp^\top \mathbf{x}\|_2^2 \\ \leq \; & 2\lambda_1^2 \|\mathbf{x}_\perp^\top \mathbf{v}\|_2^2 \|\mathbf{v}^\top \mathbf{x}\|_2^2 + 2\lambda_2^2 \|\mathbf{x}_\perp^\top \mathbf{v}_\perp\|_2^2 \|\mathbf{v}_\perp^\top \mathbf{x}\|_2^2 \\ \leq \; & 2\lambda_1^2 \|\mathbf{x}_\perp^\top \mathbf{v}\|_2^2 + 2\lambda_2^2 \|\mathbf{v}_\perp^\top \mathbf{x}\|_2^2 \\ = \; & 2\lambda_1^2 \mathbf{v}^\top \mathbf{x}_\perp \mathbf{x}_\perp^\top \mathbf{v} + 2\lambda_2^2 \mathbf{x}^\top \mathbf{v}_\perp \mathbf{v}_\perp^\top \mathbf{x} \\ = \; & 2\lambda_1^2 \mathbf{v}^\top (\mathbf{I} - \mathbf{x}\mathbf{x}^\top)\mathbf{v} + 2\lambda_2^2 \mathbf{x}^\top (\mathbf{I} - \mathbf{v}\mathbf{v}^\top)\mathbf{x} \\ = \; & 2(\lambda_1^2 + \lambda_2^2)(1 - (\mathbf{v}^\top \mathbf{x})^2). \end{aligned}$$

Since the above inequality holds for any unit vector $\mathbf{v} \in \mathrm{span}(\mathbf{V}_p)$, we have

$$\|\tilde{\nabla} f(\mathbf{x}_t)\|_2^2 \leq 2(\lambda_1^2 + \lambda_2^2) \min_{\|\mathbf{v}\|_2 = 1} \min_{\mathbf{v} \in \mathrm{span}(\mathbf{V}_p)} (1 - (\mathbf{v}^\top \mathbf{x})^2).$$

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

$$1 - \|\mathbf{V}_p^\top \mathbf{x}\|^2 = \min_{\|\mathbf{v}\|_2 = 1} \min_{\mathbf{v} \in \mathrm{span}(\mathbf{V}_p)} (1 - (\mathbf{v}^\top \mathbf{x})^2).$$

One thus has

$$\|\tilde{\nabla} f(\mathbf{x}_t)\|_2^2 \leq 2(\lambda_1^2 + \lambda_2^2) \sin^2 \theta(\mathbf{x}, \mathbf{V}_p).$$

□

Proof of Lemma 4

Proof. 1) For any x, it holds that $1 + x \leq e^x$. Then for any $x > -1$,

$$\log(1 + x) \leq x.$$

If one lets $y = 1 + x$ in the above inequality, then $\log y \leq y - 1$. Further letting $y = \frac{1}{z}$ yields $\log z \geq -\frac{1}{z} + 1 = \frac{z-1}{z}$. Last, setting $z = 1 + x$ gives us

$$\log(1 + x) \geq \frac{x}{1 + x}.$$

2) Note that $\log(1 + x) = \sum_{i=0}^{\infty} (-1)^i \frac{x^{i+1}}{i+1}$ for $|x| < 1$. One then can write for $x \in (0, 1)$ that

$$\begin{aligned} \frac{x}{-\log(1 - x)} &= \frac{x}{-\sum_{i=0}^{\infty} (-1)^i \frac{(-x)^{i+1}}{i+1}} \\ &= \frac{1}{\sum_{i=0}^{\infty} \frac{x^i}{i+1}} = \frac{1}{1 + \sum_{i=1}^{\infty} \frac{x^i}{i+1}} \\ &\geq \frac{1}{1 + \sum_{i=1}^{\infty} \frac{x^i}{i}} = \frac{1}{1 - \sum_{i=1}^{\infty} (-1)^{i-1} \frac{(-x)^i}{i}} \\ &= 1/(1 - \log(1 - x)). \end{aligned}$$

