Noisy Riemannian Gradient Descent for Eigenvalue Computation with Application to Inexact Stochastic Recursive Gradient Algorithm

You-Lin Chen CYOULIN.TW@GMAIL.COM
Zhiqiang Xu ZHIQIANGXU2001@GMAIL.COM
Ping Li PINGLI98@GMAIL.COM
Cognitive Computing Lab, Baidu Research
10900 NE 8th St. Bellevue, WA 98004, USA

Editors: Emtiyaz Khan and Mehmet Gönen

Abstract

We provide a robust convergence analysis of the Riemannian gradient descent algorithm for computing the leading eigenvector of a real symmetric matrix. Our result characterizes the convergence behavior of the algorithm under the noisy updates, where noises can be generated by a stochastic process or could be chosen adversarially. The noisy Riemannian gradient descent has a broad range of applications in machine learning and statistics, e.g., streaming principal component analysis or privacy-preserving spectral analysis. In particular, we demonstrate the usefulness of our convergence bound with a new eigengap-dependent sample complexity of the inexact Riemannian stochastic recursive gradient algorithm, which utilizes mini-batch gradients instead of full gradients in outer loops. Our robust convergence paradigm strictly improves the state-of-the-art sample complexity in terms of the gap dependence.

Keywords: Sample complexity, principal component analysis, stochastic optimization

1. Introduction

Computing the top eigenvector of a real symmetric matrix is one of the fundamental problems in numerical linear algebra and has enormous applications to machine learning and statistics. For instance, principal component analysis (PCA) (Jolliffe and Cadima, 2016) is the most popular dimensionality reduction tool and widely used in data preprocessing and unsupervised learning. Formally, given a set of n independent and identically distributed (i.i.d.) samples $\{a_i\}_{i=1}^n$ drawn from some distribution with zero mean, define the empirical covariance matrix as $A_n = (1/n)\sum_{i=1}^n a_i a_i^\top \in \mathbb{R}^{d\times d}$, $\{u_i\}_{i=1}^d$ as the d eigenvectors of A_n, their corresponding eigenvalues as $\{\lambda_i\}_{i=1}^d$ such that $\lambda_1 = \cdots = \lambda_k > \lambda_{k+1} \geq \cdots \geq \lambda_d$ for some $k \geq 1$, and the i-th eigengap of A_n as $\Delta_i = \lambda_i - \lambda_{i+1}$. The goal of this paper is to recover the first eigenvector u_1, or, equivalently, to minimize the empirical risk as follows:

$$\min_{w\in\mathbb{R}^d:\|w\|_2=1} f(w) = -w^\top A_n w. \tag{1}$$

Although deterministic algorithms are well-studied for eigenvector computation, many applications in machine learning and statistics involve a variety of noise sources including

sampling, missing data, and privacy constraints. Several works (Hardt and Price, 2014; Balcan et al., 2016; Xu and Li, 2022) tackled this problem by providing a robust convergence analysis of the well-known power method without the need for ad-hoc analysis for different applications. However, given a tolerance tol and supposing that $\Delta_1 > 0$, Xu and Li (2022) applied robustness analysis with momentum acceleration to streaming PCA, which only leads to a sub-optimal sample complexity $\mathcal{O}\left(1/(\Delta_1^{5/2}\,\mathrm{tol}^2)\right)$ for achieving a tol-optimal solution, i.e. finding a solution w such that $1-(w^\top u_1)^2 \leq \mathrm{tol}$. Here, $a_n = \mathcal{O}(b_n)$ means there exists c such that $a_n \leq cb_n$ and a_n, b_n are positive sequences. In contrast, the optimal sample complexity is $\mathcal{O}\left(1/(\Delta_1^2\,\mathrm{tol})\right)$ and can be achieved by Oja's algorithm (i.e., the projected stochastic gradient descent for the leading eigenvector computation) (Jain et al., 2016). This gap raises a natural question:

Is it possible to improve rate's eigengap dependency from $1/(\Delta_1^{5/2}\,\mathrm{tol}^2)$ *to* $1/(\Delta_1^2\,\mathrm{tol})$ *for the robustness analysis framework of the eigenvector computation?*

Our work fills this missing but important part of the literature by revisiting the Riemannian gradient descent (RGD) for the eigenvector computation. The crucial observation in our work is that the noisy power method (Hardt and Price, 2014; Balcan et al., 2016; Xu and Li, 2022) can only reduce the noise by increasing the sample size, while our gradient descent method introduces a step size parameter which is important in the presence of noise since the noise magnitude can be reduced by decreasing the step size to achieve convergence. This critical insight leads to the improvement in our robustness analysis framework.

In this work, we start by investigating the noisy Riemannian gradient descent (NRGD), a general framework for the leading eigenvector computation when gradients can only be accessed through inaccurate matrix-vector products. That is, the gradient noises can be generated by a stochastic process or could be chosen adversarially. We then demonstrate the power of our robust convergence analysis by presenting a novel result for the stochastic variance-reduced gradient method without imposing any initial condition. Several works (Zhang et al., 2016; Kasai et al., 2018) developed variance reduction methods to general Riemannian manifolds and used the eigenproblem as their important examples. However, several facts reveal that very different techniques are needed for eigenproblem. First, the eigenvector computation is guaranteed to get the first eigenvector with a simple random initialization, but the conventional optimization analysis can only show convergence to stationary points. Second, the power method converges linearly, but Eq. (1) is not even locally strictly convex (Shamir, 2016). It is worth mentioning that one can characterize convexity-like properties in an open neighborhood of the global optimum with eigengap (Zhang et al., 2016, Theorem 4). The step-size however introduces an extra eigengap. Consequently, most convergence results in non-convex optimization only cover sub-linear convergence, and the analysis relying on convexity-like properties is not tight in terms of eigengap since it is impossible to prove a convergence rate that is independent of the eigengap. Third, the conventional optimization analysis relies on the condition number, which is equal to the ratio between the largest and smallest eigenvalue for Eq. (1). This convergence rate is not tight since the power method only depends on the linear term of the eigengap Δ_1. Moreover, the power method can work even when the smallest eigenvalue is zero. All those observations indicate that the eigenvector computation is so special that different analysis

approaches are required, and general theories in optimization literature fail to explain the theoretical behavior of Eq. (1).

