Stochastic Variance-Reduced Algorithms for PCA with Arbitrary Mini-Batch Sizes

Cheolmin Kim
Northwestern University

Diego Klabjan
Northwestern University

Abstract

We present two stochastic variance-reduced algorithms for PCA and provide their convergence analyses. By deriving explicit forms of step size, epoch length and batch size, we show that the proposed algorithms can attain the optimal runtime with arbitrary batch sizes. We also establish global convergence of the algorithms based on a novel analysis, which studies the optimality gap as a ratio of two expectation terms. The framework in our analysis is general and can be applied to analyze other stochastic variance-reduced PCA algorithms and improve their analyses. Moreover, we introduce practical implementations of the algorithms which require no hyper-parameters. The experimental results show that the proposed algorithms outperform other stochastic variance-reduced PCA algorithms regardless of the batch size.

1 Introduction

Principal component analysis (PCA) (Jolliffe, 2011) is a fundamental tool for dimensionality reduction in machine learning and statistics. Given a data matrix $A = [a_1 a_2 \dots a_n] \in \mathbb{R}^{d \times n}$ consisting of n data vectors $a_1, a_2, \dots, a_n$ in $\mathbb{R}^d$, PCA finds a direction w onto which the projections of the data vectors have the largest variance. Assuming that the data vectors are standardized with a mean of zero and standard deviation of one, the PCA problem can be formulated as

$$\begin{aligned} \text{maximize} \quad & f(w) = \frac{1}{2n}\sum_{i=1}^{n}(a_i^T w)^2 = \frac{1}{2}w^T C w \\ \text{subject to} \quad & \|w\|_2 = 1 \end{aligned} \tag{1}$$

Proceedings of the 23[rd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

where $C = \frac{1}{n}AA^T \in \mathbb{R}^{d \times d}$ is the covariance matrix of data matrix A. As the largest eigenvector u_1 of C maximizes $f(w)$, (1) can be solved by computing the singular value decomposition (SVD) of A. However, the runtime of SVD is $\mathcal{O}(\min\{nd^2, n^2 d\})$, which can be expensive in a large-scale setting. An alternative way to solve (1) is to use power iteration (Golub and Van Loan, 2012) which repeatedly applies $w_{t+1} = Cw_t/\|Cw_t\|$ at each iteration. The sequence of iterates $\{w_t\}$ generated by power iteration is guaranteed to obtain an ϵ-optimal solution after $\mathcal{O}\big(\frac{1}{\Delta}\log\frac{1}{\epsilon}\big)$ iterations where $\lambda_1 > \lambda_2 \geq \dots \geq \lambda_d \geq 0$ are the eigenvalues of C and $\Delta = 1 - \lambda_2/\lambda_1$ represents the eigen-gap. Since each iteration involves multiplying vector w_t with the matrix C, the runtime becomes $\mathcal{O}\big(nd\frac{1}{\Delta}\log\frac{1}{\epsilon}\big)$. When n and d are both large, the runtime of power iteration is better than that of SVD. Nonetheless, it still largely depends on n and can be prohibitive when Δ is small.

In order to reduce the dependence on Δ or n, the following variants of power iteration have been developed. To reduce the dependence on Δ, Xu et al. (2018) propose power iteration with momentum (Power+M) utilizing the momentum idea of Polyak (1964). With the optimal choice of the momentum parameter $\beta = \lambda_2^2/4$, the total runtime improves to $\mathcal{O}\big(nd\frac{1}{\sqrt{\Delta}}\log\frac{1}{\epsilon}\big)$. Also, a stochastic algorithm utilizing the stochastic gradient $a_{i_t}a_{i_t}^T w_t$ rather than a full gradient Cw_t is introduced in Oja (1982). Since it requires just one data vector at a time, the computational cost per iteration is significantly reduced. However, due to the variance of stochastic gradients, a sequence of diminishing step sizes needs to be adopted, making its progress slow near the optimum.

Built on the recent stochastic variance-reduced gradient (SVRG) technique (Johnson and Zhang, 2013), Shamir (2015, 2016) present a stochastic variance-reduced version of Oja's algorithm (VR-PCA) and its extension to find $k \geq 1$ principal components. Utilizing stochastic variance-reduced gradients, VR-PCA works with a constant step size and converges at an exponential rate, reducing the total runtime to $\mathcal{O}(d(n + \frac{1}{\Delta^2})\log\frac{1}{\epsilon})$. The analysis of VR-PCA considers a mini-batch of size one,

Table 1: Comparison of stochastic variance-reduced methods for PCA and their convergence analyses. Types of convergence and complexity results are summarized. "Local" means that there is a restriction on the angle between an initial iterate and the first eigenvector u_1 and "global" implies no such restriction. For VR Power and VR HB Power, $\mu \geq 0$ is a parameter which controls the progress of the algorithms through step size $\eta = \Delta^\mu$ depending on batch size.

Algorithm	Convergence	Iteration	Batch Size	Total Runtime	Reference
VR-PCA	Local	$\mathcal{O}\left(\frac{1}{\Delta^2}\log\frac{1}{\epsilon}\right)$	$\mathcal{O}(1)$	$\mathcal{O}\left(d\left(n+\frac{1}{\Delta^2}\right)\log\frac{1}{\epsilon}\right)$	(Shamir, 2015)
VR Power+M	Local	$\mathcal{O}\left(\frac{1}{\Delta^{1/2}}\log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{\sqrt{d}}{\Delta^{3/2}}\right)$	$\mathcal{O}\left(d\left(n+\frac{\sqrt{d}}{\Delta^2}\right)\log\frac{1}{\epsilon}\right)$	(Xu et al., 2018)
Fast PCA	Global	$\mathcal{O}\left(\frac{1}{\Delta^2}\text{poly}\left(\log\frac{1}{\epsilon}\right)\right)$	$\mathcal{O}(1)$	$\mathcal{O}\left(d\left(n+\frac{1}{\Delta^2}\right)\text{poly}\left(\log\frac{1}{\epsilon}\right)\right)$	(Garber and Hazan, 2015)
VR Power	**Global**	$\mathcal{O}\left(\frac{1}{\Delta^{1+\mu}}\log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{1}{\Delta^{1-\mu}}\right)$	$\mathcal{O}\left(d\left(n+\frac{1}{\Delta^2}\right)\log\frac{1}{\epsilon}\right)$	**[This Paper]**
VR HB Power	**Global**	$\mathcal{O}\left(\frac{1}{\Delta^{1/2+\mu}}\log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{1}{\Delta^{3/2-\mu}}\right)$	$\mathcal{O}\left(d\left(n+\frac{1}{\Delta^2}\right)\log\frac{1}{\epsilon}\right)$	**[This Paper]**

which implies that it works with any size of mini-batch. However, conditions for the step size and the epoch size are not precisely given, making it hard to attain the theoretically optimal runtime in practice.

A stochastic variance-reduced version of Power+M (VR Power+M) is introduced by Xu et al. (2018). Due to the momentum term, the iteration complexity is improved to $\mathcal{O}\left(\frac{1}{\Delta^{1/2}}\log\frac{1}{\epsilon}\right)$. However, a batch size of $\mathcal{O}\left(\frac{\sqrt{d}}{\Delta^{3/2}}\right)$ is required to achieve such iteration complexity, leading to the total runtime of $\mathcal{O}\left(d\left(n+\frac{\sqrt{d}}{\Delta^2}\right)\log\frac{1}{\epsilon}\right)$. Note that the runtime of VR Power+M is worse than that of VR-PCA due to the extra dependency on $\sqrt{d}$. Moreover, unless the batch size is sufficiently large, VR Power+M may diverge, which makes it hard to use.

On other other hand, Garber and Hazan (2015) reduce the PCA problem to inexactly solving a sequence of convex optimization problems. Each convex optimization problem has the form of the least square problem and amounts to one step of inverse power iteration (Golub and Van Loan, 2012). Due to the finite sum structure of the objective function, the SVRG algorithm (Johnson and Zhang, 2013) can be used to solve it. However, solving this strongly convex optimization problem can be as hard as the original PCA problem since the objective function is $(\lambda_1 - \lambda_2)$-stronly convex and $(2\lambda_1 - \lambda_2 - \lambda_d)$-smooth in the accurate regime. By inexactly solving these problems, an ϵ-optimal solution can be obtained after a poly-logarithmic number of iterations.

The shifted-and-inverted approach is also introduced for the leading eigenvector problem (Garber et al., 2016) and a number of solvers such as coordinate-descent (Wang et al., 2018), SVRG (Garber et al., 2016), accelerated gradient descent, accelerated SVRG (Allen-Zhu and Li, 2016) and Riemannian gradient descent (Xu, 2018) have been developed to solve the least square problem. Other works on power iteration include the noisy (Hardt and Price, 2014) and coordinate-wise (Lei et al., 2016) power methods. The noisy power method considers the power method in a noise setting, which Balcan et al. (2016) extend to provide an improved gap-dependency analysis. Also, power iteration has been analyzed for incremental or online PCA in many works (Allen-Zhu and Li, 2017; Li et al., 2018; Balsubramani et al., 2013; Arora et al., 2012; Boutsidis et al., 2015; Jain et al., 2016; Mitliagkas et al., 2013).

In this paper, we present two mini-batch stochastic variance-reduced algorithms for PCA (VR Power, VR HB Power) and their convergence analyses. They are mini-batch versions of stochastic variance-reduced algorithms for power (Golub and Van Loan, 2012) and power with momentum (Xu et al., 2018) iteration methods. While VR-PCA (Shamir, 2015) takes a data vector at a time, VR Power works with any batch size and the accompanying analysis reveals that whatever the batch size is, VR Power can always achieve the optimal runtime by appropriately choosing the step size and epoch length. Explicit conditions for the step size, the epoch length and the batch size to ensure the optimal runtime are derived for VR Power. On the other hand, VR HB Power is an enhanced algorithm of VR Power+M. By adding the step size, VR HB Power can work with any batch size while VR Power+M can fail if the batch size is not sufficiently large. In the analysis of VR HB Power, we prove that for any batch size, VR HB Power can achieve the optimal runtime by appropriately choosing the step size, the epoch length and the momentum parameter. Explicit expressions for theses parameters are provided. In addition, our analysis removes the dependency on $\sqrt{d}$ for the batch size, which improves the analysis of VR Power+M. For the comparison of stochastic variance-reduced PCA algorithms and their convergence analyses see Table 1.

In the convergence analyses, we introduce a novel framework of analyzing stochastic variance-reduced PCA algorithms. For an inner-loop iterate w_t, we decompose $E[(u_k^T w_t)^2]$ with u_k an eigenvector with respect to λ_k into two parts where the first one is the expectation term and the second one is the variance term. To obtain tight bounds for the variance term, we analyze its growth over an epoch rather than focusing on iteration-by-iteration behavior. Using the Binomial expansion of matrices, we come up with compact bounds of the variance term. Based on the compact representation of the variance term, we establish a bound for $(E[\|w_t\|^2] - E[(u_1^T w_t)^2])/E[(u_1^T w_t)^2] = \sum_{k=2}^{d} E[(u_k^T w_t)^2]/E[(u_1^T w_t)^2]$ and derive conditions for the step size, epoch length and batch size to ensure its sufficient decrease.

The concept of representing the optimality gap as the ratio of two expectations has been never used for analyzing stochastic PCA algorithms. However, it results in much simpler convergence statements than probabilistic statements appearing in Shamir (2015) and Xu et al. (2018). Note that probabilistic statements can be easily derived from expectation bounds using the Chebyshev inequality. With the expectation bounds, we can establish global convergence of stochastic PCA algorithms. Although stochastic PCA algorithms have been observed to work well with random initialization (Shamir, 2015), an initial condition of $|u_1^T \tilde{w}_0| \geq 1/2$ is required in previous probabilistic analyses. In our framework, such condition is not necessary and the rate of convergence does not depend on how far an iterate is from u_1 but is kept the same across iterations, as in the case of deterministic power iteration. The framework introduced in this work is not specific to the presented algorithms; it can be easily applied to analyze other stochastic variance-reduced PCA algorithms such as VR-PCA or VR Power+M, deriving expectation bounds for them and resolving their initialization issues.

Our work has the following contributions.

1. We present two mini-batch stochastic variance-reduced PCA algorithms. For any batch size, our algorithms achieve the optimal runtime by appropriately choosing algorithm parameters. Explicit expressions for these parameters are provided.

2. We provide novel convergence analyses for the algorithms where we establish global convergence by deriving a bound for the ratio of two expectation terms. The framework in our convergence analyses is general, therefore can be used to analyze other stochastic variance-reduced PCA algorithms. To this end, we are the first to establish convergence of VR-PCA and VR Power+M for any initial vector and in expectation.

3. We introduce practical implementations of the algorithms and report numerical experiments on diverse datasets. Experimental results show that our algorithms outperform other stochastic variance-reduced algorithms for any batch size.

The paper is organized as follows. We introduce the algorithms in Section 2 and the convergence analyses in Section 3. Some practical considerations regarding the implementations of the algorithms are discussed in Section 4 and the experimental results are followed in Section 5.

2 Stochastic Variance-Reduced Algorithms for PCA

We consider two mini-batch stochastic variance-reduced algorithms for PCA. The first one is a mini-batch version of VR-PCA (Shamir, 2015) and the second one is an enhanced version of VR Power+M (Xu et al., 2018) with a step size incorporated. For eigenpairs (λ_k, u_k) of $C = \frac{1}{n}\sum_{i=1}^{n} a_i a_i^T$, we assume that the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_d$ satisfy $\lambda_1 > \lambda_2 \geq \ldots \geq \lambda_d \geq 0$ and the eigenvectors $u_1, u_2, \ldots, u_d$ form an orthonormal basis. Since a symmetric matrix is orthogonally diagonalizable, we can assume that such eigenvectors exist without loss of generality. We assume that all norms are L_2 for vectors and spectral for matrices.

Variance reduction algorithms have an outer loop and an inner loop. They periodically compute exact gradients at each outer iteration and use it in inner iterations to reduce the variance of stochastic gradients. Let $\tilde{w}_s$ and w_t denote an outer-loop and inner-loop iterate, respectively. To get a stochastic variance-reduced gradient of an inner loop iterate w_t, we first decompose the inner loop iterate w_t it into two parts as

$$w_t = \frac{(\tilde{w}_s^T w_t)}{\|\tilde{w}_s\|^2}\tilde{w}_s + \left(I - \frac{\tilde{w}_s \tilde{w}_s^T}{\|\tilde{w}_s\|^2}\right) w_t$$

using the outer loop iterate $\tilde{w}_s$. In the above decomposition, the former term represents the projection of w_t on $\tilde{w}_s$ while the latter term represents the remaining vector. Utilizing the exact gradient $\tilde{g}_s$ at $\tilde{w}_s$, the exact gradient at the first term can be computed as

$$\nabla f\left(\frac{(\tilde{w}_s^T w_t)}{\|\tilde{w}_s\|^2}\tilde{w}_s\right) = \frac{(\tilde{w}_s^T w_t)}{\|\tilde{w}_s\|^2} C\tilde{w}_s = \frac{(\tilde{w}_s^T w_t)}{\|\tilde{w}_s\|^2}\tilde{g}_s.$$

On the other hand, a stochastic sample S_t is used to compute a stochastic gradient at the second term as

$$\frac{1}{|S_t|}\sum_{l \in S_t} a_l a_l^T \left(I - \frac{\tilde{w}_s \tilde{w}_s^T}{\|\tilde{w}_s\|^2}\right) w_t.$$

This results in the following stochastic variance-reduced gradient g_t at w_t as

$$g_t = \frac{(\tilde{w}_s^T w_t)}{\|\tilde{w}_s\|^2}\tilde{g}_s + \frac{1}{|S_t|}\sum_{l \in S_t} a_l a_l^T \left(I - \frac{\tilde{w}_s \tilde{w}_s^T}{\|\tilde{w}_s\|^2}\right) w_t. \quad (2)$$

2.1 VR Power

Using the stochastic variance-reduced gradient g_t, we obtain a stochastic variance reduced version of Power iteration as

$$w_{t+1} \leftarrow (1-\eta)w_t + \eta g_t. \tag{3}$$

This update rule has a similar form as the one in VR-PCA, which repeats

$$w_{t+1} \leftarrow w_t + \bar{\eta}\left(a_{i_t}(a_{i_t}^T w_t - a_{i_t}^T \tilde{w}_s) + \tilde{g}_s\right). \tag{4}$$

Note that (3) generalizes (4) in the following two senses. First, we can obtain an update rule of (4) by letting $\eta = (1+\bar{\eta})/\bar{\eta}$ in (3). Second, with the choice of $\eta = 1$, we can recover deterministic power iteration from (3) while (4) does not. Using update rule (3), we have VR Power exhibited in Algorithm 1.