□

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

$$\|\mathbf{V}_p^\top \mathbf{x}_0\|_2 = \frac{\|\mathbf{V}_p^\top \mathbf{y}\|_2}{\|\mathbf{y}\|_2} = \frac{\sigma_{\min}(\mathbf{V}_p^\top \mathbf{y})}{\sigma_{\max}(\mathbf{y})},$$

where $\sigma_{\max}(\mathbf{y}) > 0$ almost surely. Since entries of $\mathbf{y}$ are i.i.d. samples from $\mathcal{N}(0, 1)$ and $\mathbf{V}_p$ is orthonormal, entries of $\mathbf{V}_p^\top \mathbf{y}$ are i.i.d. standard normal samples as well. By Equation (3.2) in [Rudelson and Vershynin, 2010], $\sigma_{\min}(\mathbf{V}_1^\top \mathbf{y}) > 0$ almost surely (i.e., with probability 1). By Theorem 3.3 in [Rudelson and Vershynin, 2010], we have $\sigma_{\min}(\mathbf{V}_p^\top \mathbf{y}) > 0$ with probability at least $1 - \nu^p$, where $0 < \nu < 1$ is a constant about $\mathbf{V}_p^\top \mathbf{y}$. □

4 Experiments

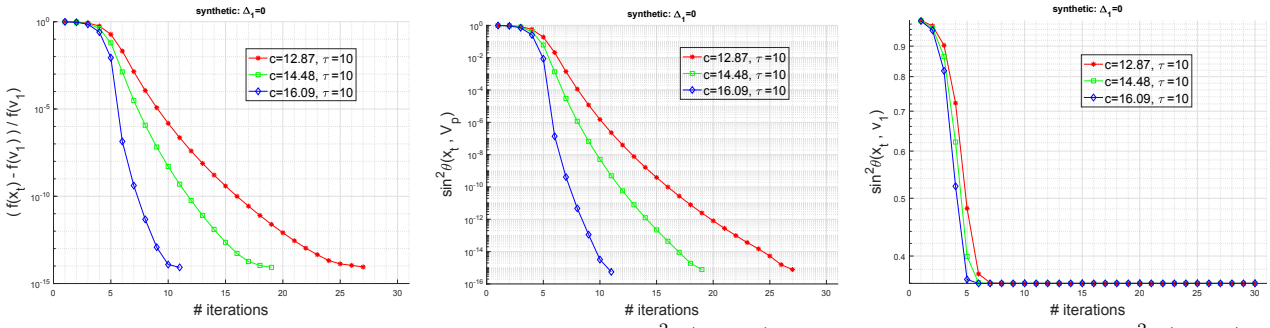
Theoretical properties of the considered solver were already observed and verified empirically in, e.g., [Wen and Yin, 2013; Liu *et al.*, 2016], albeit without thorough theoretical support. For the sake of completeness, we reproduce a few experiments on both synthetic and real data here, with new testing on matrices of zero eigengap. All the ground truth information, e.g., $\mathbf{v}_1$ or $\mathbf{V}_p$, 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 $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top$. $\mathbf{U}$ is an orthogonal matrix and is set the same way as $\mathbf{x}_0$. $\mathbf{\Sigma} = \mathrm{diag}(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2)$, where $\mathbf{\Sigma}_2 = \mathrm{diag}(\frac{|g_1|}{n}, \cdots, \frac{|g_{n-r}|}{n})$ with $g_i \sim \mathcal{N}(0, 1)$ and r being the order of $\mathbf{\Sigma}_1$. In addition, the following two settings are considered:

- $p = 1$: $\mathbf{\Sigma}_1 = \mathrm{diag}(1, 1 - \eta, 1 - 1.1\eta, 1 - 1.2\eta, 1 - 1.3\eta, 1 - 1.4\eta)$, and then $\frac{\Delta_p}{\lambda_1} = \eta > 0$, where $\eta \in \{0.2, 0.5, 0.8\}$.
- $p = 3$: $\mathbf{\Sigma}_1 = \mathrm{diag}(1, 1, 1)$, and then $\frac{\Delta_p}{\lambda_1} = 1 - \frac{|g_1|}{n} > 0$.