Our main **contributions** are summarized as follows:

1. To the best of our knowledge, we are the first to provide a robustness analysis of the Riemannian first-order optimization method for the eigenvector computation problem. Our conditions match those in the existing robustness analysis for power method (Hardt and Price, 2014; Balcan et al., 2016; Xu and Li, 2022). Moreover, in the two dimensional space, we provide the first result of noisy gradient methods achieving the same convergence rate as the power method, which matches the result in Ding et al. (2020) under the deterministic case.
2. We obtain the eigengap-dependent sample complexity of the Riemannian stochastic recursive gradient algorithm (Kasai et al., 2018) as a novel application of the framework of robust convergence analysis for the eigenvector computation problem. Our result strictly improves the state-of-the-art gap-dependent sample complexity of steaming PCA in Hardt and Price (2014); Balcan et al. (2016); Xu and Li (2022).
3. Our convergence result is global and works for arbitrary step sizes as long as the sample size is large enough when the step size is small.

Organization. Section 2 provides literature review. Section 3 presents the noisy Riemannian gradient descent and its analysis. Section 4 applies the robust convergence to stochastic recursive gradient algorithm and gets its eigengap-dependent sample complexity. Section 5 provides experiments for comparing existing methods on both synthetic and real datasets. Section 6 gives conclusions and directions for future works. The proofs are deferred to the appendix.

2. Related Works

The power method (Golub and Van Loan, 2012) serves as the standard method for solving Eq. (1). It requires a total number of samples at $O\left((n/\Delta_1)\log(1/\operatorname{tol})\right)$, if we regard the given samples as being duplicated in each iteration. For gradient descent (Oja's rule) or Riemannian gradient descent (Krasulina's methed), Ding et al. (2020) showed the sample complexity of gradient methods are the same as the power method, i.e., $O\left((n/\Delta_1)\log(1/\operatorname{tol})\right)$. Xu and Li (2021a) further improved the sample complexity to $O\left((n/\max\{\Delta_1, \operatorname{tol}\})\log(1/\operatorname{tol})\right)$ for RGD. However, it is hard to apply the result in Xu and Li (2021a) to the robustness analysis since their analysis involves many calculations on some complicated products or sums of sequences. The noise in the robustness analysis introduces an extra term of error, and it is not easy to handle the effect of the error term on recursive formulations from gradient algorithms. This is also the reason why we cannot extend the argument in Ding et al. (2020) for the two-dimensional space to higher dimensional spaces (see more details in Section 3.3.)

In real-world applications, datasets might be too large to have a full pass (n is extremely large). Stochastic methods such as Oja's algorithm (Oja and Karhunen, 1985) and incremental algorithms (Arora et al., 2012, 2013) were developed to address this issue of the

Table 1: This table shows a comparison of first-order methods for eigenvalue computation. Note that $a \wedge b = \min\{a, b\}$. "I" means if the algorithm imposes condition on initialization or not. "G" represents if the algorithm converges to global optimum or not. "M" indicates if the convergence analysis of the algorithm can use mini-batch gradients only. "S" demonstrates if the algorithm is stochastic. "V" reveals if the algorithm employs variance reduction

Reference	gap-dependent sample complexity	I	G	M	S	V
Power method (Golub and Van Loan, 2012)	$\mathcal{O}\left(\frac{n}{\Delta_1}\log(1/\operatorname{tol})\right)$	no	yes	no	no	no
Lanczos algorithm (Golub and Van Loan, 2012)	$\mathcal{O}\left(\frac{n}{\sqrt{\Delta_1}}\log(1/\operatorname{tol})\right)$	no	yes	no	no	no
RGD (Xu and Li, 2021a)	$\mathcal{O}\left(\frac{n}{\max\{\Delta_1,\operatorname{tol}\}}\log(1/\operatorname{tol})\right)$	no	yes	no	no	no
Oja's Algorithm (Jain et al., 2016)	$\mathcal{O}\left(\frac{1}{\Delta_1^2\operatorname{tol}}\right)$	no	yes	yes	yes	no
streaming PCA (Hardt and Price, 2014)	$\mathcal{O}\left(\frac{1}{\Delta_1^3\operatorname{tol}^2}\right)$	no	yes	yes	yes	no
Faster Noisy Power Method (Xu and Li, 2022)	$\mathcal{O}\left(\frac{1}{\Delta_1^{5/2}\operatorname{tol}^2}\right)$	no	yes	yes	yes	no
LazySVD (Allen-Zhu and Li, 2016)	$\mathcal{O}\left(\left(n+\frac{n^{3/4}}{\Delta_k^{1/2}}\right)\log\left(\frac{1}{\operatorname{tol}}\right)\right)$	no	yes	no	yes	no
VR-PCA (Shamir, 2015)	$\mathcal{O}\left(\left(n+\frac{1}{\Delta_k^2}\right)\log\left(\frac{1}{\operatorname{tol}}\right)\right)$	yes	yes	no	yes	yes
RSVRG (Xu and Gao, 2018)	$\mathcal{O}\left(\left(n+\frac{1}{\Delta_1^2}\right)\log\left(\frac{1}{\operatorname{tol}}\right)\right)$	no	yes	no	yes	yes
Corollary 5	$\mathcal{O}\left(\left(\frac{\log(1/\operatorname{tol})}{\Delta_k^2\operatorname{tol}}\wedge n+\frac{1}{\Delta_k^3}\right)\log\left(\frac{1}{\operatorname{tol}}\right)\right)$	no	yes	yes	yes	yes
Corollary 5 (2-dim)	$\mathcal{O}\left(\left(\frac{\log(1/\operatorname{tol})}{\Delta_k^2\operatorname{tol}}\wedge n+\frac{1}{\Delta_k^2}\right)\log\left(\frac{1}{\operatorname{tol}}\right)\right)$	no	yes	yes	yes	yes

large scale data. Stochastic algorithms like stochastic gradient descent (SGD) assume we can query some fresh data points from some distribution at each iteration. While the computational cost of SGD-type algorithms is much cheaper at each iteration, typically, their convergence rate is significantly slower than deterministic methods. In particular, if only one data point is accessed at each iteration, it can be shown by matrix Bernstein inequality that the minimax lower bound of the subspace distance is $\Omega(\sigma^2/(n\Delta_1^2))$ (Vu et al., 2013), where σ^2 is the variance of data and $a_n = \Omega(b_n)$ means there exists c such that $a_n \geq cb_n$. Thus, online algorithms can only get a sublinear convergence rate with a quadratic term of the eigengap on general data distributions.

Stochastic variance-reduced methods serve as a remedy for a sublinear convergence rate of SGD and are superior in terms of the sample complexity (runtime) compared to deterministic algorithms or SGD (Gower et al., 2020). Inspired by stochastic variance-reduced gradient (SVRG) (Johnson and Zhang, 2013), Shamir (2015) proposed a stochastic variance-reduced version of Oja's algorithm (VR-PCA). With variance reduction, VR-PCA works with a constant step size and has a linear convergence rate but with a quadratic term in eigengap, i.e., the sample complexity is $\mathcal{O}\left(\left(n+1/\Delta_k^2\right)\log(1/\operatorname{tol})\right)$. However, the understanding of the gap-dependent sample complexity still remains inadequate, and it was conjectured that the sample complexity should linearly depend on the eigengap (Shamir, 2015).