Algorithm 1 VR Power

Parameters: step size η, mini-batch size $|S|$, epoch length m
Input: data vectors $a_i, i = 1, \dots, n$
randomly initialize outer iterate $\tilde{w}_0$
for $s = 0, 1, \dots$ **do**
  $\tilde{g} \leftarrow C\tilde{w}_s$
  $w_0 \leftarrow \tilde{w}_s$
  $w_1 \leftarrow (1-\eta)w_0 + \eta\tilde{g}$
  for $t = 1, 2, \dots, m-1$ **do**
    sample a mini-batch $S_t \subset \{1, \cdots, n\}$ of size $|S|$ uniformly at random
    $g_t \leftarrow \frac{1}{|S_t|}\sum_{l\in S_t} a_l a_l^T \left(I - \frac{w_0 w_0^T}{\|w_0\|^2}\right) w_t + \frac{(w_t^T w_0)}{\|w_0\|^2}\tilde{g}$
    $w_{t+1} \leftarrow (1-\eta)w_t + \eta g_t$
  end for
  $\tilde{w}_{s+1} \leftarrow w_m$
end for

When per sample cost is as expensive as per iteration cost, VR Power is an efficient algorithm since it attains the optimal sample complexity. However, if per sample cost is cheap, it might not be effective since its iteration complexity does not improve beyond $\mathcal{O}(\frac{1}{\Delta}\log(\frac{1}{\epsilon}))$. For this reason, we introduce VR HB Power which works better in the latter setting.

2.2 VR HB Power

Using g_t, we obtain a stochastic variance-reduced heavy ball power iteration as

$$w_{t+1} \leftarrow 2\big((1-\eta)w_t + \eta g_t\big) - \beta w_{t-1} \tag{5}$$

where $\eta \in (0, 1]$ is the step size and β is the momentum parameter. Note that we can recover the deterministic heavy ball power iteration from (5) when the step size η is set to 1 and the exact gradient $g_t = Cw_t$ is used. The mechanism of controlling the progress of the algorithm using the step size η is not present in VR Power+M (Xu et al., 2018). As a result, it fails to converge unless the mini-batch size $|S|$ is sufficiently large. To the contrary, our algorithm works with any mini-batch size $|S|$ due to the presence of the step size η. By selecting an appropriate value of η depending on the size of $|S|$ and m, we can always ensure that the variance terms do not grow faster than expectation terms. Having update rule (5), VR HB Power is described in Algorithm 2.

Algorithm 2 VR HB Power

Parameters: step size η, momentum β, mini-batch size $|S|$, epoch length m
Input: data vectors $a_i, i = 1, \dots, n$
randomly initialize outer iterate $\tilde{w}_0$
for $s = 0, 1, \dots$ **do**
  $\tilde{g} \leftarrow C\tilde{w}_s$
  $w_0 \leftarrow \tilde{w}_s$
  $w_1 \leftarrow (1-\eta)w_0 + \eta\tilde{g}$
  for $t = 1, 2, \dots, m-1$ **do**
    sample a mini-batch $S_t \subset \{1, \cdots, n\}$ of size $|S|$ uniformly at random
    $g_t \leftarrow \frac{1}{|S_t|}\sum_{l\in S_t} a_l a_l^T \left(I - \frac{w_0 w_0^T}{\|w_0\|^2}\right) w_t + \frac{(w_t^T w_0)}{\|w_0\|^2}\tilde{g}$
    $w_{t+1} \leftarrow 2\big((1-\eta)w_t + \eta g_t\big) - \beta w_{t-1}$
  end for
  $\tilde{w}_{s+1} \leftarrow w_m$
end for

3 Convergence Analyses

In this section, we provide convergence analyses for VR Power and VR HB Power. Before presenting the convergence analyses, we first introduce some notation.

3.1 Notation

Let C_t and P be the sample covariance matrix at inner iteration t and the projection matrix to the space orthogonal to the outer iterate $w_0 = \tilde{w}_s$ as

$$C_t = \frac{1}{|S_t|}\sum_{l\in S_t} a_l a_l^T, \quad P = I - \frac{w_0 w_0^T}{\|w_0\|^2}. \tag{6}$$

Using (6), we can write g_t as $g_t = \eta C w_t + \eta(C_t - C)Pw_t$. Next, we characterize the variance of sample covariance matrix C_t as

$$K = E[\|(C_t - C)^2\|], \quad \sigma^2 = E[\|a_{i_t}a_{i_t}^T - C\|^2].$$

Then, for $M_k = E[(C_t - C)u_k u_k^T(C_t - C)]$, we have

$$\|M_k\| \le K = \frac{\sigma^2}{|S|}. \tag{7}$$

For the analysis of VR HB Power, we define

$$\alpha_k(\eta) = 4(1-\eta+\eta\lambda_k)^2,\ \beta(\eta) = (1-\eta+\eta\lambda_2)^2. \quad (8)$$

Also, we let $p_t(\alpha,\beta)$ and $q_t(\alpha,\beta)$ be the Chebyshev polynomials of the first and the second kind (Mason and Handscomb, 2002) respectively such that

$$p_t(\alpha,\beta) = (\alpha-\beta)p_{t-1}(\alpha,\beta) - \beta(\alpha-\beta)p_{t-2}(\alpha,\beta) + \beta^3 p_{t-3}(\alpha,\beta), \quad (9)$$

$$q_t(\alpha,\beta) = (\alpha-\beta)q_{t-1}(\alpha,\beta) - \beta(\alpha-\beta)q_{t-2}(\alpha,\beta) + \beta^3 q_{t-3}(\alpha,\beta) \quad (10)$$

for $t \geq 3$ and

$$p_0(\alpha,\beta) = 1,\ p_1(\alpha,\beta) = \frac{\alpha}{4},\ p_2(\alpha,\beta) = \left(\frac{\alpha}{2}-\beta\right)^2, \quad (11)$$

$$q_0(\alpha,\beta) = 1,\ q_1(\alpha,\beta) = \alpha,\ q_2(\alpha,\beta) = (\alpha-\beta)^2. \quad (12)$$

Since the first eigenvector u_1 of the covariance matrix C is an optimal solution to (1), the optimality gap is measured as $\sum_{k=2}^d (u_k^T w_t)^2/(u_1^T w_t)^2$, representing how closely w_t is aligned with u_1. Note that this ratio is zero if $w_t = u_1$. Our analysis studies it in expectation, providing a bound for $\theta_t = \sum_{k=2}^d E[(u_k^T w_t)^2]/E[(u_1^T w_t)^2]$ given fixed s and $\tilde{\theta}_s = \sum_{k=2}^d E[(u_k^T \tilde{w}_s)^2]/E[(u_1^T \tilde{w}_s)^2]$ for an inner loop iterate w_t and an outer loop iterate $\tilde{w}_s$, respectively.

3.2 VR Power

In Lemmas 3.1, 3.2 and 3.3, we consider a single epoch, which corresponds to one inner loop iteration starting with w_0.

Lemma 3.1. *For any $\eta \in (0,1]$, $1 \leq k \leq d$ and $1 \leq t \leq m$, we have*

$$E[(u_k^T w_t)^2] = (1-\eta+\eta\lambda_k)^{2t} E[(u_k^T w_0)^2] + \eta^2 \sum_{i=1}^{t-1} (1-\eta+\eta\lambda_k)^{2(t-i-1)} E[w_i^T P M_k P w_i].$$

Lemma 3.1 decomposes $E[(u_k^T w_t)^2]$ into two parts. The first part represents the expectation term which grows at a rate of $(1-\eta+\eta\lambda_k)^2$ and the second part is the variance term which increases as w_t strides away from w_0 as captured by $E[w_t^T P M_k P w_t]$.

Lemma 3.2. *For any $\eta \in (0,1]$, $1 \leq k \leq d$ and $1 \leq t \leq m$, we have*

$$\sum_{k=2}^d E[w_t^T P M_k P w_t] \leq 2K \cdot \sum_{k=2}^d E[(u_k^T w_0)^2] \cdot \left((1-\eta+\eta\lambda_1)^2 + \eta^2 K\right)^t.$$

Moreover, if $0 < \dfrac{\eta^2 K m}{(1-\eta+\eta\lambda_1)^2} < 1$, then we have

$$\theta_m \leq \left[\left(\frac{1-\eta+\eta\lambda_2}{1-\eta+\eta\lambda_1}\right)^{2m} + \frac{4\eta^2 K m}{(1-\eta+\eta\lambda_1)^2}\right] \cdot \theta_0.$$

Lemma 3.2 provides a bound for $\sum_{k=2}^d E[w_t^T P M_k P w_t]$, which grows at a rate not greater than $(1-\eta+\eta\lambda_1)^2 + \eta^2 K$. Using this bound and assuming some condition on η, K, and m, a bound on θ_m is derived as a function of θ_0, η, m, and K. In Lemma 3.3, we present explicit conditions for η, m, and $|S|$ to ensure a sufficient decrease of θ_m.

Lemma 3.3. *Let $\eta = \Delta^\mu$ for some $\mu \geq 0$. If m and $|S|$ satisfy*

$$m = \left\lceil \frac{(1-\eta+\eta\lambda_1)\log 2}{2\eta\lambda_1\Delta} \right\rceil \quad (13)$$

and

$$|S| \geq \frac{16\eta^2\sigma^2 m}{(1-\eta+\eta\lambda_1)^2}, \quad (14)$$

then we have $\theta_m \leq 3/4 \cdot \theta_0$.

For any $\mu \geq 0$ such that $\eta = \Delta^\mu$, Lemma 3.3 provides explicit values of m and $|S|$ to ensure a sufficient decrease of θ_m. In the analysis of VR-PCA, exact values of η and m to ensure the optimal runtime have not been provided. Instead, only the orders of η and m have been provided such that $\eta = c_1\Delta$ and $m = c_2/\Delta^2$, making it hard to obtain the optimal runtime in practice. Contrary to it, our analysis provides explicit expressions for m and $|S|$, being more practical. Moreover, since the term on the right-hand side of (14) goes to zero as μ increases, it can be also stated that for any $|S| \geq 1$, there exists some $\mu \geq 0$ and thus $\eta = \Delta^\mu$ and m (see (14)) such that $\theta_m \leq 3/4 \cdot \theta_0$ holds. This implies that VR Power can always attain a sufficient decrease of θ_m no matter what $|S|$ is used. We next give the main result.

Theorem 3.4. *Suppose that an initial vector $\tilde{w}_0$ satisfies $u_1^T \tilde{w}_0 \neq 0$ and let $\tilde{\theta}_0 = (1-(u_1^T\tilde{w}_0)^2)/(u_1^T\tilde{w}_0)^2 \geq \epsilon$ for some $\epsilon > 0$. If $\eta = \Delta^\mu$ and m and $|S|$ satisfy (13) and (14), after $\tau = \lceil \log(\tilde{\theta}_0/\epsilon)/\log(4/3) \rceil$ epochs of VR Power, we have $\tilde{\theta}_\tau \leq \epsilon$.*

Theorem 3.4 present a convergence result for τ epochs. Note that our result requires only a trivial assumption on $\tilde{\theta}_0$ and thus establishes global convergence. Also, since $\tau = \mathcal{O}(\log(\frac{1}{\epsilon}))$, only a logarithmic number of inner loops is needed to be performed to obtain ϵ-accuracy.

3.3 VR HB Power

The following Lemmas 3.5, 3.6 and 3.7 are counterparts of Lemmas 3.1, 3.2 and 3.3 for VR HB Power. For the momentum parameter β, we let $\beta = \beta(\eta)$ which is defined in (8). As in the analysis of VR Power, we first consider a single epoch with an initial inner loop iterate w_0.

Lemma 3.5. *For any $\eta \in (0, 1]$, $1 \le k \le d$ and $1 \le t \le m$, we have*

$$E[(u_k^T w_t)^2] = p_t(\alpha_k(\eta), \beta(\eta))E[(u_k^T w_0)^2] + 4\eta^2 \sum_{r=1}^{t-1} q_{t-r-1}(\alpha_k(\eta), \beta(\eta))E[w_r^T PM_kPw_r].$$

Lemma 3.5 breaks $E[(u_k^T w_t)^2]$ into the sum of expectation part and variance part. While the expectation term is a function of the Chebyshev polynomial of the first kind, the variance part is a function of the Chebyshev polynomials of the second kind. That being said, the variance term grows faster and thus we need a careful analysis for it.

Lemma 3.6. *For any $\eta \in (0, 1]$, $1 \le k \le d$, and $1 \le t \le m$, we have*

$$\sum_{k=2}^{d} E[w_t^T PM_kPw_t] \le 4K \cdot \left(1 + \frac{4\eta^2 K}{\alpha_1(\eta) - 4\beta(\eta)}\right)^{t-1} \cdot \left(\frac{\sqrt{\alpha_1(\eta)}}{2} + \frac{\sqrt{\alpha_1(\eta) - 4\beta(\eta)}}{2}\right)^{2t} \cdot \sum_{k=2}^{d} E[(u_k^T w_0)^2].$$

Moreover, if $0 < \dfrac{4\eta^2 Km}{\alpha_1(\eta) - 4\beta(\eta)} < 1$, *then we have*

$$\theta_m \le \left(\frac{p_m(\alpha_2(\eta), \beta(\eta))}{p_m(\alpha_1(\eta), \beta(\eta))} + \frac{128\eta^2 Km}{\alpha_1(\eta) - 4\beta(\eta)}\right) \cdot \theta_0.$$

Lemma 3.6 provides a bound for $\sum_{k=2}^{d} E[w_t^T PM_kPw_t]$. Note that it depends on Δ and blows up as Δ goes to zero due to the term involving $1/(\alpha_1(\eta) - 4\beta(\eta))$. Due to this dependency, VR HB Power tends to require a larger batch size than VR Power given the same values of η and m. Lemma 3.6 also establishes a bound for θ_m as a function of θ_0, η, m and K under some assumption.

Lemma 3.7. *For some $\mu \ge 0$, let $\eta = \Delta^\mu$ and*

$$m = \left\lceil \left(\frac{1 - \eta + \eta\lambda_1}{\eta\lambda_1\Delta + \sqrt{\eta\lambda_1\Delta(2(1-\eta) + \eta(\lambda_1 + \lambda_2))}} + \frac{\sqrt{\eta\lambda_1\Delta(2(1-\eta) + \eta(\lambda_1 + \lambda_2))}}{\eta\lambda_1\Delta + \sqrt{\eta\lambda_1\Delta(2(1-\eta) + \eta(\lambda_1 + \lambda_2))}} \right) \frac{\log 8}{2} \right\rceil \tag{15}$$

and

$$|S| \ge \frac{128\eta\sigma^2 m}{\lambda_1\Delta\left[2(1-\eta) + \eta(\lambda_1 + \lambda_2)\right]}. \tag{16}$$

Then, we have $\theta_m \le 3/4 \cdot \theta_0$.

Lemma 3.7 provides explicit conditions for m and $|S|$ to ensure a sufficient decrease of θ_m. Note that when $\mu = 0$, we have $|S| \ge \mathcal{O}(\frac{1}{\Delta^{3/2}})$, which improves the analysis of VR Power+M in Xu et al. (2018) by removing the dependency on $\sqrt{d}$. Also, for any $|S| \ge 1$, there exists some η and m satisfying the conditions in Lemma 3.7. This implies that VR HB Power works with any batch size while VR Power+M does not. The overall convergence is established next.

Theorem 3.8. *Suppose that an initial vector $\tilde{w}_0$ satisfies $u_1^T \tilde{w}_0 \neq 0$ and let $\tilde{\theta}_0 = (1 - (u_1^T\tilde{w}_0)^2)/(u_1^T\tilde{w}_0)^2 \ge \epsilon$ for some $\epsilon > 0$. If $\eta = \Delta^\mu$ and m and $|S|$ satisfy* (15) *and* (16)*, after $\tau = \lceil \log(\tilde{\theta}_0/\epsilon)/\log(4/3) \rceil$ epochs of VR HB Power, we have $\tilde{\theta}_\tau \le \epsilon$.*

The global convergence result in Theorem 3.8 is based on the single epoch result in Lemma 3.7. Since $\tau = \mathcal{O}(\log(\frac{1}{\epsilon}))$, the iteration complexity of VR HB Power is $\tau m = \mathcal{O}(\frac{1}{\Delta^{1/2+\mu/2}} \log(\frac{1}{\epsilon}))$. On the other hand, from $|S| = \mathcal{O}(\frac{1}{\Delta^{3/2-\mu/2}})$, the sample complexity amounts to $\mathcal{O}((n + \frac{1}{\Delta^2}) \log(\frac{1}{\epsilon}))$. Note that VR HB Power has the same sample complexity as VR Power but may have small iteration complexity. Therefore, if per sample cost is cheaper than per iteration cost, VR HB Power can be more efficient than VR Power.

4 Practical Considerations

In this section, we discuss some practical aspects implementing the proposed algorithms. First, to ensure that the algorithms are numerically stable, we consider normalizations as introduced in Shamir (2015) and Xu et al. (2018). After updating w_{t+1}, we normalize w_{t+1} as $w_{t+1} \leftarrow w_{t+1}/\|w_{t+1}\|_2$ in VR Power and update w_t and w_{t+1} as $w_t \leftarrow w_t/\|w_{t+1}\|_2$ and $w_{t+1} \leftarrow w_{t+1}/\|w_{t+1}\|_2$ in VR HB Power. Since these scaling schemes do not impact the sample paths of $w_t/\|w_t\|$, we can obtain the same results with numerical stability.