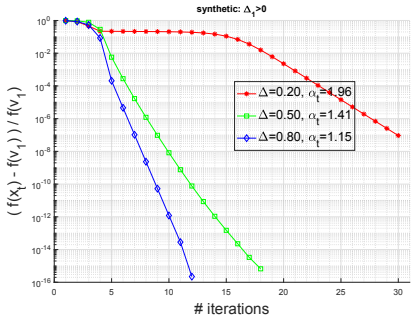
For the conventionally considered case that $\Delta_1 > 0$, two types of convergence curves, in terms of the relative function error $\frac{f(\mathbf{x}_t) - f(\mathbf{v}_1)}{f(\mathbf{v}_1)}$ and the commonly used potential function $\sin^2 \theta(\mathbf{x}_t, \mathbf{v}_1)$, respectively, are shown 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 $\alpha_{t+1} = \frac{c}{\tau + t}$ (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)-1(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 $\mathbf{x}_t$ will converge to finally, it is necessary for us to change the potential from the commonly used $\sin^2 \theta(\mathbf{x}_t, \mathbf{v}_1)$ to $\sin^2 \theta(\mathbf{x}_t, \mathbf{V}_p)$ in this case. As demonstrated in Figure 1(c),

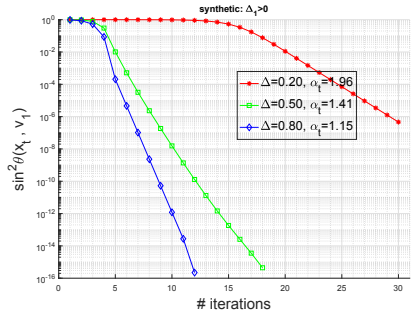


(a) Relative function error (b) Potential $\sin^2\theta(\mathbf{x}_t, \mathbf{V}_p)$ (c) Improper potential $\sin^2\theta(\mathbf{x}_t, \mathbf{v}_1)$

Figure 1: Synthetic data: $\Delta_1 = 0$

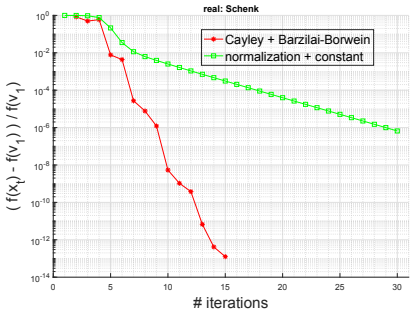


(a) Relative function error

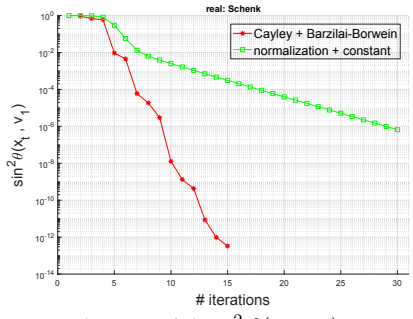


(b) Potential $\sin^2\theta(\mathbf{x}_t, \mathbf{v}_1)$

Figure 2: Synthetic data: $\Delta_1 > 0$



(a) Relative function error



(b) Potential $\sin^2\theta(\mathbf{x}_t, \mathbf{v}_1)$

Figure 3: Real data.

$\sin^2\theta(\mathbf{x}_t, \mathbf{v}_1)$ is unable to converge, which implies that $\mathbf{x}_t$ has converged to another leading eigenvector in $\mathrm{span}(\mathbf{V}_p)$.

We run two implementations of the solver on a real symmetric matrix $\mathbf{A}$, named **Schenk**[1], 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.

[1]www.cise.ufl.edu/research/sparse/matrices/

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 Δ_p eliminates the limitation of the commonly considered eigengap Δ_1. When Δ_p is large, the convergence is linear, which is described by the first general analysis. If it is as small as ϵ, the second general analysis shows that it is sub-linear $O(\frac{1}{\epsilon})$ rather than $O(\frac{1}{\epsilon^2})$. 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

[Absil *et al.*, 2008] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

[Arora *et al.*, 2012] Raman Arora, Andrew Cotter, Karen Livescu, and Nathan Srebro. Stochastic optimization for PCA and PLS. In *50th Annual Allerton Conference on Communication, Control, and Computing, Allerton*, pages 861–868, 2012.

[Arora *et al.*, 2013] Raman Arora, Andrew Cotter, and Nati Srebro. Stochastic optimization of PCA with capped MSG. In *NIPS*, pages 1815–1823, 2013.

[Balcan *et al.*, 2016] Maria-Florina Balcan, Simon Shaolei Du, Yining Wang, and Adams Wei Yu. An improved gap-dependency analysis of the noisy power method. In *COLT*, pages 284–309, 2016.