We end this section by summarizing a comparison of recent eigensolvers in terms of the eigengap-dependent sample complexity and several perspectives in Table 1.

3. Noisy Riemannian Gradient Descent for Eigenvalue Computation

This section gives our main result of the noisy Riemannian gradient descent for the eigenvector computation. Before proceeding with our algorithm and analysis, essential notations are introduced, including a potential function for measuring the closeness of iterates to a globally optimal solution. Due to the space limitations, we put related concepts, such as Riemannian optimization and stochastic recursive gradients, in the appendix.

3.1. Notations and notions

We use $\|\cdot\|$ to denote the 2-norm of a matrix or vector. Define $[n] = \{1, 2, \ldots, n\}$. For a function f, we denote ∇f as the gradient of f. Let $U_k = (u_1, \ldots, u_k)$ be the leading eigenspace. Given two unit vectors u, v, i.e. $\|u\| = 1 = \|v\|$, we define the angle between u and v as $\theta(u, v) = \arccos(u^\top v)$. Then, the principal angle between w and the space $\mathrm{span}(U_k)$ is $\theta(w, U_k) = \min_{u \in \mathrm{span}(U_k)} \theta(w, u)$, where $\mathrm{span}(U_k)$ is the subspace spanned by columns of U_k.

Instead of focusing on a conventional assumption that $\Delta_1 > 0$, we target a general framework relying on the generalized eigengap Δ_k and the leading eigenspace U_k. Thus, it suffices to show the convergence to one of the globally optimal solutions, i.e., a unit vector $u \in \mathrm{span}(U_k)$, instead of a specific solution, e.g., u_1. Given a sequence of unit vectors, $\{w_t\}$, for measuring the progress of iterates to the leading eigenspace, we define a potential function $\psi(w_t, U_k) = -2 \log \|U_k^\top w_t\|$ (Xu and Li, 2021b). We let $\psi_t = \psi(w_t, U_k)$ to simplify notation. Note that $\|U_k^\top w_t\| \leq \|U_k\| \|w_t\| = 1$, so $\psi(w_t, U_k) \geq 0$. Thus, ψ indeed is a well-defined potential function since $\psi(u, U_k) = 0$ for all u in the column space of U_k.

Since $\|U_k^\top w_t\|^2 = \cos^2 \theta(w_t, U_k)$, we can write $\psi(w_t, U_k) = \min_{u \in \mathrm{span}(U_k): \|u\|=1} \psi(w_t, u)$, where $\psi(w_t, u) = -2 \log |u^\top w_t| = -\log \cos^2 \theta(w_t, u)$. With above equations and the simple fact that $\psi(w_t, u) = -\log(1 - \sin^2 \theta(w_t, u)) \geq \sin^2 \theta(w_t, u)$, we have the following relation between the conventional potential function $\sin^2 \theta(w_t, u)$ and our proposed potential function $\psi(w_t, u)$.

$$\cos^2 \theta(w_t, u) = \exp(-\psi(w_t, u)), \quad \sin^2 \theta(w_t, u) \leq \psi(w_t, u). \tag{2}$$

As we can see, the conventional potential function $\sin^2 \theta(w_t, u)$ is dominated by our potential function, which provides a relationship between the conventional potential function and our proposed potential function.

3.2. Convergence argument

This section aims to develop the convergence theorem of the noisy Riemannian gradient descent (NRGD) which is defined as follows:

$$w'_{t+1} = w_t + \eta P_t(A_n w_t + \varepsilon_t), \quad w_{t+1} = w'_{t+1} / \|w'_{t+1}\|. \tag{3}$$

where $P_t := I_d - w_t w_t^\top$ and ε_t is the noise that can be generated by a stochastic process or could be chosen adversarially. Compared to SGD (Oja, 1982; Jain et al., 2016), stochastic RGD (Krasulina, 1969) remains much less studied. Nevertheless, recent works provide evidence that RGD may have better theoretical properties than Oja's rule. For example, Li

et al. (2018) used stochastic RGD to obtain a finer estimate than the minimax lower bound for sub-Gaussian data i.e. $\mathcal{O}\left((\operatorname{tol}\Delta_1)^{-1}\sum_{k=2}^{d}(\lambda_k/\lambda_1-\lambda_k))\right)$, compared to the result for Oja's rule (Jain et al., 2016), $\mathcal{O}\left(d(\Delta_1^2\operatorname{tol})^{-1}\right)$.

In what follows, we state our main theorem which establishes the convergence in terms of the deterministic error bound. Specifically, this theorem says that if η is small enough and the norm of noise ε is upper bounded by the initial value of potential function ψ_0, we can decrease the potential function ψ_t at every step.

Theorem 1 *Assume that w_t is updated by Eq. (3). Given $\rho>0$, $\beta\in(0,1)$, and an initial point w_0 such that $\cos^2\theta_0>\gamma^2$ for some $\gamma>0$, suppose that $\|\varepsilon_t\|^2\le\rho^2\sin^2\theta_0$ for all t and*

$$\eta<\min\left\{\frac{\gamma}{2\rho},\frac{\Delta_k}{16\lambda_1(1+\eta\Delta_k)},\frac{\Delta_k}{\rho^2(8+8\eta\Delta_k)}\frac{\beta}{1+\psi_0}-\frac{(1+\psi_0)}{\gamma\rho-2\eta\rho^2}\right\}. \tag{4}$$

Then, we have $\psi_{t+1}<\psi_0$. Moreover, it holds that

$$\psi_{t+1}\le\left((1-r)^t+\frac{\beta}{2}\right)\psi_0, \tag{5}$$

where $r=\frac{\eta\Delta_k}{2+2\eta\Delta_k}$ and $\psi_t=\psi(w_t,U_k)$.