Another practical issue with the implementations of VR Power and VR HB Power is to estimate λ_1 and λ_2. As appearing in Lemma 3.3 and Lemma 3.7, accurate values of λ_1 and λ_2 are essential to determine the values of η, m, and β (for VR HB Power). In the experiments, the mini-batch size $|S|$ is given as some percentage of n, so no estimation is required for $|S|$. In order to estimate λ_1 and λ_2 at a regular interval (at the start of each inner-loop), we use the exact gradients of two consecutive outer-loop iterates $\tilde{w}_{s-1}$ and $\tilde{w}_s$. Since we expect that $\tilde{w}_s$ approaches u_1 as the iterations advance, using the Rayleigh quotient, we estimate λ_1 as

$$\hat{\lambda}_1 = \frac{(\tilde{w}_s)^T C(\tilde{w}_s)}{(\tilde{w}_s)^T \tilde{w}_s}. \tag{17}$$

To estimate λ_2 in the same way, we need an estimate of u_2. In Power iteration, an iterate first approaches the subspace spanned by u_1 and u_2 before converging to u_1. That being said, after a number of iterations, we can approximate it by a linear combination of u_1 and u_2. Based on this observation, we estimate u_2 as

$$\hat{u}_2 = \tilde{w}_{s-1} - (\tilde{w}_{s-1}^T \tilde{w}_s)\tilde{w}_s. \quad (18)$$

The idea of the above estimation is to project $\tilde{w}_{s-1}$ to the space orthogonal to $\tilde{w}_s$. If $\tilde{w}_s \approx u_1$ and $\tilde{w}_{s-1} \approx \alpha_1 u_1 + \alpha_2 u_2$ for some $\alpha_1, \alpha_2 (\neq 0)$, we have $\hat{u}_2 \approx u_2$. Using the Rayleigh quotient of $\hat{u}_2$, we estimate λ_2 as

$$\hat{\lambda}_2 = \frac{\tilde{w}_{s-1}^T C \tilde{w}_{s-1} - 2\theta_s \tilde{w}_s^T C \tilde{w}_{s-1} + \theta_s^2 \tilde{w}_s^T C \tilde{w}_s}{1 - \theta_s^2} \quad (19)$$

where $\theta_s = \tilde{w}_{s-1}^T \tilde{w}_s$. While two matrix-vector multiplications, $C\tilde{w}_{s-1}$ and $C\tilde{w}_s$, are involved in computing (17) and (19), they incur no extra computation since they are the exact gradients of $\tilde{w}_{s-1}$ and $\tilde{w}_s$, which are computed regardless of the estimation. As a result, we can obtain $\hat{\lambda}_1$ and $\hat{\lambda}_2$ by only computing some inner products. For initial estimation of $\hat{\lambda}_1$ and $\hat{\lambda}_2$, we run Power iteration five times and use the last two iterates. Note that the exact gradient of the last iterate is computed at the start of the very first outer-loop iteration.

Given $|S|$ and estimates of λ_1 and λ_2, we use bisection search to find $\eta \in (0, 1]$ such that the terms on the right-hand sides of (14) and (16) are almost equal to $|S|$. After η is found, we use (13) and (15) to determine m.

5 Numerical Experiments

In this section, we test the performance of VR Power and VR HB Power with that of (i) VR-PCA (Shamir, 2015), (ii) VR Power+M (Xu et al., 2018) and (iii) Fast PCA (Garber and Hazan, 2015) for finding the first eigenvector u_1 of the covariance matrix C constructed by data vectors $a_i, i = 1, \dots, n$ from real world datasets. Note that all present stochastic variance-reduced PCA algorithms are compared in this experiment.

5.1 Datasets

The datasets include ijcnn (Prokhorov, 2001), covertype (Blackard and Dean, 1999), YearPredictionMSD (Bertin-Mahieux et al., 2011) and MNIST (LeCun et al., 1998) as summarized in Tabel 2. All of them are obtained either from the UCI repository (Dheeru and Karra Taniskidou, 2017) or the LIBSVM library (Chang and Lin, 2011). They are carefully chosen to incorporate a variety of datasets in terms of size and eigen-gap. The first three datasets are standardized with a mean

Table 2: A summary of datasets

DATASET	n	d	Δ
ICJNN(TEST)	91,701	22	0.0079
COV	581,012	54	0.2106
MSD	463,715	90	0.3224
MNIST	70,000	764	0.8851

of zero and standard deviation of one while the last one is scaled to the range between 0 and 1 to preserve its sparsity.

5.2 Settings

In order to report a comprehensive comparison of the algorithms, we consider two settings for selecting hyper-parameters. In the first setting, we use hyper-parameter tuning. Specifically, we use a grid search to find the best values of η, m and $|S| = \rho\%$ of each algorithm and dataset where $\eta \in \{0.01, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 1.0\}$, $m \in \{25, 50, 100, 200\}$ and $\rho \in \{1, 2, 5, 10\}$.

In the second setting, we use the following theoretically derived or recommended hyper-parameter values.

- VR-PCA: $\eta = \sqrt{n}/\sum_{i=1}^n \|a_i\|^2$, $m = n$, $|S| = 1$.
- VR Power+M: $\beta = \lambda_2^2/4$, $\sigma^2 = \sum_{i=1}^n \|a_i\|^2/n$,
$$|S| = \frac{\lambda_2 \log 16}{\sqrt{\lambda_1^2 - \lambda_2^2}}, \quad T = \frac{512 \log 16 \lambda_2 \sigma^2 \sqrt{d}}{\sqrt{\lambda_1^2 - \lambda_2^2}}.$$
- Fast PCA: $\delta = \lambda_1 - \lambda_2$. We only consider the accurate regime. In order to solve each problem, we use SVRG (Johnson and Zhang, 2013) with $\tilde{\epsilon} = 10^{-6}$,
$$\eta = \frac{\lambda_1 - \lambda_2}{7(2\lambda_1 + \lambda_2)^2}, \quad m = \left\lceil \frac{1}{2\eta^2(2\lambda_1 + \lambda_2)^2} \right\rceil.$$
- VR Power, VR HB Power: $|S| = \rho\% \cdot n$ for $\rho \in \{1, 2\}$ and $\sigma^2 = \sum_{i=1}^n \|a_i\|^2/n$. For η and m, we use bisection search explained in Section 4. Also, the scaling schemes in Section 4 are used to ensure numerical stability. The exact values of λ_1 and λ_2 are used to find η and m.
- PF VR Power, PF VR HB Power: As opposed to VR Power and VR HB Power, adaptive estimates of $\hat{\lambda}_1$ and $\hat{\lambda}_2$ obtained by the procedure in Section 4 are used to find η and m.

5.3 Results

Figure 1 displays the experimental result with hyper-parameter tuning. In the figure, the x-axis represents time in seconds and the y-axis represents the optimality

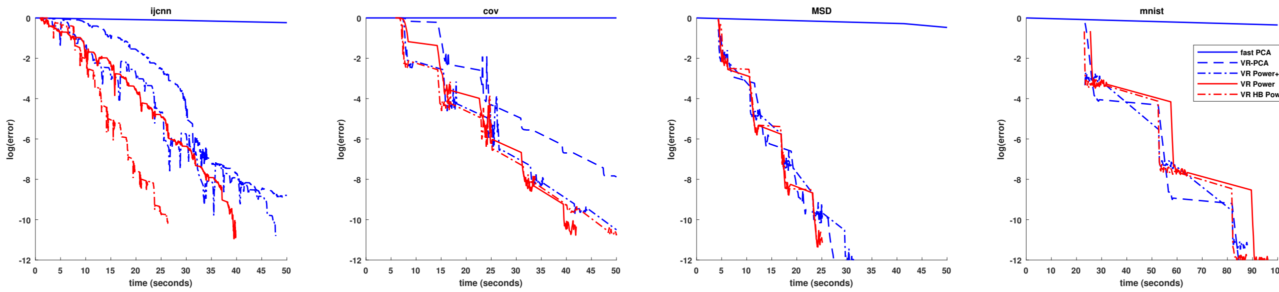

Figure 1: The comparison of stochastic variance-reduced PCA algorithms with hyper-parameters tuned

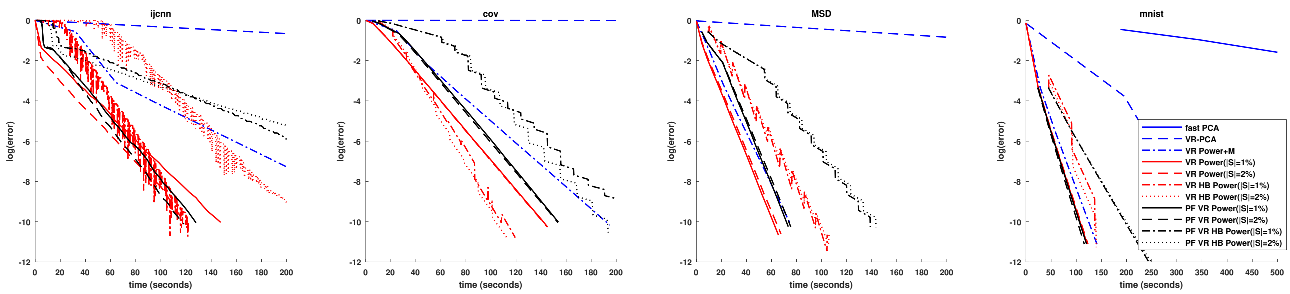

Figure 2: The comparison of stochastic variance-reduced PCA algorithms with recommended hyper-parameters and parameter-free algorithms

gap, $1-(\tilde{w}_s^T u_1)^2$, in the log-scale. Since VR-PCA and VR Power are related algorithms, their performances are similar except for cov where the step size of VR-PCA is tuned to the largest possible value of 1.0. If some larger values are included in the grid, VR-PCA would have a similar performance to VR Power even for cov. On the other hand, VR HB Power always performs better than VR Power+M due to its additional control through the step size. VR HB Power works particularly well for ijcnn which has the smallest eigen-gap. If the eigen-gap is large, the performance of VR HB Power is not much different from the performances of VR Power+M, VR-PCA and VR Power. We were not able to find good hyperparameters for Fast PCA.

Figure 2 shows the experimental result without parameter tuning. In the figure, regardless of the batch size, VR Power and VR HB Power outperform VR-PCA, VR Power+M and Fast PCA. Although VR Power and VR-PCA are similar algorithms, the performance of VR Power is much better than that of VR-PCA due to the choice of η and m. While VR Power precisely choose the values of η and m depending on the values of λ_1, λ_2 and $|S|$, VR-PCA does not utilize such information and let them depend only on n. As a result, the step size is too small and the epoch length is too large, leading to slow convergence. On the other hand, due to the extra dependency on $\sqrt{d}$, VR Power+M requires too large samples and thus it is slower than VR Power even for ijcnn which has the smallest eigen-gap. The epoch length m of SVRG in Fast PCA is of the order of $1/\Delta^2$. Therefore, Fast PCA takes a significant amount of time to solve each convex sub-problem and therefore it does not appear in the figures of ijcnn, cov, and MSD. While it appears in the figure of mnist, its optimality gap does not decrease as sharply as other algorithms. On the other hand, PF VR HB Power takes more than 50 seconds than VR HB Power while the performance of PF VR Power looks very similar to that of VR Power. This is because VR HB Power has the additional momentum parameter β, which makes its performance more affected by estimation errors. Nevertheless, both parameter-free algorithms work very well compared to other algorithms.

6 Conclusion

In this paper, we present two mini-batch stochastic variance-reduced algorithms for PCA and derive exact forms of their parameters to attain the optimal runtime. Our results show that for any batch size, the optimal runtime can be achieved by appropriately choosing the step size and epoch length. We also introduce practical implementations which automatically find such values depending on batch sizes. The framework used in our analysis is not specific to the proposed algorithms but can be applied to analyze other stochastic variance-reduced PCA algorithms and improve their results. In our framework, the optimality gap is measured as the ratio of two expectation terms and this enables us to develop global convergence statements. Experimental results show that the proposed algorithms work well for arbitrary batch sizes.

References

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

Allen-Zhu, Z. and Li, Y. (2017). First efficient convergence for streaming k-PCA: a global, gap-free, and near-optimal rate. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science*, pages 487–492. IEEE.

Arora, R., Cotter, A., Livescu, K., and Srebro, N. (2012). Stochastic optimization for PCA and PLS. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 861–868. IEEE.

Balcan, M.-F., Du, S. S., Wang, Y., and Yu, A. W. (2016). An improved gap-dependency analysis of the noisy power method. In *Conference on Learning Theory*, pages 284–309.

Balsubramani, A., Dasgupta, S., and Freund, Y. (2013). The fast convergence of incremental PCA. In *Advances in Neural Information Processing Systems*, pages 3174–3182.

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., and Lamere, P. (2011). The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval*.

Blackard, J. A. and Dean, D. J. (1999). Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and Electronics in Agriculture*, 24(3):131–151.

Boutsidis, C., Garber, D., Karnin, Z., and Liberty, E. (2015). Online principal components analysis. In *Proceedings of the Twenty-Sixth annual ACM-SIAM Symposium on Discrete Algorithms*, pages 887–901. Society for Industrial and Applied Mathematics.

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27.

Dheeru, D. and Karra Taniskidou, E. (2017). UCI machine learning repository.

Garber, D. and Hazan, E. (2015). Fast and simple PCA via convex optimization. *arXiv preprint arXiv:1509.05647*.

Garber, D., Hazan, E., Jin, C., Kakade, S. M., Musco, C., Netrapalli, P., and Sidford, A. (2016). Faster eigenvector computation via shift-and-invert preconditioning. In *International Conference on Machine Learning*, pages 2626–2634.

Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*, volume 3. JHU Press.

Hardt, M. and Price, E. (2014). The noisy power method: A meta algorithm with applications. In *Advances in Neural Information Processing Systems*, pages 2861–2869.

Jain, P., Jin, C., Kakade, S. M., Netrapalli, P., and Sidford, A. (2016). Streaming PCA: Matching matrix Bernstein and near-optimal finite sample guarantees for Ojas algorithm. In *Conference on Learning Theory*, pages 1147–1164.

Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323.

Jolliffe, I. (2011). *Principal component analysis*. Springer.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lei, Q., Zhong, K., and Dhillon, I. S. (2016). Coordinate-wise power method. In *Advances in Neural Information Processing Systems*, pages 2064–2072.

Li, C. J., Wang, M., Liu, H., and Zhang, T. (2018). Near-optimal stochastic approximation for online principal component estimation. *Mathematical Programming*, 167(1):75–97.

Mason, J. C. and Handscomb, D. C. (2002). *Chebyshev polynomials*. Chapman and Hall/CRC.

Mitliagkas, I., Caramanis, C., and Jain, P. (2013). Memory limited, streaming PCA. In *Advances in Neural Information Processing Systems*, pages 2886–2894.

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

Polyak, B. T. (1964). Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17.

Prokhorov, D. (2001). Ijcnn 2001 neural network competition. *Slide Presentation in International Joint Conference on Neural Networks*, 1:97.

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

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

Wang, J., Wang, W., Garber, D., and Srebro, N. (2018). Efficient coordinate-wise leading eigenvector computation. In *Algorithmic Learning Theory*, pages 806–820.

Xu, P., He, B., De Sa, C., Mitliagkas, I., and Re, C. (2018). Accelerated stochastic power iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 58–67.

Xu, Z. (2018). Gradient descent meets shift-and-invert preconditioning for eigenvector computation. In *Advances in Neural Information Processing Systems*, pages 2830–2839.

A Supplementary Material

In the proofs below, for $\alpha, \beta \geq 0$, we let $Y_t(A,\beta)$ and $Z_t(A,\beta)$ be matrix polynomials such that

$$Y_t(A,\beta) = 2AY_{t-1}(A,\beta) - \beta Y_{t-2}(A,\beta),\ t \geq 2,\ Y_1(A,\beta) = A,\ Y_0(A,\beta) = I, \tag{20}$$

$$Z_t(A,\beta) = 2AZ_{t-1}(A,\beta) - \beta Z_{t-2}(A,\beta),\ t \geq 2,\ Z_1(A,\beta) = 2A,\ Z_0(A,\beta) = I. \tag{21}$$

and let $y_t(\alpha,\beta)$ and $z_t(\alpha,\beta)$ be recurrence polynomials such that

$$y_t(\alpha,\beta) = \sqrt{\alpha} y_{t-1}(\alpha,\beta) - \beta y_{t-2}(\alpha,\beta),\ t \geq 2,\ y_1(\alpha,\beta) = \frac{\sqrt{\alpha}}{2},\ y_0(\alpha,\beta) = 1, \tag{22}$$

$$z_t(\alpha,\beta) = \sqrt{\alpha} z_{t-1}(\alpha,\beta) - \beta z_{t-2}(\alpha,\beta),\ t \geq 2,\ z_1(\alpha,\beta) = \sqrt{\alpha},\ z_0(\alpha,\beta) = 1. \tag{23}$$

For a sequence of matrices $B_0, B_1, B_2, \cdots$, let

$$\prod_{i=j}^{k} B_i = \begin{cases} B_j B_{j-1} \cdots B_k & \text{if } j \geq k \\ I, & \text{otherwise} \end{cases}.$$

Since the eigenvectors $u_1, u_2, \ldots, u_d$ form an orthogonal basis, we frequently use the fact that for $w \in \mathbb{R}^d$, we have $\|w\|^2 = \sum_{k=1}^d (u_k^T w)^2$.