[Balsubramani *et al.*, 2013] Akshay Balsubramani, Sanjoy Dasgupta, and Yoav Freund. The fast convergence of incremental pca. In *NIPS*, pages 3174–3182. Curran Associates, Inc., 2013.

[Bonnabel, 2013] Silvere Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Trans. Automat. Contr.*, 58(9):2217–2229, 2013.

[Edelman *et al.*, 1999] Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20(2):303–353, April 1999.

[Fan *et al.*, 2018] Jianqing Fan, Qiang Sun, Wen-Xin Zhou, and Ziwei Zhu. Principal component analysis for big data. *arXiv preprint arXiv:1801.01602*, 2018.

[Garber *et al.*, 2016] Dan Garber, Elad Hazan, Chi Jin, Sham M. Kakade, Cameron Musco, Praneeth Netrapalli, and Aaron Sidford. Faster eigenvector computation via shift-and-invert preconditioning. In *ICML*, pages 2626–2634, 2016.

[Golub and Van Loan, 1996] Gene H. Golub and Charles F. Van Loan. *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, Baltimore, MD, USA, 1996.

[Hardt and Price, 2014] Moritz Hardt and Eric Price. The noisy power method: A meta algorithm with applications. In *NIPS*, pages 2861–2869, 2014.

[Liu *et al.*, 2016] Huikang Liu, Weijie Wu, and Anthony Man-Cho So. Quadratic optimization with orthogonality constraints: Explicit lojasiewicz exponent and linear convergence of line-search methods. In *ICML*, pages 1158–1167, 2016.

[Musco and Musco, 2015] Cameron Musco and Christopher Musco. Randomized block krylov methods for stronger and faster approximate singular value decomposition. In *NIPS*, pages 1396–1404, 2015.

[Ng *et al.*, 2002] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, pages 849–856. MIT Press, 2002.

[Pitaval *et al.*, 2015] Renaud-Alexandre Pitaval, Wei Dai, and Olav Tirkkonen. Convergence of gradient descent for low-rank matrix approximation. *IEEE Trans. Information Theory*, 61(8):4451–4457, 2015.

[Rudelson and Vershynin, 2010] Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices: extreme singular values. *arXiv preprint arXiv:1003.2990*, 2010.

[Sa *et al.*, 2015] Christopher De Sa, Christopher Re, and Kunle Olukotun. Global convergence of stochastic gradient descent for some non-convex matrix problems. In *ICML*, pages 2332–2341, 2015.

[Shamir, 2015] Ohad Shamir. A stochastic PCA and SVD algorithm with an exponential convergence rate. In *ICML*, pages 144–152, 2015.

[Shamir, 2016a] Ohad Shamir. Convergence of stochastic gradient descent for PCA. In *ICML*, pages 257–265, 2016.

[Shamir, 2016b] Ohad Shamir. Fast stochastic algorithms for SVD and PCA: convergence properties and convexity. In *ICML*, pages 248–256, 2016.

[Sra *et al.*, 2011] Suvrit Sra, Sebastian Nowozin, and Stephen J. Wright. *Optimization for Machine Learning*. The MIT Press, 2011.

[Wen and Yin, 2013] Zaiwen Wen and Wotao Yin. A feasible method for optimization with orthogonality constraints. *Math. Program.*, 142(1-2):397–434, 2013.

[Xu and Gao, 2018] Zhiqiang Xu and Xin Gao. On truly block eigensolvers via riemannian optimization. In *AISTATS*, pages 168–177, 2018.

[Xu and Ke, 2016a] Zhiqiang Xu and Yiping Ke. Effective and efficient spectral clustering on text and link data. In *CIKM*, pages 357–366, 2016.

[Xu and Ke, 2016b] Zhiqiang Xu and Yiping Ke. Stochastic variance reduced riemannian eigensolver. *CoRR*, abs/1605.08233, 2016.

[Xu *et al.*, 2017] Zhiqiang Xu, Yiping Ke, and Xin Gao. A fast algorithm for matrix eigen-decompositionn. In *UAI 2017, Sydney, Australia, August 11-15, 2017*, 2017.

[Zhang *et al.*, 2016] Hongyi Zhang, Sashank J. Reddi, and Suvrit Sra. Riemannian SVRG: fast stochastic optimization on riemannian manifolds. In *NIPS*, pages 4592–4600, 2016.