To the best of our knowledge, Theorem 1 is the first robustness analysis of the Riemannian first-order optimization method for the eigenvector computation problem. Choosing

$$t=\frac{\log(2/\beta)}{-\log(1-r)}\approx\frac{\log(2/\beta)}{r}=\widetilde{O}\left(\frac{1}{\eta\Delta_k}\right),$$

where $\widetilde{O}$ hiddens log factors. Eq. (5) becomes $\psi_{t+1}\le\beta\psi_0$. Hence, given any $\operatorname{tol}>0$, it takes $\widetilde{O}(1/(\eta\Delta_k))$ steps to have $\psi_{t+1}<\operatorname{tol}$, which is linear convergence as the same as power methods. Furthermore, if we can carefully decrease the noise level, the iterates in Eq. (3) could approach the leading eigenspace U_k. We end this section by providing more remarks for our theorem:

1. The condition (4) ensures that $(1-r)$ in Eq. (5) is positive. For a sufficiently small β, Theorem 1 requires the noise level ρ to be also sufficiently small. In other words, there always exists a small enough ρ such that the right-hand side of Eq. (4) is satisfied.

2. If $k=1$, i.e. $\Delta_1>0$, then $\cos^2\theta_0>0$ almost surely with a random initialization. Hence, our theorem is global convergence by convention. Xu and Gao (2018, Lemma 4.7) further provides the probability of the event that $\cos^2\theta_0>\gamma$ when $\Delta_1=0$ and $\Delta_k>0$.

3. We can derive the sufficient condition implying Eq. (4) is positive. Since $\gamma/2\rho$ and $\Delta_k/(16\lambda_1(1+\eta\Delta_k))$ are positive, it suffices to show $\frac{\Delta_k}{\rho^2(8+8\eta\Delta_k)}\frac{\beta}{1+\psi_0}-\frac{(1+\psi_0)}{\gamma\rho-2\eta\rho^2}$ is positive. Assuming $\eta\le\frac{\gamma}{4\rho}$ and $\rho\le\frac{\gamma\Delta_k\beta}{16(1+\eta\Delta_k)(1+\psi_0)^2}$, we have $\gamma-2\eta\rho\ge\frac{\gamma}{2}$ which implies

$$\frac{\Delta_k\beta}{8(1+\eta\Delta_k)(1+\psi_0)^2}\ge\frac{\rho}{\gamma/2}\ge\frac{\rho}{\gamma-2\eta\rho}. \tag{6}$$

It is not hard to see that Eq.(6) implies Eq.(4) is positive, so ρ should be less than $\mathcal{O}(\Delta_k)$ when Δ_k is small. This result matches the noise bound of power iterations (Hardt and Price, 2014, Corollary 1.1).

4. Our result extends the conventionally considered eigengap Δ_1 to the generalized eigengap $\Delta_k = \lambda_k - \lambda_{k-1} > 0$, which is critical because the gap-dependent bound for the noise may be restrictive for the algorithm's applicability and hurt the convergence and sample complexity.

5. Theorem 1 requires that η needs to be $\mathcal{O}(\Delta_k)$, which is unsatisfactory since the theoretical upper bound for RGD in the deterministic case is $1/\lambda_1$. We will show how the step size can be improved for the two-dimensional case in the next section.

3.3. Refined analysis: two-dimensional case

The key bottleneck of the previous result is that η needs to be $\mathcal{O}(\Delta_k)$. However, the theoretical upper bound of η for RGD should be $1/\lambda_1$ (see Appendix A) in the deterministic case. As a result, setting $\eta = \mathcal{O}(\Delta_k)$ slows down the convergence considerably and is not optimal. In fact, Ding et al. (2020) improved this upper bound of the step size in the deterministic case based on the observation that the worst-case convergence rate of RGD occurs when the initial vector lies in the space spanned by the first two largest eigenvectors. The crucial insight is that the iterates of the deterministic power method and gradient descent only stay in the space spanned by the first two largest eigenvectors as long as the initial vector is in such space, which is not true for stochastic algorithms. Therefore, it suffices to conduct analysis in a two-dimensional space for deterministic algorithms, but this argument cannot be easily extended to the stochastic setting. In this section, we only show that it is possible to improve the step size in this two-dimensional space for NRGD. The following theorem gives the details.

Theorem 2 *Assume that w_t is updated by Eq. (3). Given $\rho_1, \rho_2 > 0$ and $\beta, \gamma \in (0, 1)$, suppose that*

$$|\cos\theta_0| \geq \gamma, \quad \|\varepsilon_t\| \leq \min\{\rho_1\Delta_1|\cos\theta_0|, \rho_2\Delta_1|\sin\theta_0|\},$$

for all t and $\eta \leq 1/\lambda_1$ such that $2\rho_2\eta \leq \gamma\beta(\eta - \rho_1)$. Then it holds that $|\tan\theta_t| \leq |\tan\theta_0|$ and

$$|\tan\theta_{t+1}| \leq \left((1-r)^t + \frac{\beta}{2}\right)|\tan\theta_0|,$$

where $\theta_t = \theta(w_t, u_1)$ and $r = \frac{(\eta-\rho_1)\Delta_1}{1+(\eta-\rho_1)\Delta_1}$.

We can see that the condition on the step size is improved from $\mathcal{O}(\Delta_1)$ to $\mathcal{O}(1/\lambda_1)$. As a result, to achieve the *tol*-optimal solution, t can be $\widetilde{O}(1/\Delta_1)$ in the two dimensional case, which achieve the same convergence rate as the power method under the case that the norm of noise is $O(\Delta_1)$. As far as we know, this is the first result of gradient methods with noise achieving the same convergence rate as the power method. The potential function is changed since we use a different technical approach and k can be only 2 in the two-dimensional space. Using the fact that $\sin\theta_t \leq \tan\theta_t$, Theorem 2 indeed provides an upper bound of the conventional potential function $\sin\theta_t$, where $\theta_t = \theta(w_t, u_1)$.

Algorithm 1 iRSRG

Data: $\{a_t\}_{t=1}^n$

Input: the initial point $\widetilde{w}_0$, the step size η, the number of iterations of the inner loop and the outer loop M, m, the mini-batch sizes of the inner loop and the outer loop B_s, b

for $s = 1 \dots M$ **do**

  if $B_s \geq n$ **then**

    $v_0 = \frac{1}{n} \sum_{t=1}^n a_t a_t^\top \widetilde{w}_{s-1}$

  else

    Draw $\{i_j\}_{j=1}^{B_s}$ uniformly from $[n]$

    Compute $v_0 = \frac{1}{B_s} \sum_{j=1}^{B_s} a_{i_j} a_{i_j}^\top \widetilde{w}_{s-1}$

  end

  $w_0 = \widetilde{w}_{s-1}$

  $w_1 = \left(w_0 + \eta(I_d - w_0 w_0^\top)v_0\right) / \|w_0 + \eta(I_d - w_0 w_0^\top)v_0\|$

  for $t = 1 \dots m$ **do**

    Draw $\{i_j\}_{j=1}^{b}$ uniformly from $[n]$

    Update $v_t = \frac{1}{b} \sum_{j=1}^{b} a_{i_j} a_{i_j}^\top (w_t - w_{t-1}) + v_{t-1}$