A.1 Main Results

Lemma A.1. *For $w \in \mathbb{R}^d$ such that $\|w\| = 1$ and $t \geq 0$, we have*

$$\|P\left[(1-\eta)I + \eta C\right]^t w\|^2 \leq 2(1-\eta+\eta\lambda_1)^{2t}(1-(u_1^T w)^2), \tag{24a}$$

$$\|PY_t((1-\eta)I + \eta C, \beta(\eta))w\|^2 \leq 4(1-(u_1^T w)^2) p_t(\alpha_1(\eta), \beta(\eta)), \tag{24b}$$

$$\|Z_t((1-\eta)I + \eta C, \beta(\eta))\|^2 \leq q_t(\alpha_1(\eta), \beta(\eta)). \tag{24c}$$

Proof. Since $u_1, u_2, \cdots, u_d$ forms an orthogonal basis in $\mathbb{R}^d$, we can write $w = \sum_{k=1}^d (u_k^T w) u_k$. From that (λ_k, u_k) are eigenpairs of C, we have

$$\left[(1-\eta)I + \eta C\right]^t w = \sum_{k=1}^{d} (u_k^T w)(1-\eta+\eta\lambda_k)^t u_k. \tag{25}$$

Since

$$\begin{aligned}
\|P\left[(1-\eta)I + \eta C\right]^t w\|^2 &= w^T \left[(1-\eta)I + \eta C\right]^t P^2 \left[(1-\eta)I + \eta C\right]^t w \\
&= w^T \left[(1-\eta)I + \eta C\right]^t P \left[(1-\eta)I + \eta C\right]^t w \\
&= w^T \left[(1-\eta)I + \eta C\right]^t (I - ww^T) \left[(1-\eta)I + \eta C\right]^t w \\
&= \|\left[(1-\eta)I + \eta C\right]^t w\|^2 - \left(w^T \left[(1-\eta)I + \eta C\right]^t w\right)^2,
\end{aligned}$$

using (25), we have

$$\begin{aligned}
\|P\left[(1-\eta)I + \eta C\right]^t w\|^2 &= \sum_{k=1}^{d} (u_k^T w)^2 (1-\eta+\eta\lambda_k)^{2t} - \Big(\sum_{k=1}^{d} (u_k^T w)^2 (1-\eta+\eta\lambda_k)^t\Big)^2 \\
&\leq (1-\eta+\eta\lambda_1)^{2t} - (u_1^T w)^4 (1-\eta+\eta\lambda_1)^{2t} \\
&\leq 2(1-(u_1^T w)^2)(1-\eta+\eta\lambda_1)^{2t}
\end{aligned}$$

where the last inequality follows from

$$1 - (u_1^T w)^4 = \big(1 + (u_1^T w)^2\big)\big(1 - (u_1^T w)^2\big) \leq 2\big(1 - (u_1^T w)^2\big). \tag{26}$$

To prove (24b), we first show that

$$Y_t((1-\eta)I + \eta C, \beta(\eta))u_k = y_t(\alpha_k(\eta), \beta(\eta))u_k. \tag{27}$$

First, consider the cases when $t = 0$ and $t = 1$. For $t = 0$, we have $Y_0((1-\eta)I + \eta C, \beta(\eta))u_k = y_0(\alpha_k(\eta), \beta(\eta))u_k$. For $t = 1$, it follows that

$$Y_1((1-\eta)I + \eta C, \beta(\eta))u_k = ((1-\eta)I + \eta C)u_k = (1-\eta+\eta\lambda_k)u_k = \frac{\sqrt{\alpha_k(\eta)}}{2}u_k = y_1(\alpha_k(\eta), \beta(\eta))u_k.$$

Suppose that (27) holds for $t-1$ and $t-2$. Using the definition of Y_t in (20), we have

$$\begin{aligned}
Y_t((1-\eta)I + \eta C, \beta(\eta))u_k &= [2((1-\eta)I + \eta C)Y_{t-1}((1-\eta)I + \eta C, \beta(\eta)) - \beta(\eta)Y_{t-2}((1-\eta)I + \eta C, \beta(\eta))]\, u_k \\
&= [2(1-\eta+\eta\lambda_k)y_{t-1}(\alpha_k(\eta), \beta(\eta)) - \beta(\eta)y_{t-2}(\alpha_k(\eta), \beta(\eta))]\, u_k \\
&= \left[\sqrt{\alpha_k(\eta)}y_{t-1}(\alpha_k(\eta), \beta(\eta)) - \beta(\eta)y_{t-2}(\alpha_k(\eta), \beta(\eta))\right] u_k \\
&= y_t(\alpha_k(\eta), \beta(\eta))u_k.
\end{aligned}$$

This completes the proof of (27).

Next, we show that

$$(y_t(\alpha_k(\eta), \beta(\eta))^2 = p_t(\alpha_k(\eta), \beta(\eta)). \tag{28}$$

For the base cases, we have

$$(y_0(\alpha_k(\eta), \beta(\eta))^2 = 1 = p_0(\alpha_k(\eta), \beta(\eta)), \quad (y_1(\alpha_k(\eta), \beta(\eta))^2 = \frac{\alpha_k}{4} = p_1(\alpha_k(\eta), \beta(\eta))$$

and

$$(y_2(\alpha_k(\eta), \beta(\eta))^2 = \left(\sqrt{\alpha_k(\eta)}y_1(\alpha_k(\eta), \beta(\eta)) - \beta(\eta)y_0(\alpha_k(\eta), \beta(\eta))\right)^2 = \left(\frac{\alpha(\eta)}{2} - \beta(\eta)\right)^2 = p_2(\alpha_k(\eta), \beta(\eta)).$$

Using the definition of y_t in (22) for t and $t-1$, we have

$$\begin{aligned}
(y_t(\alpha_k(\eta), \beta(\eta)))^2 &= (\sqrt{\alpha_k(\eta)}y_{t-1}(\alpha_k(\eta), \beta(\eta)) - \beta(\eta)y_{t-2}(\alpha_k(\eta), \beta(\eta)))^2 \\
&= \alpha_k(\eta)(y_{t-1}(\alpha_k(\eta), \beta(\eta)))^2 - 2\sqrt{\alpha_k(\eta)}\beta(\eta)y_{t-1}(\alpha_k(\eta), \beta(\eta))y_{t-2}(\alpha_k(\eta), \beta(\eta)) \\
&\quad + \beta(\eta)^2(y_{t-2}(\alpha_k(\eta), \beta(\eta)))^2
\end{aligned}$$

and

$$\begin{aligned}
(y_{t-1}(\alpha_k(\eta), \beta(\eta)))^2 &= \alpha_k(\eta)(y_{t-2}(\alpha_k(\eta), \beta(\eta)))^2 - 2\sqrt{\alpha_k(\eta)}\beta(\eta)y_{t-2}(\alpha_k(\eta), \beta(\eta))y_{t-3}(\alpha_k(\eta), \beta(\eta)) \\
&\quad + \beta(\eta)^2(y_{t-3}(\alpha_k(\eta), \beta(\eta)))^2.
\end{aligned}$$

Moreover, since

$$y_{t-1}(\alpha_k(\eta), \beta(\eta))y_{t-2}(\alpha_k(\eta), \beta(\eta)) = \sqrt{\alpha_k(\eta)}(y_{t-2}(\alpha_k(\eta), \beta(\eta)))^2 - \beta(\eta)y_{t-2}(\alpha_k(\eta), \beta(\eta))y_{t-3}(\alpha_k(\eta), \beta(\eta)),$$

we have

$$\begin{aligned}
(y_t(\alpha_k(\eta), \beta(\eta)))^2 &= \alpha_k(\eta)(y_{t-1}(\alpha_k(\eta), \beta(\eta)))^2 - 2\alpha_k(\eta)\beta(\eta)(y_{t-2}(\alpha_k(\eta), \beta(\eta)))^2 + \beta(\eta)^2(y_{t-2}(\alpha_k(\eta), \beta(\eta)))^2 \\
&\quad + 2\sqrt{\alpha_k(\eta)}\beta(\eta)^2y_{t-2}(\alpha_k(\eta), \beta(\eta))y_{t-3}(\alpha_k(\eta), \beta(\eta)) \\
&= \alpha_k(\eta)(y_{t-1}(\alpha_k(\eta), \beta(\eta)))^2 - 2\alpha_k(\eta)\beta(\eta)(y_{t-2}(\alpha_k(\eta), \beta(\eta)))^2 + \beta(\eta)^2(y_{t-2}(\alpha_k(\eta), \beta(\eta)))^2 \\
&\quad + \beta(\eta)\left(\alpha_k(\eta)(y_{t-2}(\alpha_k(\eta), \beta(\eta)))^2 + \beta(\eta)^2(y_{t-3}(\alpha_k(\eta), \beta(\eta)))^2 - (y_{t-1}(\alpha_k(\eta), \beta(\eta)))^2\right) \\
&= (\alpha_k(\eta) - \beta(\eta))(y_{t-1}(\alpha_k(\eta), \beta(\eta)))^2 - \beta(\eta)(\alpha_k(\eta) - \beta(\eta))(y_{t-2}(\alpha_k(\eta), \beta(\eta)))^2 \\
&\quad + \beta(\eta)^3(y_{t-3}(\alpha_k(\eta), \beta(\eta)))^2.
\end{aligned}$$

This proves (28).

Now, using (27), we have

$$Y_t((1-\eta)I+\eta C,\beta(\eta))w=\sum_{k=1}^{d}y_t(\alpha_k(\eta),\beta(\eta))(u_k^Tw)u_k. \tag{29}$$

Since $u_1, u_2, \cdots, u_d$ form an orthogonal basis in $\mathbb{R}^d$, we have

$$\|Y_t((1-\eta)I+\eta C,\beta(\eta))w\|^2=\sum_{k=1}^{d}(y_t(\alpha_k(\eta),\beta(\eta)))^2(u_k^Tw)^2=\sum_{k=1}^{d}p_t(\alpha_k(\eta),\beta(\eta))(u_k^Tw)^2.$$

Using (90) and (92) in Lemma A.4, for $k \geq 2$, we have

$$p_t(\alpha_k(\eta),\beta(\eta))\leq p_t(\alpha_1(\eta),\beta(\eta)) \tag{30}$$

Since $\sum_{k=1}^{d}(u_k^Tw)^2=1$, we have

$$\|Y_t((1-\eta)I+\eta C,\beta(\eta))w\|^2\leq p_t(\alpha_1(\eta),\beta(\eta)).$$

Moreover, using $(u_1^Tw)^2 \leq 1$ and (29), we obtain

$$\begin{aligned}\left(w^TY_t((1-\eta)I+\eta C,\beta(\eta))w\right)^2&=\Big(y_t(\alpha_1(\eta),\beta(\eta))(u_1^Tw)^2+\sum_{k=2}^{d}y_t(\alpha_k(\eta),\beta(\eta))(u_k^Tw)^2\Big)^2\\&\geq(y_t(\alpha_1(\eta),\beta(\eta)))^2(u_1^Tw)^4-2y_t(\alpha_1(\eta),\beta(\eta))\sum_{k=2}^{d}|y_t(\alpha_k(\eta),\beta(\eta))|(u_k^Tw)^2\\&\geq(y_t(\alpha_1(\eta),\beta(\eta)))^2(u_1^Tw)^4-2(y_t(\alpha_1(\eta),\beta(\eta)))^2(1-(u_k^Tw)^2)\end{aligned}$$

Therefore,

$$\begin{aligned}\|PY_t((1-\eta)I+\eta C,\beta(\eta))w\|^2&=\|Y_t((1-\eta)I+\eta C,\beta(\eta))w\|^2-\left(w^TY_t((1-\eta)I+\eta C,\beta(\eta))w\right)^2\\&\leq(y_t(\alpha_1(\eta),\beta(\eta)))^2(1-(u_1^Tw)^4)+2(y_t(\alpha_1(\eta),\beta(\eta)))^2(1-(u_k^Tw)^2)\\&\leq 4(y_t(\alpha_1(\eta),\beta(\eta)))^2(1-(u_k^Tw)^2)\end{aligned}$$

where the last inequality follows from (26).

Lastly, we prove (24c). In the same way we prove (27) and (28), we can show that

$$Z_t((1-\eta)I+\eta C,\beta(\eta))u_k=z_t(\alpha_k(\eta),\beta(\eta))u_k,\quad(z_t(\alpha_k(\eta),\beta(\eta))^2=q_t(\alpha_k(\eta),\beta(\eta)). \tag{31}$$

Using (91) and (92) in Lemma A.4, for $k \geq 2$, we have

$$q_t(\alpha_k(\eta),\beta(\eta))\leq q_t(\alpha_1(\eta),\beta(\eta)). \tag{32}$$

Using (31), we have

$$w^TZ_t((1-\eta)I+\eta C,\beta(\eta))w=\sum_{k=1}^{d}z_t(\alpha_k(\eta),\beta(\eta))(u_k^Tw)^2\leq\sum_{k=1}^{d}|z_t(\alpha_k(\eta),\beta(\eta))|(u_k^Tw)^2.$$

Moreover, using (32) and the fact that $\sum_{k=1}^{d}(u_k^Tw)^2=1$, we have

$$\sum_{k=1}^{d}|z_t(\alpha_k(\eta),\beta(\eta))|(u_k^Tw)^2\leq|z_t(\alpha_1(\eta),\beta(\eta))|\sum_{k=1}^{d}(u_k^Tw)^2=|z_t(\alpha_1(\eta),\beta(\eta))|.$$

This results in

$$w^TZ_t((1-\eta)I+\eta C,\beta(\eta))w\leq|z_t(\alpha_1(\eta),\beta(\eta))|,$$

leading to

$$\|Z_t((1-\eta)I+\eta C,\beta(\eta))\|^2\leq|z_t(\alpha_1(\eta),\beta(\eta))|^2=q_t(\alpha_1(\eta),\beta(\eta)).$$

This complets the proof. □

A.1.1 VR Power

Proof of Lemma 3.1. Since $Pw_0 = \left(I - w_0 w_0^T\right) w_0 = 0$, we have

$$u_k^T w_1 = (1-\eta) u_k^T w_0 + \eta u_k^T C w_0 + \eta u_k^T (C_0 - C) P w_0 = (1 - \eta + \eta\lambda_k) u_k^T w_0. \tag{33}$$

Taking the expectation of the square of (33), we obtain

$$E[(u_k^T w_1)^2] = (1 - \eta + \eta\lambda_k)^2 E[(u_k^T w_0)^2]. \tag{34}$$

For $t \geq 2$, we have

$$u_k^T w_t = (1 - \eta + \eta\lambda_k) u_k^T w_{t-1} + \eta u_k^T (C_{t-1} - C) P w_{t-1}. \tag{35}$$

Since S_t is sampled uniformly at random, C_t is independent of $S_1, \ldots, S_{t-1}$ and w_0 with $E[C_t] = C$, leading to

$$\begin{aligned} E[u_k^T w_{t-1} u_k^T (C_{t-1} - C) P w_t] &= E[E[u_k^T w_{t-1} u_k^T (C_{t-1} - C) P w_t | w_0, S_1, \ldots, S_{t-2}]] \\ &= E[u_k^T w_{t-1} u_k^T E[C_{t-1} - C] P w_t] = 0. \end{aligned}$$

Therefore, taking the expectation of the square of (35), we have

$$\begin{aligned} E[(u_k^T w_t)^2] &= (1 - \eta + \eta\lambda_k)^2 E[(u_k^T w_{t-1})^2] + \eta^2 E[w_{t-1}^T P (C_{t-1} - C) u_k u_k^T (C_{t-1} - C) P w_{t-1}] \\ &= (1 - \eta + \eta\lambda_k)^2 E[(u_k^T w_{t-1})^2] + \eta^2 E[w_{t-1}^T P M_k P w_{t-1}] \end{aligned} \tag{36}$$

where the last equality follows from

$$\begin{aligned} E[w_{t-1}^T P (C_{t-1} - C) u_k u_k^T (C_{t-1} - C) P w_{t-1}] &= E[E[w_{t-1}^T P (C_{t-1} - C) u_k u_k^T (C_{t-1} - C) P w_{t-1} | w_0, S_1, \ldots, S_{t-2}]] \\ &= E[w_{t-1}^T P E[(C_{t-1} - C) u_k u_k^T (C_{t-1} - C)] P w_{t-1}] \\ &= E[w_{t-1}^T P M_k P w_{t-1}]. \end{aligned}$$

Repeatedly applying (36) and using (34), we obtain

$$E[(u_k^T w_t)^2] = (1 - \eta + \eta\lambda_k)^{2t} E[(u_k^T w_0)^2] + \eta^2 \sum_{i=1}^{t-1} (1 - \eta + \eta\lambda_k)^{2(t-i-1)} E[w_i^T P M_k P w_i].$$

□

Proof of Lemma 3.2. By Lemma A.2, we have

$$\sum_{k=2}^{d} E[w_t^T P M_k P w_t] = \sum_{k=2}^{d} E[w_t^T P M_k P w_t] = E[w_t^T P \sum_{k=2}^{d} M_k P w_t] \leq \| \sum_{k=2}^{d} M_k \| \cdot E[\|P w_t\|^2]. \tag{37}$$