    $w_{t+1} = \left(w_t + \eta(I_d - w_t w_t^\top)v_t\right) / \|w_t + \eta(I_d - w_t w_t^\top)v_t\|$

  end

  $\widetilde{w}_s = w_m$

end

Result: $\widetilde{w}_M$

4. Application to Stochastic Recursive Gradients

In this section, we combine NRGD and stochastic recursive gradient algorithm (SARAH) (Nguyen et al., 2017a; Fang et al., 2018) as a novel application of our robustness analysis in the previous section. In particular, we will use Theorem 1 and 2 to analyze the convergence property of SARAH for the top eigenvalue computation problem. Compared to SVRG, SARAH employs a biased gradient estimation and has a better convergence rate in a non-convex setting (Nguyen et al., 2017b). Technically speaking, the martingale property of SARAH provides a better variance bound and uniform error control, which significantly improves the previous variance bound for SVRG that Jiang et al. (2017) showed (see Proposition 3). We give a comprehensive description for our proposed algorithm (inexact Riemannian stochastic recursive gradient (iRSRG)) in Algorithm 1, which can use mini-batch gradients in outer loops.

Applying the result in Section 3 requires the noise to be controlled at a certain level (meet the condition $\|\varepsilon_t\| \leq \rho^2 \sin^2 \theta_0$ in Theorem 1, for example) with respect to the eigengap at *every* step. Then Gaussian distribution concentrations are employed (Hardt and Price, 2014), so the noise is uniformly bounded with a high probability for streaming PCA provided that the sample size is large enough. We use a different approach that leverages the martingale property of stochastic recursive gradients to obtain a uniform bound of the variance with high probability via Doob's maximal inequality for submartingale (Durrett, 2019, Thm 5.4.2). The key ingredient is to show the sum of the variance of stochastic recur-

sive gradients can be bounded by the initial value $\sin^2\theta_0$. The next proposition summarizes all auxiliary results.

Proposition 3 *Suppose that* $\max_t\{\|a_t\|^2, \|a_t a_t^\top - A_n\|^2\} \le K$, *and let* $P_t := I - w_t w_t^\top$ *and* $\widehat{L} = \sqrt{10}\lambda_1 + K/2$ *and* $\tau = 1 - \frac{\eta\widehat{L}}{2} - m\eta^2\frac{K^2}{b}$ *where* w_t, v_t, b, B_s *are defined in Algorithm 1. 1) For any* $d > 2$, *we have*

$$\tau\eta\sum_{t=1}^{m}\mathbb{E}[\|P_t v_t\|^2|\mathcal{F}_0] \le \lambda_1\sin^2\theta_0 + \eta m\|v_0 - A_n w_0\|,$$

where $\mathcal{F}_0$ *is defined in Appendix C, and setting* $B_s = B/sin^2\theta_0$ *implies*

$$\mathbb{P}\left\{\max_{1\le t\le m}\frac{\|v_t - Aw_t\|^2}{\sin^2\theta_0} > \rho^2\right\} \le \frac{K^2}{\rho^2 B} + \frac{\lambda_1 K^2\eta}{\rho^2 b\tau} + \frac{m\eta^2K^4}{b\tau\rho^2 B}.$$

2) For two-dimensional space, i.e. $d = 2$, *we have*

$$\tau\eta\sum_{t=1}^{m}\mathbb{E}[\|P_t v_t\|^2|\mathcal{F}_0] \le \Delta_1\sin^2\theta_0 + \eta m\|v_0 - A_n w_0\|,$$

and setting $B_s = B/sin^2\theta_0$ *implies*

$$\mathbb{P}\left\{\max_{1\le t\le m}\frac{\|v_t - Aw_t\|^2}{\sin^2\theta_0} > \rho^2\right\} \le \frac{K^2}{\rho^2 B} + \frac{\Delta_1 K^2\eta}{\rho^2 b\tau} + \frac{m\eta^2K^4}{b\tau\rho^2 B}.$$

We have several observations. First, the boundedness condition on a_t is the typical condition for stochastic methods for eigenvalue computation (Jain et al., 2016; Shamir, 2015). Second, not only the step-size can also be improved in the two-dimensional space but also we can replace λ_1 with Δ_1, which is crucial to makes the sample complexity better in the two-dimensional case. Third, we have the formula to control the variance without any assumption on convexity. It is important to avoid using convexity-like properties for getting a sharper gap-dependent sample complexity. To see this point, Zhang et al. (2016, Theorem 4) stated that there is an open neighborhood of u_1 that Eq. (1) is gradient dominated, so it is easy to combine this result and the analysis in Riemannian optimization to get the gap-dependent sample complexity. However, this approach probably introduces an extra eigengap in the convergence rate, which is not tight in terms of eigengap since it may be impossible to prove a convergence rate that is independent of the eigengap. Finally, Jiang et al. (2017, Lemma 3.6) provided a similar result of Proposition 3 for SVRG. However, their result depends on the exponential term of m, which may not be practical. In contrast, our estimation depends linearly on m only, which significantly improves their result and increases the sample efficiency. This improvement is due to the martingale property of SARAH and illustrates the benefit of SARAH.

Next, we present our convergence analysis of Algorithm 1. Note that given sequences of unit vectors $\{w_t\}$, $\{\widetilde{w}_s\}$, we let $\theta_t = \theta(w_t, U_k)$ and $\widetilde{\theta}_s = \theta(\widetilde{w}_s, U_k)$.

Theorem 4 *Suppose that* $\max_t\{\|a_t\|^2, \|a_t a_t^\top - A_n\|^2\} \le K$. *Given* $\beta, \gamma \in (0, 1)$, *we run Algorithm 1 with an initial point* w_0 *such that* $\cos^2\theta_0 \ge \gamma^2$.