Using the Jensen's inequality and the fact that $\| \sum_{k=2}^{d} u_k u_k^T \| = 1$, we have

$$\| \sum_{k=2}^{d} M_k \| = \| \sum_{k=2}^{d} E[(C_t - C) u_k u_k^T (C_t - C)] \| \leq E[\|C_t - C\|^2] = E[\|(C_t - C)^2\|] = K,$$

resulting in

$$\sum_{k=2}^{d} E[w_t^T P M_k P w_t] \leq K E[\|P w_t\|^2]. \tag{38}$$

Let

$$B_i = (1-\eta) I + \eta C + \eta (C_i - C) P.$$

Since $Pw_0 = 0$ and

$$\begin{aligned} \prod_{i=t-1}^{0} B_i &= \prod_{i=t-1}^{1} B_i \eta (C_0 - C) P + \prod_{i=t-1}^{1} B_i \left((1-\eta) I + \eta C\right) \\ &= \prod_{i=t-1}^{1} B_i \eta (C_0 - C) P + \sum_{j=1}^{t-1} \prod_{i=t-1}^{j+1} B_i \eta (C_j - C) P \left[(1-\eta) I + \eta C\right]^j + \left[(1-\eta) I + \eta C\right]^t, \end{aligned}$$

which can be seen by elementary manipulation, we have

$$w_t = \prod_{i=t-1}^{0} B_i w_0 = \left[\sum_{j=1}^{t-1} \prod_{i=t-1}^{j+1} B_i \eta (C_j - C) P \left[(1-\eta)I + \eta C\right]^j + \left[(1-\eta)I + \eta C\right]^t \right] w_0,$$

resulting in

$$Pw_t = P \prod_{i=t-1}^{0} B_i w_0 = \left[\sum_{j=1}^{t-1} P \prod_{i=t-1}^{j+1} B_i \eta (C_j - C) P \left[(1-\eta)I + \eta C\right]^j + P\left[(1-\eta)I + \eta C\right]^t \right] w_0. \tag{39}$$

Since $C_0, \cdots, C_{t-1}$ are independent with $E[C_i] = C$ for all $1 \le i \le t-1$, we obtain

$$E\Big[w_0^T \left[(1-\eta)I + \eta C\right]^t P^2 \prod_{i=t-1}^{j+1} B_i \eta (C_j - C) P \left[(1-\eta)I + \eta C\right]^j w_0 \Big] = 0 \tag{40}$$

$$E\Big[w_0^T \left[(1-\eta)I + \eta C\right]^{j_1} P (C_{j_1} - C)\eta \prod_{i=j_1+1}^{t-1} B_i P^2 \prod_{i=t-1}^{j_2+1} B_i \eta (C_{j_2} - C) P \left[(1-\eta)I + \eta C\right]^{j_2} w_0 \Big] = 0 \tag{41}$$

where $1 \le j, j_1, j_2 \le t-1$ and $j_1 \ne j_2$. Therefore, we have

$$E[\|Pw_t\|^2] = \sum_{j=1}^{t-1} E\big[\big\|P \prod_{i=t-1}^{j+1} B_i \eta (C_j - C) P \left[(1-\eta)I + \eta C\right]^j w_0 \big\|^2\big] + E[\|P\left[(1-\eta)I + \eta C\right]^t w_0\|^2] \tag{42}$$

due to cross-terms being 0 from (40) and (41) when "squaring" (39). Using Lemma A.1 with $w = w_0/\|w_0\|$ and the fact that $\|w_0\|^2(1 - (u_1^T w_0)^2/\|w_0\|^2) = \sum_{k=2}^d (u_k^T w_0)^2$, we have

$$E[\|P[(1-\eta)I + \eta C]^t w_0\|^2] \le 2(1 - \eta + \eta\lambda_1)^{2t} \sum_{k=2}^{d} E[(u_k^T w_0)^2]. \tag{43}$$

By Lemma A.2 and $\|P\| = 1$, we have

$$\big\|P \prod_{i=t-1}^{j+1} B_i \eta (C_j - C) P \left[(1-\eta)I + \eta C\right]^j w_0 \big\|^2 \le \eta^2 \big\| \prod_{i=t-1}^{j+1} B_i (C_j - C) P \left[(1-\eta)I + \eta C\right]^j w_0 \big\|^2. \tag{44}$$

Moreover, by repeatedly using first the property that B_i is independent of $w_0, C_j, B_{j+1}, \cdots, B_{i-1}$ and Lemma A.2, we have

$$\begin{aligned}
&E\big[\big\| \prod_{i=t-1}^{j+1} B_i (C_j - C) P \big[(1-\eta)I + \eta C\big]^j w_0 \big\|^2\big] \\
&= E[w_0^T [(1-\eta)I + \eta C]^j P (C_j - C) \Big(\prod_{i=t-2}^{j+1} B_i \Big)^T B_{t-1}^T B_{t-1} \prod_{i=t-2}^{j+1} B_i P (C_j - C) [(1-\eta)I + \eta C]^j w_0] \\
&= E[w_0^T [(1-\eta)I + \eta C]^j P (C_j - C) \Big(\prod_{i=t-2}^{j+1} B_i \Big)^T E[B_{t-1}^T B_{t-1}] \prod_{i=t-2}^{j+1} B_i P (C_j - C) [(1-\eta)I + \eta C]^j w_0] \\
&\le \|E[B_{t-1}^T B_{t-1}]\| \cdot E\big[\big\| \prod_{i=t-2}^{j+1} B_i (C_j - C) P \big[(1-\eta)I + \eta C\big]^j w_0 \big\|^2\big] \\
&\le \prod_{i=t-1}^{j+1} \big\|E[B_i^T B_i]\big\| \cdot E[\|(C_j - C) P \left[(1-\eta)I + \eta C\right]^j w_0\|^2].
\end{aligned}$$

In the same way, using the fact that C_j is independent of w_0 and Lemma A.2, we have

$$E[\|(C_j - C) P \left[(1-\eta)I + \eta C\right]^j w_0\|^2] \le \|E[(C_j - C)^2]\| \cdot E[\|P \left[(1-\eta)I + \eta C\right]^j w_0\|^2],$$

resulting in

$$E\big[\|\prod_{i=t-1}^{j+1} B_i(C_j - C)P[(1-\eta)I + \eta C]^j w_0\big\|^2\big] \leq \prod_{i=t-1}^{j+1} \|E[B_i^T B_i]\| \cdot \|E[(C_j - C)^2]\| \cdot E[\|P\,[(1-\eta)I + \eta C]^j\, w_0\|^2]. \tag{45}$$

Since C_i is independent of w_0 and $E[C_i] = C$, we have

$$\|E[B_i^T B_i]\| \leq \|\,[(1-\eta)I + \eta C]^2\,\| + \eta^2 \|E[P(C_i - C)^2 P]\|.$$

Since all induced norms are convex, using the Jensen's inequality, we have

$$\|E[P(C_i - C)^2 P]\|] \leq E[\|P(C_i - C)^2 P\|] \leq E[\|(C_i - C)^2\|] = K,$$

leading to

$$\|E[B_i^T B_i]\| \leq \|\,[(1-\eta)I + \eta C]^2\,\| + \eta^2 \|E[P(C_i - C)^2 P]\| \leq (1 - \eta + \eta\lambda_1)^2 + \eta^2 K. \tag{46}$$

In the same way, we obtain

$$\|E[(C_j - C)^2]\| \leq E[\|(C_j - C)^2\|] = K. \tag{47}$$

Using (46), (47) and (43) for (45), we have

$$E\big[\|\prod_{i=t-1}^{j+1} B_i(C_j - C)P[(1-\eta)I + \eta C]^j w_0\big\|^2\big] \leq K\left[(1-\eta+\eta\lambda_1)^2 + \eta^2 K\right]^{t-j-1} (1-\eta+\eta\lambda_1)^{2j} \sum_{k=2}^{d} E[(u_k^T w_0)^2]. \tag{48}$$

From (42), (43), (44) and (48), we finally have

$$\begin{aligned} E[\|Pw_t\|^2] &\leq 2\left[\sum_{j=1}^{t-1} \eta^2 K\left[(1-\eta+\eta\lambda_1)^2 + \eta^2 K\right]^{t-j-1} (1-\eta+\eta\lambda_1)^{2j} + (1-\eta+\eta\lambda_1)^{2t}\right] \cdot \sum_{k=2}^{d} E[(u_k^T w_0)^2] \\ &\leq 2\left[(1-\eta+\eta\lambda_1)^2 + \eta^2 K\right]^t \cdot \sum_{k=2}^{d} E[(u_k^T w_0)^2], \end{aligned}$$

where the last inequality can be checked by elementary manipulation. This results in

$$\sum_{k=2}^{d} E[w_t^T P M_k P w_t] \leq 2K\left[(1-\eta+\eta\lambda_1)^2 + \eta^2 K\right]^t \cdot \sum_{k=2}^{d} E[(u_k^T w_0)^2]. \tag{49}$$

This proves the first part of the proof.

Next, we have

$$\sum_{k=2}^{d}\sum_{i=1}^{t-1} (1-\eta+\eta\lambda_k)^{2(t-i-1)} E[w_i^T P M_k P w_i] \leq (1-\eta+\eta\lambda_1)^{2t} \cdot \sum_{i=1}^{t-1} (1-\eta+\eta\lambda_1)^{-2(i+1)} \sum_{k=2}^{d} E[w_i^T P M_k P w_i]$$

and

$$\begin{aligned} \sum_{i=1}^{t-1} (1-\eta+\eta\lambda_1)^{-2(i+1)}\left[(1-\eta+\eta\lambda_1)^2 + \eta^2 K\right]^i &\leq \frac{1}{(1-\eta+\eta\lambda_1)^2} \sum_{i=1}^{t-1} \left(\frac{(1-\eta+\eta\lambda_1)^2 + \eta^2 K}{(1-\eta+\eta\lambda_1)^2}\right)^i \\ &\leq \frac{1}{\eta^2 K}\left[\left(1 + \frac{\eta^2 K}{(1-\eta+\eta\lambda_1)^2}\right)^{t-1} - 1\right]\left(1 + \frac{\eta^2 K}{(1-\eta+\eta\lambda_1)^2}\right) \\ &\leq \frac{1}{\eta^2 K}\left[\exp\left(\frac{\eta^2 K t}{(1-\eta+\eta\lambda_1)^2}\right) - 1\right]. \end{aligned}$$

Using the condition that

$$0 < \frac{\eta^2 K m}{(1-\eta+\eta\lambda_1)^2} < 1$$

and the fact $\exp(x) - 1 \le 2x$ for all $x \in (0,1)$, we further obtain

$$\sum_{i=1}^{t-1} (1-\eta+\eta\lambda_1)^{-2(i+1)} \left[(1-\eta+\eta\lambda_1)^2 + \eta^2 K\right]^i \le \frac{2t}{(1-\eta+\eta\lambda_1)^2}.$$

Combined with (49), this results in

$$\begin{aligned}\eta^2 \sum_{k=2}^{d}\sum_{i=1}^{m-1} (1-\eta+\eta\lambda_k)^{2(m-i-1)} E[w_i^T P M_k P w_i] &\le \eta^2 \sum_{i=1}^{m-1} (1-\eta+\eta\lambda_k)^{2(m-i-1)} \sum_{k=2}^{d} E[w_i^T P M_k P w_i] \\ &\le 4\eta^2 K m (1-\eta+\eta\lambda_1)^{2(m-1)} \cdot \sum_{k=2}^{d} E[(u_k^T w_0)^2].\end{aligned}$$

Using Lemma 3.1 for $t = m$ and the fact that $(1-\eta+\eta\lambda_k)^{2m} \le (1-\eta+\eta\lambda_2)^{2m}$ for $k \ge 2$, we finally have

$$\begin{aligned}\sum_{k=2}^{d} E[(u_k^T w_m)^2] &= \sum_{k=2}^{d} (1-\eta+\eta\lambda_k)^{2m} E[(u_k^T w_0)^2] + \eta^2 \sum_{k=2}^{d}\sum_{i=1}^{m-1} (1-\eta+\eta\lambda_k)^{2(m-i-1)} E[w_i^T P M_k P w_i] \\ &\le \left((1-\eta+\eta\lambda_2)^{2m} + 4\eta^2 K m (1-\eta+\eta\lambda_1)^{2(m-1)}\right) \cdot \sum_{k=2}^{d} E[(u_k^T w_0)^2]. \qquad (50)\end{aligned}$$

On the other hand, by Lemma 3.1 and the fact that PM_kP is positive semi-definite, we have

$$(1-\eta+\eta\lambda_1)^{2m} E[(u_1^T w_0)^2] \le E[(u_1^T w_m)^2]. \qquad (51)$$

Combining (51) with (50), we obtain

$$\frac{\sum_{k=2}^{d} E[(u_k^T w_m)^2]}{E[(u_1^T w_m)^2]} \le \left[\left(\frac{1-\eta+\eta\lambda_2}{1-\eta+\eta\lambda_1}\right)^{2m} + \frac{4\eta^2 K m}{(1-\eta+\eta\lambda_1)^2}\right] \cdot \frac{\sum_{k=2}^{d} E[(u_k^T w_0)^2]}{E[(u_1^T w_0)^2]}.$$

□

Proof of Lemma 3.3. From the conditions on η, m and $|S|$, we have

$$0 < \frac{\eta^2 K m}{(1-\eta+\eta\lambda_1)^2} < \frac{1}{16}.$$

Therefore, using Lemma 3.2, we have

$$\frac{\sum_{k=2}^{d} E[(u_k^T w_m)^2]}{E[(u_1^T w_m)^2]} \le \left[\left(\frac{1-\eta+\eta\lambda_2}{1-\eta+\eta\lambda_1}\right)^{2m} + \frac{4\eta^2 K m}{(1-\eta+\eta\lambda_1)^2}\right] \cdot \frac{\sum_{k=2}^{d} E[(u_k^T w_0)^2]}{E[(u_1^T w_0)^2]}.$$

By the choice of η and m, we have

$$\left(\frac{1-\eta+\eta\lambda_2}{1-\eta+\eta\lambda_1}\right)^{2m} = \left(1 - \frac{\eta(\lambda_1-\lambda_2)}{1-\eta+\eta\lambda_1}\right)^{2m} \le \exp\left(-\frac{2\eta(\lambda_1-\lambda_2)m}{1-\eta+\eta\lambda_1}\right) \le \exp(-\log 2) = \frac{1}{2}.$$

Also, by the choice of η, m and $|S|$, we have

$$\frac{4\eta^2 K m}{(1-\eta+\eta\lambda_1)^2} = \frac{4\sigma^2\eta^2 m}{|S|(1-\eta+\eta\lambda_1)^2} \le \frac{1}{4}.$$

Therefore, we have

$$\frac{\sum_{k=2}^{d} E[(u_k^T w_m)^2]}{E[(u_1^T w_m)^2]} \le \frac{3}{4} \cdot \frac{\sum_{k=2}^{d} E[(u_k^T w_0)^2]}{E[(u_1^T w_0)^2]}.$$

□

Proof of Theorem 3.4. By repeatedly applying Lemma 3.3, we have

$$\frac{\sum_{k=2}^{d} E[(u_k^T \tilde{w}_\tau)^2]}{E[(u_1^T \tilde{w}_\tau)^2]} \leq \left(\frac{3}{4}\right)^\tau \frac{\sum_{k=2}^{d} E[(u_k^T \tilde{w}_0)^2]}{E[(u_1^T \tilde{w}_0)^2]} = \left(\frac{3}{4}\right)^\tau \tilde{\theta}_0.$$

Since $\tau = \lceil \log(\tilde{\theta}_0/\epsilon)/\log(4/3) \rceil$, we have

$$\tau \log\left(\frac{3}{4}\right) \leq \log\left(\frac{\epsilon}{\tilde{\theta}_0}\right),$$

resulting in

$$\frac{\sum_{k=2}^{d} E[(u_k^T \tilde{w}_\tau)^2]}{E[(u_1^T \tilde{w}_\tau)^2]} \leq \epsilon.$$

□

A.1.2 VR HB Power

Proof of Lemma 3.5. From

$$\begin{aligned} w_1 &= (1-\eta)w_0 + \eta\tilde{g} \\ &= (1-\eta)w_0 + \eta C w_0, \end{aligned}$$

we have

$$\begin{aligned} u_k^T w_1 &= (1-\eta)u_k^T w_0 + \eta u_k^T C w_0 \\ &= (1-\eta)u_k^T w_0 + \eta\lambda_k u_k^T w_0 \\ &= (1-\eta+\eta\lambda_k)u_k^T w_0. \end{aligned} \tag{52}$$

Taking the expectation of the square of (52), we obtain

$$E[(u_k^T w_1)^2] = (1-\eta+\eta\lambda_k)^2 E[(u_k^T w_0)^2] = \frac{\alpha_k(\eta)}{4} E[(u_k^T w_0)^2]. \tag{53}$$

Next, from (5), we have

$$\begin{aligned} w_{t+1} &= 2\bigg((1-\eta)w_t + \eta\frac{1}{|S_t|}\sum_{i_t\in S_t} a_{i_t}a_{i_t}^T\Big(w_t - \frac{(w_t^T w_0)}{\|w_0\|^2}w_0\Big) + \frac{(w_t^T w_0)}{\|w_0\|^2}\tilde{g}\bigg) - \beta(\eta)w_{t-1} \\ &= 2\bigg((1-\eta)w_t + \eta\frac{1}{|S_t|}\sum_{i_t\in S_t} a_{i_t}a_{i_t}^T\Big(I - \frac{w_0w_0^T}{\|w_0\|^2}\Big)w_t + C\frac{w_0w_0^T}{\|w_0\|^2}w_t\bigg) - \beta(\eta)w_{t-1} \\ &= 2\bigg((1-\eta)w_t + \eta C w_t + \eta\frac{1}{|S_t|}\sum_{i_t\in S_t}(a_{i_t}a_{i_t}^T - C)\Big(I - \frac{w_0w_0^T}{\|w_0\|^2}\Big)w_t\bigg) - \beta(\eta)w_{t-1} \\ &= 2\big((1-\eta)w_t + \eta C w_t + \eta(C_t - C)Pw_t\big) - \beta(\eta)w_{t-1}, \end{aligned} \tag{54}$$

leading to

$$u_k^T w_{t+1} = 2\big((1-\eta+\eta\lambda_k)u_k^T w_t + \eta u_k^T(C_t - C)Pw_t\big) - \beta(\eta)u_k^T w_{t-1}. \tag{55}$$

Taking the square of (55), we have

$$\begin{aligned} (u_k^T w_{t+1})^2 &= 4(1-\eta+\eta\lambda_k)^2(u_k^T w_t)^2 + 4\eta^2 w_t^T P(C_t - C)u_k u_k^T(C_t - C)Pw_t + (\beta(\eta))^2(u_k^T w_{t-1})^2 \\ &\quad + 8\eta(1-\eta+\eta\lambda_k)u_k^T w_t u_k^T(C_t - C)Pw_t - 4(1-\eta+\eta\lambda_k)\beta(\eta)u_k^T w_t u_k^T w_{t-1} \\ &\quad - 4\eta\beta(\eta)u_k^T(C_t - C)Pw_t u_k^T w_{t-1}. \end{aligned} \tag{56}$$

Since S_t is sampled uniformly at random, C_t is independent of $S_1, \ldots, S_{t-1}$ and identically distributed with $E[C_t] = C$. Therefore,

$$E[u_k^T w_t u_k^T(C_t - C)Pw_t] = E[E[u_k^T w_t u_k^T(C_t - C)Pw_t | w_0, S_1, \ldots, S_{t-1}]] = E[u_k^T w_t u_k^T E[C_t - C]Pw_t] = 0.$$

Similarly, we have

$$E[u_k^T(C_t - C)Pw_t u_k^T w_{t-1}] = 0. \tag{57}$$

As a result, we obtain

$$\begin{aligned} E[(u_k^T w_{t+1})^2] &= \alpha_k(\eta)E[(u_k^T w_t)^2] - 2\sqrt{\alpha_k(\eta)}\beta(\eta)E[(u_k^T w_t)(u_k^T w_{t-1})] + (\beta(\eta))^2 E[(u_k^T w_{t-1})^2] \\ &\quad + 4\eta^2 E[w_t^T P M_k P w_t]. \end{aligned} \tag{58}$$

Using (52) and (53) in (58) for $t = 1$, we have

$$E[(u_k^T w_2)^2] = \Big(\frac{\alpha_k(\eta)}{2} - \beta(\eta)\Big)^2 E[(u_k^T w_0)^2] + 4\eta^2 E[w_1^T P M_k P w_1]. \tag{59}$$

Moreover, by using (55) with $t-1$, multiplying it with $u_k^T w_{t-1}$, taking expectation and using (57) with w_t being w_{t-1} (which can be derived in the same way as (57)) , we have

$$E[(u_k^T w_t)(u_k^T w_{t-1})] = \sqrt{\alpha_k(\eta)} E[(u_k^T w_{t-1})^2] - \beta(\eta) E[(u_k^T w_{t-1})(u_k^T w_{t-2})]. \tag{60}$$

Using (60), we can further write (58) as

$$\begin{aligned} E[(u_k^T w_{t+1})^2] = {} & \alpha_k(\eta) E[(u_k w_t)^2] - \beta(\eta)(2\alpha_k(\eta) - \beta(\eta)) E[(u_k^T w_{t-1})^2] \\ & + 2\sqrt{\alpha_k(\eta)}(\beta(\eta))^2 E[(u_k^T w_{t-1})(u_k^T w_{t-2})] + 4\eta^2 E[w_t^T P M_k P w_t]. \end{aligned} \tag{61}$$

With $t-1$ in (58), we have

$$\begin{aligned} E[(u_k^T w_t)^2] = {} & \alpha_k(\eta) E[(u_k^T w_{t-1})^2] - 2\sqrt{\alpha_k(\eta)}\beta(\eta) E[(u_k^T w_{t-1})(u_k^T w_{t-2})] + (\beta(\eta))^2 E[(u_k^T w_{t-2})^2] \\ & + 4\eta^2 E[w_{t-1}^T P M_k P w_{t-1}]. \end{aligned} \tag{62}$$

Adding (62) multiplied by $\beta(\eta)$ to (61), we obtain

$$\begin{aligned} E[(u_k^T w_{t+1})^2] = {} & (\alpha_k(\eta) - \beta(\eta)) E[(u_k^T w_t)^2] - \beta(\eta)(\alpha_k(\eta) - \beta(\eta)) E[(u_k^T w_{t-1})^2] + (\beta(\eta))^3 E[(u_k^T w_{t-2})^2] \\ & + 4\eta^2 E[w_t^T P M_k P w_t] + 4\eta^2 \beta(\eta) E[w_{t-1}^T P M_k P w_{t-1}]. \end{aligned} \tag{63}$$

With $t-1$ in (63), we finally have

$$\begin{aligned} E[(u_k^T w_t)^2] = {} & (\alpha_k(\eta) - \beta(\eta)) E[(u_k^T w_{t-1})^2] - \beta(\eta)(\alpha_k(\eta) - \beta(\eta)) E[(u_k^T w_{t-2})^2] + (\beta(\eta))^3 E[(u_k^T w_{t-3})^2] \\ & + 4\eta^2 E[w_{t-1}^T P M_k P w_{t-1}] + 4\eta^2 \beta(\eta) E[w_{t-2}^T P M_k P w_{t-2}] \end{aligned} \tag{64}$$

for $t \geq 3$.

Using Lemma A.4 for $E[(u_k^T w_t)^2]$ defined by (53), (59), and (64) with

$$\alpha = \alpha_k(\eta), \quad \beta = \beta(\eta), \quad L_0 = E[(u_k^T w_0)^2], \quad L_t = 4\eta^2 E[w_t^T P M_k P w_t],$$

we have

$$E[(u_k^T w_t)^2] = p_t(\alpha_k(\eta), \beta(\eta)) E[(u_k^T w_0)^2] + 4\eta^2 \sum_{r=1}^{t-1} q_{t-r-1}(\alpha_k(\eta), \beta(\eta)) E[w_r^T P M_k P w_r].$$

□

Proof of Lemma 3.6. Since $\| \sum_{k=2}^d u_k u_k^T \| \leq 1$, we have

$$\| \sum_{k=2}^d M_k \| = \| \sum_{k=2}^d E[(C_t - C) u_k u_k^T (C_t - C)] \| \leq E[\|C_t - C\|^2] = E[\|(C_t - C)^2\|] = K.$$

By Lemma A.2, this leads to

$$\sum_{k=2}^d E[w_t^T P M_k P w_t] = E[w_t^T P \sum_{k=2}^d M_k P w_t] \leq \| \sum_{k=2}^d M_k \| E[\|P w_t\|^2] \leq K E[\|P w_t\|^2]. \tag{65}$$

Let

$$F = \begin{bmatrix} I \\ 0 \end{bmatrix}, \; G = \begin{bmatrix} 2\left[(1-\eta) I + \eta C\right] & -\beta(\eta) I \\ I & 0 \end{bmatrix}, \; G_0 = \begin{bmatrix} (1-\eta) I + \eta C & -\beta(\eta) I \\ I & 0 \end{bmatrix}, \; H_t = 2\eta \begin{bmatrix} (C_t - C) P & 0 \\ 0 & 0 \end{bmatrix}.$$

From the update rule in Algorithm 2 expressed in (54), we can write

$$w_t = F^T (G + H_{t-1})(G + H_{t-2}) \cdots (G + H_1)(G_0 + H_0) F w_0.$$

Using Lemma A.3 for the expansion of $(G+H_{t-1})(G+H_{t-2})\cdots(G+H_1)(G_0+H_0)$, we have

$$Pw_t = PF^T\left(G^{t-1}G_0 + \sum_{i=1}^{t-1}\left[\prod_{j=t-1}^{i+1}(G+H_j)H_iG^{i-1}G_0\right] + \prod_{j=t-1}^{1}(G+H_j)H_0\right)Fw_0. \tag{66}$$

Since $C_0, C_1, \cdots, C_{t-1}$ are independent and identically distributed with mean C, so are $H_0, H_1, \cdots, H_{t-1}$ with mean 0. Therefore, the expectation of all cross-terms in the "square" of (66) are zero. Using the fact that $H_0Fw_0 = 0$, we have

$$E[\|Pw_t\|^2] = E[\|PF^TG^{t-1}G_0Fw_0\|^2] + \sum_{i=1}^{t-1}E\Big[\Big\|PF^T\prod_{j=t-1}^{i+1}(G+H_j)H_iG^{i-1}G_0Fw_0\Big\|^2\Big]. \tag{67}$$

Note that this result is analogous to (42) in the analysis of VR Power. From $F^TG^{t-1}G_0F = Y_t((1-\eta)I+\eta C, \beta(\eta))$ (see (20) for the definition of Y_t) and (24b) in Lemma A.1 with $w = w_0/\|w_0\|$ and the fact that $\|w_0\|^2(1-(u_1^Tw_0)^2/\|w_0\|^2) = \sum_{k=2}^d(u_k^Tw_0)^2$, we have

$$E\big[\|PF^TG^{t-1}G_0Fw_0\|^2\big] = 4p_t(\alpha_1(\eta), \beta(\eta))\cdot\sum_{k=2}^{d}E[(u_k^Tw_0)^2]. \tag{68}$$

Using Lemma A.2, $\|P\| = 1$, $H_t = 2\eta F(C_t - C)PF^T$, we have

$$\begin{aligned}E\big[\big\|PF^T\prod_{j=t-1}^{i+1}(G+H_j)H_iG^{i-1}G_0Fw_0\big\|^2\big] &\le 4\eta^2\|P\|^2\cdot E\big[\big\|F^T\prod_{j=t-1}^{i+1}(G+H_j)F(C_i-C)PF^TG^{i-1}G_0Fw_0\big\|^2\big]\\ &\le \big\|E\big[F^T\big[\prod_{j=t-1}^{i+1}(G+H_j)\big]^TFF^T\prod_{j=t-1}^{i+1}(G+H_j)F\big]\big\|\\ &\quad\cdot 4\eta^2E\big[\big\|(C_i-C)PF^TG^{i-1}G_0Fw_0\big\|^2\big].\end{aligned} \tag{69}$$

Using mathematical induction on i, we prove that

$$E\big[\big[\prod_{j=t-1}^{i+1}(G+H_j)\big]^TFF^T\prod_{j=t-1}^{i+1}(G+H_j)\big] = \sum_{\substack{(v_{i+1},\cdots,v_{t-1})\\ \in\{0,1\}^{t-i-1}}}E\big[\big[\prod_{j=t-1}^{i+1}H_j^{1-v_j}G^{v_j}\big]^TFF^T\prod_{j=t-1}^{i+1}H_j^{1-v_j}G^{v_j}\big] \tag{70}$$

for any $i \le t-2$ and fixed $t \ge 2$. Since $E[H_{t-1}] = 0$, we have

$$E[(G^T+H_{t-1}^T)FF^T(G+H_{t-1})] = G^TFF^TG + E[H_{t-1}^TFF^TH_{t-1}].$$

This proves the base case for $i = t-2$.

Suppose that (70) holds for $i = k$. Then, since H_k is independent from $H_{k+1}, \cdots, H_{t-1}$ and $E[H_k] = 0$, we have

$$\begin{aligned}E\big[\big[\prod_{j=t-1}^{k}(G+H_j)\big]^TFF^T\prod_{j=t-1}^{k}(G+H_j)\big] &= G^TE\big[\big[\prod_{j=t-1}^{k+1}(G+H_j)\big]^TFF^T\prod_{j=t-1}^{k+1}(G+H_j)\big]G\\ &\quad + E\big[H_k^T\big[\prod_{j=t-1}^{k+1}(G+H_j)\big]^TFF^T\prod_{j=t-1}^{k+1}(G+H_j)H_k\big].\end{aligned}$$

From (70), we have

$$\begin{aligned}&G^TE\big[\big[\prod_{j=t-1}^{k+1}(G+H_j)\big]^TFF^T\prod_{j=t-1}^{k+1}(G+H_j)\big]G\\ &= \sum_{\substack{(v_{k+1},\cdots,v_{t-1})\\ \in\{0,1\}^{t-k-1}}}E\big[\big[\big(\prod_{j=t-1}^{k+1}H_j^{1-v_j}G^{v_j}\big)G\big]^TFF^T\big(\prod_{j=t-1}^{k+1}H_j^{1-v_j}G^{v_j}\big)G\big].\end{aligned}$$

Also, by the independence of H_k from $H_{k+1}, \cdots, H_{t-1}$ and (70), we have

$$\begin{aligned}
&E\big[H_k^T\big[\prod_{j=t-1}^{k+1}(G+H_j)\big]^T FF^T \prod_{j=t-1}^{k+1}(G+H_j)H_k\big] \\
&= E\big[H_k^T E\big[\big[\prod_{j=t-1}^{k+1}(G+H_j)\big]^T FF^T \prod_{j=t-1}^{k+1}(G+H_j)\big]H_k\big] \\
&= E\big[H_k^T \sum_{\substack{(v_{k+1},\cdots,v_{t-1}) \\ \in\{0,1\}^{t-i-1}}} E\big[\big[\prod_{j=t-1}^{k+1} H_j^{1-v_j}G^{v_j}\big]^T FF^T \prod_{j=t-1}^{k+1} H_j^{1-v_j}G^{v_j}\big]H_k\big] \\
&= \sum_{\substack{(v_{k+1},\cdots,v_{t-1}) \\ \in\{0,1\}^{t-k-1}}} E\big[\big[\big(\prod_{j=t-1}^{k+1} H_j^{1-v_j}G^{v_j}\big)H_k\big]^T FF^T \big(\prod_{j=t-1}^{k+1} H_j^{1-v_j}G^{v_j}\big)H_k\big].
\end{aligned}$$

Therefore, we have

$$E\big[\big[\prod_{j=t-1}^{k}(G+H_j)\big]^T FF^T \prod_{j=t-1}^{k}(G+H_j)\big] = \sum_{\substack{(v_k,\cdots,v_{t-1}) \\ \in\{0,1\}^{t-k}}} E\big[\big[\prod_{j=t-1}^{k} H_j^{1-v_j}G^{v_j}\big]^T FF^T \prod_{j=t-1}^{k} H_j^{1-v_j}G^{v_j}\big],$$

which completes the proof of (70).

Using the Jensen's inequality and the norm property of a symmetric matrix, we have

$$\|E\big[F^T\big[\prod_{j=t-1}^{i+1}[H_j^{1-v_j}G^{v_j}]\big]^T FF^T \prod_{j=t-1}^{i+1}[H_j^{1-v_j}G^{v_j}]F\big]\| \le E\big[\|F^T \prod_{j=t-1}^{i+1}[H_j^{1-v_j}G^{v_j}]F\|^2\big]. \tag{71}$$

For $(v_{i+1}, \cdots, v_{t-1}) \in \{0,1\}^{t-i-1}$, let $J = \{j_1, j_2, \cdots, j_{\bar{k}}\}$ be a set of indices such that $j_1 < j_2 < \cdots < j_{\bar{k}}$ and $v_j = 0$ if $j \in J$ and $v_j = 1$ otherwise. Also, let $j_0 = i$. Using that $H_j = FF^T H_j FF^T$, we have

$$\begin{aligned}
E\big[\|F^T \prod_{j=t-1}^{i+1}[H_j^{1-v_j}G^{v_j}]F\|^2\big] &= E\big[\|F^T G^{t-j_{\bar{k}}-1}F \prod_{l=\bar{k}}^{1}\big(F^T H_{j_l} FF^T G^{j_l-j_{l-1}-1}F\big)\|^2\big] \\
&\le E\big[\|F^T G^{t-j_{\bar{k}}-1}F\|^2 \prod_{l=\bar{k}}^{1} \|F^T H_{j_l}F\|^2 \|F^T G^{j_l-j_{l-1}-1}F\|^2\big].
\end{aligned} \tag{72}$$

Since $F^T G^t F = Z_t((1-\eta)I + \eta C, \beta(\eta))$, using (24c) in Lemma A.1, we have

$$\|F^T G^t F\|^2 \le q_t(\alpha_1(\eta), \beta(\eta)). \tag{73}$$

Also, from that $F^T H_t F = 2\eta(C_t - C)P$, we have

$$E\big[\|F^T H_t F\|^2\big] \le 4\eta^2 E\big[\|(C_t - C)P\|^2\big] \le 4\eta^2 E\big[\|(C_t - C)\|^2\big] = 4\eta^2 E\big[\|(C_t - C)^2\|\big] = 4\eta^2 K. \tag{74}$$

where the last inequality follows from $\|P\| = 1$ and the second last equality follows from the symmetry of $C_t - C$. Using (73) and Lemma A.5, we have

$$\|F^T G^{t-j_{\bar{k}}-1}F\|^2 \prod_{l=\bar{k}}^{1} \|F^T G^{j_l-j_{l-1}-1}F\|^2 \le \left(\frac{1}{\alpha_1(\eta) - 4\beta(\eta)}\right)^{\bar{k}} q_{t-i-1}(\alpha_1(\eta), \beta(\eta)). \tag{75}$$

Note that there are $\bar{k}+1$ terms of the form $\|F^T G^t F\|^2$ for some $t \ge 0$ on the left-hand side of the above inequality and we use Lemma A.5 $\bar{k}$ times to obtain the term on the right-hand side.