1) For any $d > 2$, given $\rho > 0$, we set

$$m = r^{-1}\log(2/\beta), \quad M = \log(\psi_0/\operatorname{tol})/\log(1/\beta),$$

and a step size η such that

$$\eta \leq \min\left\{\frac{\gamma}{2\rho}, \frac{\Delta_k}{16\lambda_1(1+\eta\Delta_k)}, \frac{\Delta_k}{\rho^2(8+8\eta\Delta_k)}\frac{\beta}{1+\psi_0} - \frac{1+\psi_0}{\gamma\rho - 2\eta\rho^2}\right\}, \tag{7}$$

and $\tau > 0$. Then it holds that $\psi(\widetilde{w}_M, U_k) < \operatorname{tol}$ with probability at least

$$1 - M\left(\frac{K^2}{\rho^2 B} + \frac{\lambda_1 K^2 \eta}{\rho^2 b\tau} + \frac{m\eta^2 K^4}{b\tau\rho^2 B}\right) - p_k(\gamma). \tag{8}$$

where $\tau = 1 - \frac{\eta\widehat{L}}{2} - m\eta^2\frac{K^2}{b}$ and $r = \eta\Delta_k/(2+2\eta\Delta_k)$ and $p_k(\gamma)$ is defined on Lemma 8.
2) For two-dimensional space, $d = 2$, given $\rho_1, \rho_2 \geq 0$, set

$$m = r^{-1}\log(2/\beta), M = \log(|\tan\widetilde{\theta}_0|/\operatorname{tol})/\log(1/\beta),$$

and step size $\eta \leq 1/\lambda_1$ such that $\tau_1 > 0$ and $\tau_2 > 0$, then we have $|\tan\widetilde{\theta}_M| \leq \operatorname{tol}$ with probability at least

$$1 - p_k(\gamma) - M\left(\frac{K^2}{\gamma^2\rho_1^2\rho_2^2\Delta_1^2 B} + \frac{K^2\eta}{\gamma^2\rho_1^2\rho_2^2\Delta_1\tau_2 b} + \frac{m\eta^2K^4}{b\tau\gamma^2\rho_1^2\rho_2^2\Delta_1^2 B}\right). \tag{9}$$

where $\tau_1 = \gamma\beta(\eta-\rho_1) - 2\rho_2\eta$, $\tau_2 = 1 - \frac{\eta\widehat{L}}{2} - m\eta^2\frac{K^2}{b}$, $r = (\eta-\rho_1)\Delta_1/(1+(\eta-\rho_1)\Delta_1)$ and $p_k(\gamma)$ is defined on Lemma 8.

We make some remarks about our theorems.

1. For any k, $0 < p(\gamma) < 1$. Moreover, $p_k(\gamma) \to 0$ as $\gamma \to 0$.

2. Most theorems (Johnson and Zhang, 2013; Nguyen et al., 2021) for stochastic variance reduction methods use the averaging iterates $(1/m)\sum_{t=1}^{m} w_t$ to update $\widetilde{w}_s$ instead of the last iterate w_m. On the other hand, our algorithm works and analysis holds for the last iterate.

3. Compared to the result of VR-PCA (Shamir, 2015), we have the same order on the parameter of η, m. However, Theorem 4 does not require that the initial point is close to U_k. As the convergence holds with high probability for any random initial iterate, it is global by convention.

4. By choosing B and b large enough, the probability that iRSRG fails decreases, which means our result could always be non-trivial.

5. We can choose η small enough to satisfy Eq. (7). Since the step size η depends on the initial value γ, larger γ allows a large step size and implies a faster convergence rate. However, it is well-known that $\cos\theta(\widetilde{w}_0, u) < O(1/\sqrt{d})$ with high probability if $\widetilde{w}_0$ is uniformly sampled (Hardt and Price, 2014). Warm-start solvers or Oja's algorithm

can be used to speed up the algorithm at the beginning. Based on our theorem, there is no need to access all or large amounts of data to achieve linear convergence at the beginning. Thus, fewer samples are required to get the same accuracy compared to other stochastic variance reduction algorithms. This is equivalent to warm-start schemes, and our theorem provides a uniform perspective by considering the inexact version of the stochastic recursive gradient.

6. Our theorem does not necessarily require full gradients in the outer loop, which implies our proposed algorithm can work for streaming data. However, to achieve arbitrary accuracy, increasing B is necessary, which is typical for variance-reduced methods using mini-batch gradients in the outer loop (Lei and Jordan, 2017; Nguyen et al., 2021).

The next corollary shows the existence and the order of parameters in Theorem 4 and derives the total sample complexity of Algorithm 1 under the circumstance that Δ_k is small and $\widetilde{w}_0$ is close to U_k.

Corollary 5 *1) For any $d > 2$, suppose that $\beta = 0.5$, $\Delta_k \leq 1$ and $\widetilde{w}_0$ such that $\cos^2\theta_0 \geq \gamma^2$ and $1 - 1.99p_k(\gamma) \geq 0.5$. Then, there exist η, B, b, ρ such that*

$$\eta = \mathcal{O}(\Delta_k), \quad b = \Omega(1/\Delta_k), \quad B = \Omega(\log(1/\operatorname{tol})/\Delta_k^2), \quad \rho = \mathcal{O}(\Delta_k),$$

Eq. (7) is satisfied, and $\tau > 0$. Furthermore, the sample complexity of Algorithm 1 to achieve tol*-accuracy, i.e, $\sin^2\widetilde{\theta}_s \leq \operatorname{tol}$, with a probability in Eq. (8) that is larger than $0.99p_k(\gamma)$, is*

$$\mathcal{O}\left(\left(\frac{\log(1/\operatorname{tol})}{\Delta_k^2 \operatorname{tol}} \wedge n + \frac{1}{\Delta_k^3}\right)\log\left(\frac{1}{\operatorname{tol}}\right)\right).$$

2) For two-dimensional space, $d = 2$, suppose that $\beta = 0.5$, $\Delta_k \leq \lambda_1$ and $\widetilde{w}_0$ such that $\cos^2\theta_0 \geq \gamma^2$ and $1 - 1.99p_k(\gamma) \geq 0.5$. Then, there exist η, B, b, ρ such that

$$\eta = \mathcal{O}(1), \quad b = \Omega(1/\Delta_k), \quad B = \Omega(\log(1/\operatorname{tol})/\Delta_k^2), \quad \rho_1 = \mathcal{O}(1), \quad \rho_2 = \mathcal{O}(1),$$

and $\tau_1 > 0$ and $\tau_2 > 0$. Furthermore, the sample complexity of Algorithm 1 in two-dimensional space to achieve tol*-accuracy, i.e, $\sin^2\widetilde{\theta}_s \leq \operatorname{tol}$, with a probability in Eq. (9) which is larger than $0.99p_k(\gamma)$, is*

$$\mathcal{O}\left(\left(\frac{\log(1/\operatorname{tol})}{\Delta_k^2 \operatorname{tol}} \wedge n + \frac{1}{\Delta_k^2}\right)\log\left(\frac{1}{\operatorname{tol}}\right)\right). \tag{10}$$

Corollary 5 strictly improves the result or robustness analysis in Hardt and Price (2014); Xu and Li (2022). In particular, for general d, since $\eta = \mathcal{O}(\Delta_k)$ and $m = \Omega(\Delta_k^{-2})$, letting $n \to \infty$, we get $O((\log(1/\operatorname{tol}))^2/(\Delta_k^{-2}\operatorname{tol}))$ that dominates Eq. (10) when tol is small. This rate is superior to the sub-optimal gap dependency $1/(\Delta_1^{5/2}\operatorname{tol}^2)$ in Xu and Li (2022) and near to the optimal lower bound $1/(\Delta_1^2\operatorname{tol})$ in Jain et al. (2016). We further refine our analysis in the two-dimensional space where we can choose $\eta = \mathcal{O}(1)$ and match the typical bound of other stochastic variance-reduced gradient methods (Shamir, 2015; Xu and Gao, 2018) $\mathcal{O}\left(\left(n + 1/\Delta_k^2\right)\log\left(1/\operatorname{tol}\right)\right)$. Moreover, our proposed method does not require full gradients when n is large.