Using (71), (72), (75), and the independence of $C_0, C_1, \cdots, C_{t-1}$, we obtain

$$\|E\big[F^T\big[\prod_{j=t-1}^{i+1}[H_j^{1-v_j}G^{v_j}]\big]^T FF^T \prod_{j=t-1}^{i+1}[H_j^{1-v_j}G^{v_j}]F\big]\| \le \left(\frac{4\eta^2 K}{\alpha_1(\eta) - 4\beta(\eta)}\right)^{\bar{k}} q_{t-i-1}(\alpha_1(\eta), \beta(\eta)).$$

Combined with (70), this results in

$$\begin{aligned}
&\|E\big[F^T\big[\prod_{j=t-1}^{i+1}(G+H_j)\big]^T FF^T\prod_{j=t-1}^{i+1}(G+H_j)F\big]\| \\
&=\|\sum_{\substack{(v_{i+1},\cdots,v_{t-1})\\ \in\{0,1\}^{t-i-1}}} E\big[F^T\big[\prod_{j=t-1}^{i+1}[H_j^{1-v_j}G^{v_j}]\big]^T FF^T\prod_{j=t-1}^{i+1}[H_j^{1-v_j}G^{v_j}]F\big]\| \\
&\le \sum_{\substack{(v_{i+1},\cdots,v_{t-1})\\ \in\{0,1\}^{t-i-1}}} \|E\big[F^T\big[\prod_{j=t-1}^{i+1}[H_j^{1-v_j}G^{v_j}]\big]^T FF^T\prod_{j=t-1}^{i+1}[H_j^{1-v_j}G^{v_j}]F\big]\| \\
&\le \sum_{\bar{k}=0}^{t-i-1}\binom{t-i-1}{\bar{k}}\left(\frac{4\eta^2K}{\alpha_1(\eta)-4\beta(\eta)}\right)^{\bar{k}} q_{t-i-1}(\alpha_1(\eta),\beta(\eta)) \\
&= q_{t-i-1}(\alpha_1(\eta),\beta(\eta))\left(1+\frac{4\eta^2K}{\alpha_1(\eta)-4\beta(\eta)}\right)^{t-i-1}.
\end{aligned} \tag{76}$$

On the other hand, using Lemma A.2 and (68) for $t=i$, we have

$$\begin{aligned}
\eta^2E\big[\big\|(C_i-C)PF^TG^{i-1}G_0Fw_0\big\|^2\big] &= \eta^2E[w_0F^TG_0^T(G^{i-1})^TFP^TE[(C_i-C)^2]PF^TG^{i-1}G_0Fw_0] \\
&\le \eta^2\|E[(C_i-C)^2]\|E[\|PF^TG^{i-1}G_0Fw_0\|^2] \\
&\le 4\eta^2K\cdot p_i(\alpha_1(\eta),\beta(\eta))\cdot\sum_{k=2}^{d}E[(u_k^Tw_0)^2].
\end{aligned} \tag{77}$$

Using (76) and (77) to bound (69), we have

$$\begin{aligned}
&E\big[\big\|PF^T\prod_{j=t-1}^{i+1}(G+H_j)H_iG^{i-1}G_0Fw_0\big\|^2\big] \\
&\quad\le 16\eta^2K\cdot p_i(\alpha_1(\eta),\beta(\eta))\cdot q_{t-i-1}(\alpha_1(\eta),\beta(\eta))\left(1+\frac{4\eta^2K}{\alpha_1(\eta)-4\beta(\eta)}\right)^{t-i-1}\cdot\sum_{k=2}^{d}E[(u_k^Tw_0)^2]
\end{aligned} \tag{78}$$

Using (68) and (78) for (67), we finally have

$$\begin{aligned}
E[\|Pw_t\|^2] \le &\left[4p_t(\alpha_1(\eta),\beta(\eta))+16\eta^2K\sum_{i=1}^{t-1}p_i(\alpha_1(\eta),\beta(\eta))\cdot q_{t-i-1}(\alpha_1(\eta),\beta(\eta))\left(1+\frac{4\eta^2K}{\alpha_1(\eta)-4\beta(\eta)}\right)^{t-i-1}\right] \\
&\cdot\sum_{k=2}^{d}E[(u_k^Tw_0)^2].
\end{aligned}$$

By (90) and (91) in Lemma A.4, we have

$$\begin{aligned}
p_t(\alpha_1(\eta),\beta(\eta)) &\le \left(\frac{\sqrt{\alpha_1(\eta)}}{2}+\frac{\sqrt{\alpha_1(\eta)-4\beta(\eta)}}{2}\right)^{2t}, \\
q_t(\alpha_1(\eta),\beta(\eta)) &\le \left(\frac{1}{\alpha_1(\eta)-\beta(\eta)}\right)\left(\frac{\sqrt{\alpha_1(\eta)}}{2}+\frac{\sqrt{\alpha_1(\eta)-4\beta(\eta)}}{2}\right)^{2(t+1)}.
\end{aligned}$$

Therefore, we obtain

$$\left[4p_t(\alpha_1(\eta),\beta(\eta)) + 16\eta^2 K \sum_{i=1}^{t-1} p_i(\alpha_1(\eta),\beta(\eta)) \cdot q_{t-i-1}(\alpha_1(\eta),\beta(\eta)) \left(1+\frac{4\eta^2 K}{\alpha_1(\eta)-4\beta(\eta)}\right)^{t-i-1}\right]$$
$$\leq 4\left[1+\frac{4\eta^2 K}{\alpha_1(\eta)-4\beta(\eta)}\sum_{i=1}^{t-1}\left(1+\frac{4\eta^2 K}{\alpha_1(\eta)-4\beta(\eta)}\right)^{t-i-1}\right]\cdot\left(\frac{\sqrt{\alpha_1(\eta)}}{2}+\frac{\sqrt{\alpha_1(\eta)-4\beta(\eta)}}{2}\right)^{2t}$$
$$= 4\left(1+\frac{4\eta^2 K}{\alpha_1(\eta)-4\beta(\eta)}\right)^{t-1}\cdot\left(\frac{\sqrt{\alpha_1(\eta)}}{2}+\frac{\sqrt{\alpha_1(\eta)-4\beta(\eta)}}{2}\right)^{2t},$$

which results in

$$E[\|Pw_t\|^2] \leq 4\left(1+\frac{4\eta^2 K}{\alpha_1(\eta)-4\beta(\eta)}\right)^{t-1}\cdot\left(\frac{\sqrt{\alpha_1(\eta)}}{2}+\frac{\sqrt{\alpha_1(\eta)-4\beta(\eta)}}{2}\right)^{2t}\cdot\sum_{k=2}^{d}E[(u_k^T w_0)^2].$$

Finally, from (65), we have

$$\sum_{k=2}^{d}E[w_t^T P M_k P w_t] \leq 4K\left(1+\frac{4\eta^2 K}{\alpha_1(\eta)-4\beta(\eta)}\right)^{t-1}\cdot\left(\frac{\sqrt{\alpha_1(\eta)}}{2}+\frac{\sqrt{\alpha_1(\eta)-4\beta(\eta)}}{2}\right)^{2t}\cdot\sum_{k=2}^{d}E[(u_k^T w_0)^2]. \tag{79}$$

This completes the proof of the first statement.

Next, from $\alpha_2(\eta) = 4\beta(\eta) \geq \alpha_k(\eta)$ for $k \geq 2$ and (92) in Lemma A.4,

$$\sum_{k=2}^{d}p_m(\alpha_k(\eta),\beta(\eta))E[(u_k^T w_0)^2] \leq p_m(\alpha_2(\eta),\beta(\eta))\cdot\sum_{k=2}^{d}E[(u_k^T w_0)^2]. \tag{80}$$

Also, using (91) and (92) in Lemma A.4 and (79), we have

$$4\eta^2\sum_{k=2}^{d}\sum_{r=1}^{m-1}q_{m-r-1}(\alpha_k(\eta),\beta(\eta))E[w_r^T P M_k P w_r]$$
$$\leq \frac{16\eta^2 K}{\alpha_1(\eta)-4\beta(\eta)}\sum_{r=1}^{m-1}\left(1+\frac{4\eta^2 K}{\alpha_1(\eta)-4\beta(\eta)}\right)^{r-1}\cdot\left(\frac{\sqrt{\alpha_1(\eta)}}{2}+\frac{\sqrt{\alpha_1(\eta)-4\beta(\eta)}}{2}\right)^{2m}\cdot\sum_{k=2}^{d}E[(u_k^T w_0)^2]$$
$$\leq 4\left[\left(1+\frac{4\eta^2 K}{\alpha_1(\eta)-4\beta(\eta)}\right)^{m-1}-1\right]\cdot\left(\frac{\sqrt{\alpha_1(\eta)}}{2}+\frac{\sqrt{\alpha_1(\eta)-4\beta(\eta)}}{2}\right)^{2m}\cdot\sum_{k=2}^{d}E[(u_k^T w_0)^2].$$

Since $0 < \dfrac{4\eta^2 Km}{\alpha_1(\eta)-\beta(\eta)} < 1$, using that $\exp(x) \leq 1+2x$ for $x \in [0,1]$ we have

$$\left(1+\frac{4\eta^2 K}{\alpha_1(\eta)-4\beta(\eta)}\right)^{m-1}-1 \leq \left(1+\frac{4\eta^2 K}{\alpha_1(\eta)-4\beta(\eta)}\right)^{m}-1 \leq \exp\left(\frac{4\eta^2 Km}{\alpha_1(\eta)-4\beta(\eta)}\right)-1 \leq \frac{8\eta^2 Km}{\alpha_1(\eta)-4\beta(\eta)},$$

leading to

$$4\eta^2\sum_{k=2}^{d}\sum_{r=1}^{m-1}q_{m-r-1}(\alpha_k(\eta),\beta(\eta))E[w_r^T P M_k P w_r] \leq \frac{32\eta^2 Km}{\alpha_1(\eta)-4\beta(\eta)}\cdot\left(\frac{\sqrt{\alpha_1(\eta)}}{2}+\frac{\sqrt{\alpha_1(\eta)-4\beta(\eta)}}{2}\right)^{2m}\cdot\sum_{k=2}^{d}E[(u_k^T w_0)^2]. \tag{81}$$

Using (80), (81) for Lemma 3.5, we finally have

$$\sum_{k=2}^{d}E[(u_k^T w_m)^2] \leq \left[p_m(\alpha_2(\eta),\beta(\eta))+\frac{32\eta^2 Km}{\alpha_1(\eta)-4\beta(\eta)}\cdot\left(\frac{\sqrt{\alpha_1(\eta)}}{2}+\frac{\sqrt{\alpha_1(\eta)-4\beta(\eta)}}{2}\right)^{2m}\right]\cdot\sum_{k=2}^{d}E[(u_k^T w_0)^2]. \tag{82}$$

Lastly, using Lemma 3.5 for $k = 1$, we have

$$E[(u_1^T w_m)^2] = p_m(\alpha_1(\eta), \beta(\eta))E[(u_1^T w_0)^2] + 4\eta^2 \sum_{r=1}^{m-1} q_{m-r-1}(\alpha_1(\eta), \beta(\eta))E[w_r^T PM_1 P w_r].$$

Since PM_kP is positive semi-definite and $q_t(\alpha_1(\eta), \beta(\eta)) \geq 0$ for $1 \leq t < m$ by (91) in Lemma A.4, we have

$$E[(u_1^T w_m)^2] \geq p_m(\alpha_1(\eta), \beta(\eta))E[(u_1^T w_0)^2]. \tag{83}$$

Also, from $\alpha_1(\eta) > \alpha_2(\eta) = 4\beta(\eta)$ and (90) in Lemma A.4, we have

$$p_m(\alpha_1(\eta), \beta(\eta)) \geq \frac{1}{4}\left(\frac{\sqrt{\alpha_1(\eta)}}{2} + \frac{\sqrt{\alpha_1(\eta) - 4\beta(\eta)}}{2}\right)^{2m}. \tag{84}$$

Using (82), (83) and (84), we eventually obtain

$$\frac{\sum_{k=2}^d E[(u_k^T w_m)^2]}{E[(u_1^T w_m)^2} \leq \left[\frac{p_m(\alpha_2(\eta), \beta(\eta))}{p_m(\alpha_1(\eta), \beta(\eta))} + \frac{128\eta^2 Km}{\alpha_1(\eta) - 4\beta(\eta)}\right] \cdot \frac{\sum_{k=2}^d E[(u_k^T w_0)^2]}{E[(u_1^T w_0)^2},$$

which completes the proof. □

Proof of Lemma 3.7. Using the conditions on m and $|S|$, we have

$$0 \leq \frac{4\eta^2 Km}{\alpha_1(\eta) - 4\beta(\eta)} \leq \frac{1}{128}. \tag{85}$$

Also, from

$$p_m(\alpha_2(\eta), \beta(\eta)) = (\beta(\eta))^m, \quad p_m(\alpha_1(\eta), \beta(\eta)) \geq \frac{1}{4}\left(\frac{\sqrt{\alpha_1(\eta)}}{2} + \frac{\sqrt{\alpha_1(\eta) - 4\beta(\eta)}}{2}\right)^{2m}.$$

and the choice of and m, we have

$$\begin{aligned}
\frac{p_m(\alpha_2(\eta), \beta(\eta))}{p_m(\alpha_1(\eta), \beta(\eta))} &\leq 4 \cdot \left(\frac{\sqrt{4\beta(\eta)}}{\sqrt{\alpha_1(\eta)} + \sqrt{\alpha_1(\eta) - 4\beta(\eta)}}\right)^{2m} \\
&= 4 \cdot \left(1 - \frac{\sqrt{\alpha_1(\eta)} - \sqrt{4\beta(\eta)} + \sqrt{\alpha_1(\eta) - 4\beta(\eta)}}{\sqrt{\alpha_1(\eta)} + \sqrt{\alpha_1(\eta) - 4\beta(\eta)}}\right)^{2m} \\
&= 4 \cdot \left(1 - \frac{\eta\lambda_1\Delta + \sqrt{\eta\lambda_1\Delta(2(1-\eta) + \eta(\lambda_1 + \lambda_2))}}{1 - \eta + \eta\lambda_1 + \sqrt{\eta\lambda_1\Delta(2(1-\eta) + \eta(\lambda_1 + \lambda_2))}}\right)^{2m} \\
&\leq 4 \cdot \exp\left(-2\frac{\eta\lambda_1\Delta + \sqrt{\eta\lambda_1\Delta(2(1-\eta) + \eta(\lambda_1 + \lambda_2))}}{1 - \eta + \eta\lambda_1 + \sqrt{\eta\lambda_1\Delta(2(1-\eta) + \eta(\lambda_1 + \lambda_2))}} m\right) \\
&\leq \frac{1}{2}.
\end{aligned} \tag{86}$$

Therefore, using (85) and (86) in Lemma 3.6, we finally have

$$\frac{\sum_{k=2}^d E[(u_k^T w_m)^2]}{E[(u_1^T w_m)^2]} \leq \left(\frac{p_m(\alpha_2(\eta), \beta(\eta))}{p_m(\alpha_1(\eta), \beta(\eta))} + \frac{128\eta^2 Km}{\alpha_1(\eta) - 4\beta(\eta)}\right)\left(\frac{\sum_{k=2}^d E[(u_k^T w_0)^2]}{E[(u_1^T w_0)^2]}\right) \leq \frac{3}{4}\left(\frac{\sum_{k=2}^d E[(u_k^T w_0)^2]}{E[(u_1^T w_0)^2]}\right),$$

which completes the proof. □

Proof of Theorem 3.8. By repeatedly applying Lemma 3.7, we have

$$\frac{\sum_{k=2}^{d} E[(u_k^T \tilde{w}_\tau)^2]}{E[(u_1^T \tilde{w}_\tau)^2]} \leq \left(\frac{3}{4}\right)^\tau \frac{\sum_{k=2}^{d} E[(u_k^T \tilde{w}_0)^2]}{E[(u_1^T \tilde{w}_0)^2]} = \left(\frac{3}{4}\right)^\tau \tilde{\theta}_0.$$

Since $\tau = \lceil \log(\tilde{\theta}_0/\epsilon)/\log(4/3) \rceil$, we have

$$\tau \log\left(\frac{3}{4}\right) \leq \log\left(\frac{\epsilon}{\tilde{\theta}_0}\right),$$

resulting in

$$\frac{\sum_{k=2}^{d} E[(u_k^T \tilde{w}_\tau)^2]}{E[(u_1^T \tilde{w}_\tau)^2]} \leq \epsilon.$$