A promising technique is to combine the result in Xu and Li (2022) where they improved the dependency of the convergence rate over $O(\lambda_k - \lambda'_k)$ to dependency over $O(\sqrt{\lambda_k - \lambda_{k'}})$, where $k' > k$. It is still an open problem how the linearly gap-dependent sample complexity of the inner loop can be achieved. We also conjecture that $B = \Omega(\Delta_k^{-2})$ is necessary as the lower bound of sample complexity for Oja's algorithm is $\Omega((n\Delta_k^2)^{-1})$. It is also possible to extend our theorem to more general settings such as computing top-k generalized eigenvalues Xu and Li (2020, 2021b).

5. Experiments

In this section, we present numerical experiments on synthetic and real datasets. All the ground truth information, e.g., λ_1 and U_k, is obtained by SciPy's "eigh" function or Matlab's "eig" function for the purpose of benchmarking. We only consider the case that $k = 1$ and $U_1 = u_1$ become the leading eigenvector.

For synthetic datasets, we can control the eigengap in the following way. Let

$$D = \text{diag}\left(1, 1-\Delta, \ldots, 1-5\Delta, \frac{|g_1|}{d}, \frac{|g_2|}{d}, \ldots, \frac{|g_{d-5}|}{d}\right), \tag{11}$$

where g_i is drawn from a standard normal distribution for $i = 1, \ldots, d-5$ and $\Delta = 0.01, 0.001$. Following Ding et al. (2020), we sample $\{a_i\}_{i=1}^n$ from normal distribution $\mathcal{N}(0, A)$ and compute the sample covariance matrix as well as the first eigenvector, where $A, U \in \mathbb{R}^{d\times d}$ such that $A = UDU^\top$, U is a random $d\times d$ orthogonal matrix, $\mathcal{N}(0, A)$ is the multivariate normal distribution with zero mean and covariance A. We set $n = 10000, d = 20$.

For the real datasets, we performed experiments using the popular data of COLO100 and USPS. The USPS is the dataset of images of Handwritten Digits (Hull, 1994) with a size 9298×256. The COLI100 (Nene et al., 1996) is a dataset of gray-scale images of 100 objects with size 7200×1024. See Table 2 for more details

Table 2: The summary of real datasets.

dataset	n	d
Synthetic	10000	20
COLO100	7200	1024
USPS	9298	256

5.1. Algorithms

In this experiment, we only focus on *stochastic variance reduced methods for Riemannian optimization*. In particular, we compare the following methods:

1. iRSRG is our proposed method which can employ mini-batch gradients in outer loops and stated in Algrithm 1 .
2. RSRG (Kasai et al., 2018) is the Riemannian stochastic recursive gradient algorithm.
3. RSVRG (Zhang et al., 2016) is the Riemannian stochastic variance reduced gradient (SVRG) method.

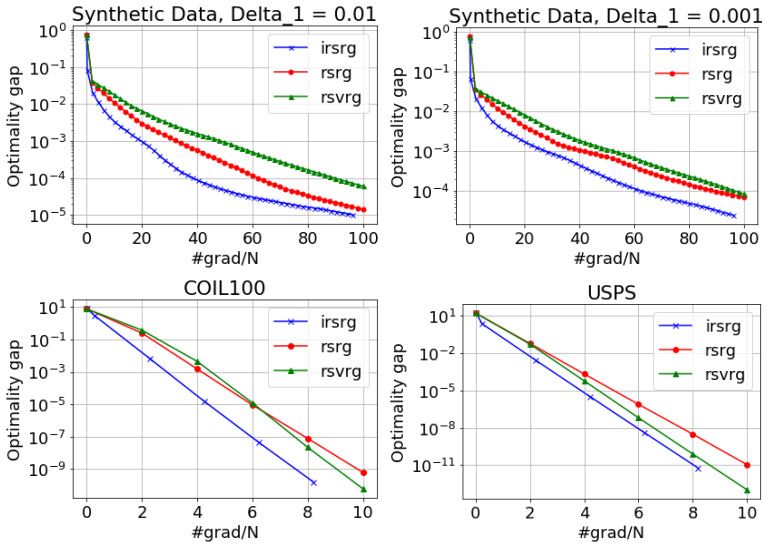

Figure 1: Experiments for Synthetic and real data. Each line is average by 100 experiments

RSRG and RSVRG are implemented in Matlab by Hiroyuki Kasai et. al. and their code can be found in https://github.com/hiroyuki-kasai/RSOpt.

5.2. Parameter setting

In what follows, we discuss how we choose parameters. Heuristically, since our algorithm converges linearly, we should increase B_s exponentially based on Theorem 4. In particular, we set $B_s = B10^s$ and B would be tuned carefully. Moreover, by Corollary 5, m and b should be of the same order, so we pick $m = \sqrt{n}$ and $b = \sqrt{n}$. The upper bound of the step size of the Riemannian gradient descent is $1/\lambda_1$. As a result, given a dataset $\{a_i \in \mathbb{R}^d\}_{i=1}^n$, we choose the step size to be $n/\sum_{i=1}^n \|a_i\|^2$, since $\frac{1}{n}\sum_{i=1}^n \|a_i\|^2 = \text{Trace}\left(\frac{1}{n}\sum_{i=1}^n a_i a_i^\top\right) \geq \lambda_1$. We fix the step size for all algorithms to get fair comparisons.

5.3. Results

In this study, all algorithms are initialized by w_0 such that $w_0 = \sqrt{r}u_1 + \sqrt{1-r}\frac{(I_d - u_1u_1^\top)\bar{w}}{\|(I_d - u_1u_1^\top)\bar{w}\|}$, where $\bar{w}$ is a unit vector uniformly chosen from the unit sphere and $r = 10^{-8}$. In other words, we fix the initial angle between w_0 and u, i.e. $u^\top w_0 = r$. We measure optimality gaps, i.e. $\lambda_1 - \widetilde{w}_s^\top A\widetilde{w}_s$, in term of sample complexity in this experiment. Note that *experiments are repeated 100 times and results are averaged* to eliminate randomness.

Figure 1 shows the results of our experiment. We can see that our inexact scheme indeed improves the sample complexity, and our proposed method, iRSRG outperforms other Riemannian stochastic variance reduced gradient methods, which also verifies our Corollary 5. This phenomenon may be due to extremely small r, and the warm start scheme is necessary. Moreover, the linear convergence rate is observed, which justifies Theorem 4.

6. Conclusion

In this paper, we provide a robustness analysis of the Riemannian gradient descent method for the eigenvector computation problem, which is the first result of noisy gradient methods achieving the same convergence rate as the power method in the two dimensional space. We leverage our novel framework of robustness analysis to study stochastic recursive gradient algorithm, which improves the gap-dependent sample complexity of robustness analysis and match the result from the state-of-the-art stochastic variance-reduced gradient methods. Our convergence result is global and works for arbitrary step sizes as long as the sample size is large enough when the step size is small. In particular, our proposed method can use mini-batch gradients in outer loops, hence it is possible to adapt to streaming data.

References

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

Zeyuan Allen-Zhu and Yuanzhi Li. LazySVD: Even faster SVD decomposition yet without agonizing pain. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 974–982, Barcelona, Spain, 2016.

Ehsan Amid and Manfred K. Warmuth. An implicit form of krasulina's k-pca update without the orthonormality constraint. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, pages 3179–3186, New York, NY, 2020.

Raman Arora, Andrew Cotter, Karen Livescu, and Nathan Srebro. Stochastic optimization for PCA and PLS. In *Proceedings of the 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 861–868, Allerton Park & Retreat Center, Monticello, IL, 2012.

Raman Arora, Andrew Cotter, and Nati Srebro. Stochastic optimization of PCA with capped MSG. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1815–1823, Lake Tahoe, NV, 2013.

Maria-Florina Balcan, Simon Shaolei Du, Yining Wang, and Adams Wei Yu. An improved gap-dependency analysis of the noisy power method. In *Proceedings of the 29th Conference on Learning Theory (COLT)*, pages 284–309, New York, NY, 2016.

Qinghua Ding, Kaiwen Zhou, and James Cheng. Tight convergence rate of gradient descent for eigenvalue computation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3276–3282, 2020, 2020.

Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. SPIDER: near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 687–697, Montréal, Canada, 2018.

Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU press, 2012.

Robert M Gower, Mark Schmidt, Francis Bach, and Peter Richtarik. Variance-reduced methods for machine learning. *Proceedings of the IEEE*, 108(11):1968–1983, 2020.

Moritz Hardt and Eric Price. The noisy power method: A meta algorithm with applications. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2861–2869, Montreal, Quebec, Canada, 2014.

Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 1994.

Prateek Jain, Chi Jin, Sham M. Kakade, Praneeth Netrapalli, and Aaron Sidford. Streaming PCA: matching matrix bernstein and near-optimal finite sample guarantees for Oja's algorithm. In *Proceedings of the 29th Conference on Learning Theory (COLT)*, pages 1147–1164, New York, NY, 2016.

Bo Jiang, Shiqian Ma, Anthony Man-Cho So, and Shuzhong Zhang. Vector transport-free svrg with general retraction for riemannian optimization: Complexity analysis and practical implementation. *arXiv preprint arXiv:1705.09059*, 2017.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems (NIPS)*, pages 315–323, Lake Tahoe, NV, 2013.

Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

Hiroyuki Kasai, Hiroyuki Sato, and Bamdev Mishra. Riemannian stochastic recursive gradient algorithm with retraction and vector transport and its convergence analysis. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 2521–2529, Stockholmsmässan, Stockholm, Sweden, 2018.

TP Krasulina. The method of stochastic approximation for the determination of the least eigenvalue of a symmetrical matrix. *USSR Computational Mathematics and Mathematical Physics*, 9(6):189–195, 1969.

Lihua Lei and Michael I. Jordan. Less than a single pass: Stochastically controlled stochastic gradient. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 148–156, Fort Lauderdale, FL, 2017.

Chris Junchi Li, Mengdi Wang, Han Liu, and Tong Zhang. Near-optimal stochastic approximation for online principal component estimation. *Math. Program.*, 167(1):75–97, 2018.

Sameer A Nene, Shree K Nayar, and Hiroshi Murase. Columbia object image library (coil-20). *Technical Report CUCS-005-96*, 1996.

Lam M. Nguyen, Jie Liu, Katya Scheinberg, and Martin Takác. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 2613–2621, Sydney, Australia, 2017a.

Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Stochastic recursive gradient algorithm for nonconvex optimization. *arXiv preprint arXiv:1705.07261*, 2017b.

Lam M. Nguyen, Katya Scheinberg, and Martin Takác. Inexact SARAH algorithm for stochastic optimization. *Optim. Methods Softw.*, 36(1):237–258, 2021.

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

Erkki Oja and Juha Karhunen. On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of Mathematical Analysis and Applications*, 106(1):69–84, 1985.

Ohad Shamir. A stochastic PCA and SVD algorithm with an exponential convergence rate. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 144–152, Lille, France, 2015.

Ohad Shamir. Fast stochastic algorithms for SVD and PCA: convergence properties and convexity. In *Proceedings of the 33nd International Conference on Machine Learning (ICML)*, pages 248–256, New York City, NY, 2016.

Vincent Q Vu, Jing Lei, et al. Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, 41(6):2905–2947, 2013.

Zhiqiang Xu and Xin Gao. On truly block eigensolvers via riemannian optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 168–177, 2018.

Zhiqiang Xu and Ping Li. A practical riemannian algorithm for computing dominant generalized eigenspace. In *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 819–828, virtual online, 2020.

Zhiqiang Xu and Ping Li. A comprehensively tight analysis of gradient descent for PCA. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 21935–21946, virtual, 2021a.

Zhiqiang Xu and Ping Li. On the riemannian search for eigenvector computation. *J. Mach. Learn. Res.*, 22:249:1–249:46, 2021b.

Zhiqiang Xu and Ping Li. Faster noisy power method. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, pages 1138–1164, Paris, France, 2022.

Hongyi Zhang, Sashank J. Reddi, and Suvrit Sra. Riemannian SVRG: fast stochastic optimization on riemannian manifolds. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4592–4600, Barcelona, Spain, 2016.