□

A.2 Technical Lemmas

Lemma A.2. *Let w be a vector in $\mathbb{R}^d$ and let M be a $d \times d$ symmetric matrix. Then, we have*

$$w^T M w \leq \|M\| \|w\|^2.$$

Proof. By the cyclic property of the trace, we have

$$w^T M w = \mathrm{Tr}[w^T M w] = \mathrm{Tr}[M w w^T].$$

Since ww^T is positive semi-definite, we have

$$\mathrm{Tr}[M w w^T] \leq \|M\| \mathrm{Tr}[w w^T].$$

Again, by the cyclic property of the trace, we finally have

$$w^T M w \leq \|M\| \mathrm{Tr}[w w^T] = \|M\| \mathrm{Tr}[w^T w] = \|M\| \|w\|^2.$$

□

Lemma A.3. *Let A_i and B_i be $d \times d$ matrices for $i = 0, \cdots, t-1$. Then, we have*

$$\prod_{i=t-1}^{0} (A_i + B_i) = (A_{t-1} + B_{t-1})(A_{t-2} + B_{t-2}) \cdots (A_0 + B_0) = \prod_{i=t-1}^{0} A_i + \sum_{i=0}^{t-1} \left[\prod_{j=t-1}^{i+1} (A_j + B_j) B_i \prod_{k=i-1}^{0} A_k \right]. \quad (87)$$

Proof. We prove the statement by induction. For $t = 1$, we have

$$\prod_{i=0}^{0} A_i + \sum_{i=0}^{0} \left[\prod_{j=0}^{i+1} (A_j + B_j) B_i \prod_{k=i-1}^{0} A_k \right] = A_0 + \left[\prod_{j=0}^{1} (A_j + B_j) B_0 \prod_{k=-1}^{0} A_k \right] = A_0 + B_0,$$

which proves the base case. Next, suppose that we have (87) for $t-2$. Then, we have

$$\begin{aligned}
\prod_{i=t-1}^{0} (A_i + B_i) &= (A_{t-1} + B_{t-1}) \prod_{i=t-2}^{0} (A_i + B_i) \\
&= (A_{t-1} + B_{t-1}) \left(\prod_{i=t-2}^{0} A_i + \sum_{i=0}^{t-2} \left[\prod_{j=t-2}^{i+1} (A_j + B_j) B_i \prod_{k=i-1}^{0} A_k \right] \right) \\
&= \prod_{i=t-1}^{0} A_i + B_{t-1} \prod_{i=t-2}^{0} A_i + \left(\sum_{i=0}^{t-2} \left[\prod_{j=t-1}^{i+1} (A_j + B_j) B_i \prod_{k=i-1}^{0} A_k \right] \right) \\
&= \prod_{i=t-1}^{0} A_i + \sum_{i=0}^{t-1} \left[\prod_{j=t-1}^{i+1} (A_j + B_j) B_i \prod_{k=i-1}^{0} A_k \right].
\end{aligned}$$

This completes the proof. □

Lemma A.4. *Let x_t be a sequence of real numbers such that*

$$x_t = (\alpha - \beta) x_{t-1} - \beta(\alpha - \beta) x_{t-2} + \beta^3 x_{t-3} + L_{t-1} + \beta L_{t-2}$$

for $t \geq 3$ and $x_0 = L_0$, $x_1 = \frac{\alpha}{4} L_0$, $x_2 = \left(\frac{\alpha}{2} - \beta\right)^2 L_0 + L_1$. Then, we have

$$x_t = p_t(\alpha, \beta) L_0 + \sum_{r=1}^{t-1} q_{t-r-1}(\alpha, \beta) L_r. \quad (88)$$

Moreover, for $t \geq 0$, we have

- *if* $0 \le \alpha = 4\beta$,

$$p_t(4\beta, \beta) = \beta^t \ge 0, \quad q_t(4\beta, \beta) = (t+1)^2 \beta^t \ge 0, \tag{89}$$

- *if* $0 \le 4\beta < \alpha$,

$$p_t(\alpha, \beta) = \left[\frac{1}{2}\left(\frac{\sqrt{\alpha}}{2} + \frac{\sqrt{\alpha - 4\beta}}{2}\right)^t + \frac{1}{2}\left(\frac{\sqrt{\alpha}}{2} - \frac{\sqrt{\alpha - 4\beta}}{2}\right)^t\right]^2 > p_t(4\beta, \beta) \ge 0, \tag{90}$$

$$q_t(\alpha, \beta) = \frac{1}{\alpha - 4\beta}\left[\left(\frac{\sqrt{\alpha}}{2} + \frac{\sqrt{\alpha - 4\beta}}{2}\right)^{t+1} - \left(\frac{\sqrt{\alpha}}{2} - \frac{\sqrt{\alpha - 4\beta}}{2}\right)^{t+1}\right]^2 > q_t(4\beta, \beta) \ge 0, \tag{91}$$

- *if* $0 \le \alpha < 4\beta$,

$$p_t(\alpha, \beta) \le p_t(4\beta, \beta), \quad q_t(\alpha, \beta) \le q_t(4\beta, \beta). \tag{92}$$

Proof. It is easy to check that x_0, x_1, and x_2 satisfy (88). Suppose that (88) holds for $t-1, t-2, t-3$. Then, we have

$$\begin{aligned}
x_t &= (\alpha - \beta)x_{t-1} - \beta(\alpha - \beta)x_{t-2} + \beta^3 x_{t-3} + L_{t-1} + \beta L_{t-2} \\
&= p_t(\alpha, \beta)L_0 + L_{t-1} + \alpha L_{t-2} + (\alpha - \beta)^2 L_{t-3} + \sum_{r=1}^{t-4} q_{t-r-1}(\alpha, \beta)L_r \\
&= p_t(\alpha, \beta)L_0 + \sum_{r=1}^{t-1} q_{t-r-1}(\alpha, \beta)L_r.
\end{aligned}$$

Therefore, (88) holds by induction.

Next, we prove (89), (90), (91) and (92). The characteristic equation of (9) is

$$r^3 - (\alpha - \beta)r^2 + \beta(\alpha - \beta)r - \beta^3 = 0. \tag{93}$$

If $0 \le \alpha = 4\beta$, (93) has a cube root of $r = \beta$. From initial conditions (11) and (12), we obtain

$$p_t(4\beta, \beta) = \beta^t \ge 0, \quad q_t(4\beta, \beta) = (t+1)^2 \beta^t \ge 0. \tag{94}$$

If $0 \le 4\beta < \alpha$, the roots of (93) are

$$r = \beta, \frac{\alpha - 2\beta}{2} + \frac{\sqrt{\alpha^2 - 4\alpha\beta}}{2}, \frac{\alpha - 2\beta}{2} - \frac{\sqrt{\alpha^2 - 4\alpha\beta}}{2}.$$

With initial conditions (11), we obtain

$$p_t(\alpha, \beta) = \frac{1}{4}\left(\frac{\alpha - 2\beta}{2} + \frac{\sqrt{\alpha^2 - 4\alpha\beta}}{2}\right)^t + \frac{1}{4}\left(\frac{\alpha - 2\beta}{2} - \frac{\sqrt{\alpha^2 - 4\alpha\beta}}{2}\right)^t + \frac{1}{2}\beta^t$$

Using the fact that $\alpha > 4\beta$ and the arithmetic-geometric mean inequality, we have

$$p_t(\alpha, \beta) > \beta^t \ge 0.$$

Moreover, we can further write $p_t(\alpha, \beta)$ as

$$p_t(\alpha, \beta) = \left[\frac{1}{2}\left(\frac{\sqrt{\alpha}}{2} + \frac{\sqrt{\alpha - 4\beta}}{2}\right)^t + \frac{1}{2}\left(\frac{\sqrt{\alpha}}{2} - \frac{\sqrt{\alpha - 4\beta}}{2}\right)^t\right]^2$$

by expanding this expression.

On the other hand, using (12), we have

$$\begin{aligned} q_t(\alpha, \beta) &= \frac{1}{\alpha - 4\beta}\left[\left(\frac{\alpha - 2\beta}{2} + \frac{\sqrt{\alpha^2 - 4\alpha\beta}}{2}\right)^{t+1} + \left(\frac{\alpha - 2\beta}{2} - \frac{\sqrt{\alpha^2 - 4\alpha\beta}}{2}\right)^{t+1} - 2\beta^{t+1}\right] \\ &= \frac{1}{\alpha - 4\beta}\left[\left(\frac{\sqrt{\alpha}}{2} + \frac{\sqrt{\alpha - 4\beta}}{2}\right)^{t+1} - \left(\frac{\sqrt{\alpha}}{2} - \frac{\sqrt{\alpha - 4\beta}}{2}\right)^{t+1}\right]^2 \geq 0. \end{aligned}$$

Using the fact that $A^{t+1} - B^{t+1} = (A - B)(A^t + A^{t-1}B + \cdots + B^t)$ for any $A, B \in \mathbb{R}$, we have

$$q_t(\alpha, \beta) = \left[\sum_{i=0}^{t}\left(\frac{\sqrt{\alpha}}{2} + \frac{\sqrt{\alpha - 4\beta}}{2}\right)^i\left(\frac{\sqrt{\alpha}}{2} - \frac{\sqrt{\alpha - 4\beta}}{2}\right)^{t-i}\right]^2.$$

Again, using the arithmetic-geometric mean inequality and the fact that $\alpha > 4\beta$, we have

$$q_t(\alpha, \beta) \geq \left[(t+1)\left(\frac{\sqrt{\alpha}}{2} + \frac{\sqrt{\alpha - 4\beta}}{2}\right)^{t/2}\left(\frac{\sqrt{\alpha}}{2} - \frac{\sqrt{\alpha - 4\beta}}{2}\right)^{t/2}\right]^2 = (t+1)^2\beta^t = q_t(4\beta, \beta).$$

If $0 \leq \alpha < 4\beta$, the roots of (93) are

$$r = \beta, \frac{\alpha - 2\beta}{2} + \frac{\sqrt{4\alpha\beta - \alpha^2}}{2}i, \frac{\alpha - 2\beta}{2} - \frac{\sqrt{4\alpha\beta - \alpha^2}}{2}i.$$

Setting

$$\cos\,\theta_p = \frac{\alpha - 2\beta}{2\beta}, \quad \sin\,\theta_p = \frac{\sqrt{4\alpha\beta - \alpha^2}}{2\beta}$$

it is easy to verify that

$$\begin{aligned} p_t(\alpha, \beta) &= \frac{1}{4}\beta^t\Big[\cos\,\theta_p + i\,\sin\,\theta_p\Big]^t + \frac{1}{4}\beta^t\Big[\cos\,\theta_p - i\,\sin\,\theta_p\Big]^t + \frac{1}{2}\beta^t \\ &= \frac{1}{4}(e^{i\theta t} + e^{-i\theta t})\beta^t + \frac{1}{2}\beta^t \\ &= \frac{1}{4}|e^{i\theta t} + e^{-i\theta t}|\beta^t + \frac{1}{2}\beta^t \\ &\leq \frac{1}{4}(|e^{i\theta t}| + |e^{-i\theta t}|)\beta^t + \frac{1}{2}\beta^t \\ &= \beta^t. \end{aligned}$$

Moreover, with

$$\cos\,\theta_q = \frac{\alpha - 2\beta}{2\beta}, \quad \sin\,\theta_q = \frac{\sqrt{4\alpha\beta - \alpha^2}}{2\beta}, \quad \cos\,\phi_q = 1 - \frac{\alpha}{2\beta}, \quad \sin\,\phi_q = -\frac{\sqrt{4\alpha\beta - \alpha^2}}{2\beta},$$

it can be seen by using elementary calculus that

$$q_t(\alpha, \beta) = \left[\frac{2\beta}{4\beta - \alpha} + \frac{2\beta}{4\beta - \alpha}\cos(\phi_q + t\theta_q)\right]\beta^t. \tag{95}$$

Let

$$Q(t) = \frac{q_t(4\beta, \beta) - q_t(\alpha, \beta)}{\beta^t}.$$

Then, from (9) and (11), we have

$$Q(0) = 0, \quad Q(1) = \frac{4\beta - \alpha}{\beta}, \quad Q(2) = \frac{(4\beta - \alpha)(2\beta + \alpha)}{\beta^2}, \quad Q(3) = \frac{(\alpha^2 + 4\beta^2)(4\beta - \alpha)}{\beta^3} \tag{96}$$

resulting in

$$Q(2) - Q(0) = \frac{(4\beta-\alpha)(2\beta+\alpha)}{\beta^2} \geq 0, \quad Q(3) - Q(1) = \frac{(\alpha^2+3\beta^2)(4\beta-\alpha)}{\beta^3} \geq 0. \tag{97}$$

In order to show $Q(t) \geq 0$ for $t \geq 0$, we prove $Q(t+2) - Q(t) \geq 0$ for $t \geq 0$. Using (94), (95) and standard trigonometric equalities, it follows that

$$Q(t+2) - 2Q(t) + Q(t-2) = 8 + \frac{2\alpha}{\beta}\cos(\phi_q + t\theta_q).$$

In turn, we have

$$\begin{aligned}
Q(t+2) - Q(t) &= Q(t) - Q(t-2) + 8 + \frac{2\alpha}{\beta}\cos(\phi_q + t\theta_q) \\
&\geq Q(t) - Q(t-2) + 8 - \frac{2\alpha}{\beta} \\
&= Q(t) - Q(t-2) + \frac{2(4\beta-\alpha)}{\beta} \\
&\geq Q(t) - Q(t-2).
\end{aligned} \tag{98}$$

From (96), (97), and (98), for $t \geq 0$, we obtain $Q(t) \geq 0$ implying

$$q_t(\alpha,\beta) \leq q_t(4\beta,\beta).$$

□

Lemma A.5. *If $\alpha > 4\beta \geq 0$, then for $0 \leq t_1 < t_2$, we have*

$$q_{t_1}(\alpha,\beta) \cdot q_{t_2}(\alpha,\beta) \leq \left(\frac{1}{\alpha-4\beta}\right) q_{t_1+t_2+1}(\alpha,\beta).$$

Proof. From (91) in Lemma A.4, we have

$$\begin{aligned}
q_{t_1}(\alpha,\beta) \cdot q_{t_2}(\alpha,\beta) = \left(\frac{1}{\alpha-4\beta}\right)^2 &\left[\left(\frac{\sqrt{\alpha}}{2} + \frac{\sqrt{\alpha-4\beta}}{2}\right)^{t_1+1} - \left(\frac{\sqrt{\alpha}}{2} - \frac{\sqrt{\alpha-4\beta}}{2}\right)^{t_1+1}\right]^2 \\
&\cdot \left[\left(\frac{\sqrt{\alpha}}{2} + \frac{\sqrt{\alpha-4\beta}}{2}\right)^{t_2+1} - \left(\frac{\sqrt{\alpha}}{2} - \frac{\sqrt{\alpha-4\beta}}{2}\right)^{t_2+1}\right]^2.
\end{aligned}$$

Since

$$0 \leq \frac{\sqrt{\alpha}}{2} - \frac{\sqrt{\alpha-4\beta}}{2} < \frac{\sqrt{\alpha}}{2} + \frac{\sqrt{\alpha-4\beta}}{2},$$

we have

$$\begin{aligned}
&\left[\left(\frac{\sqrt{\alpha}}{2} + \frac{\sqrt{\alpha-4\beta}}{2}\right)^{t_1+1} - \left(\frac{\sqrt{\alpha}}{2} - \frac{\sqrt{\alpha-4\beta}}{2}\right)^{t_1+1}\right]\left[\left(\frac{\sqrt{\alpha}}{2} + \frac{\sqrt{\alpha-4\beta}}{2}\right)^{t_2+1} - \left(\frac{\sqrt{\alpha}}{2} - \frac{\sqrt{\alpha-4\beta}}{2}\right)^{t_2+1}\right] \\
&= \left(\frac{\sqrt{\alpha}}{2} + \frac{\sqrt{\alpha-4\beta}}{2}\right)^{t_1+t_2+2} - \left(\frac{\sqrt{\alpha}}{2} - \frac{\sqrt{\alpha-4\beta}}{2}\right)^{t_1+1}\left(\frac{\sqrt{\alpha}}{2} + \frac{\sqrt{\alpha-4\beta}}{2}\right)^{t_2+1} \\
&\quad - \left(\frac{\sqrt{\alpha}}{2} + \frac{\sqrt{\alpha-4\beta}}{2}\right)^{t_1+1}\left(\frac{\sqrt{\alpha}}{2} - \frac{\sqrt{\alpha-4\beta}}{2}\right)^{t_2+1} + \left(\frac{\sqrt{\alpha}}{2} - \frac{\sqrt{\alpha-4\beta}}{2}\right)^{t_1+t_2+2} \\
&\leq \left(\frac{\sqrt{\alpha}}{2} + \frac{\sqrt{\alpha-4\beta}}{2}\right)^{t_1+t_2+2} - \left(\frac{\sqrt{\alpha}}{2} - \frac{\sqrt{\alpha-4\beta}}{2}\right)^{t_1+t_2+2}.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
q_{t_1}(\alpha,\beta) \cdot q_{t_2}(\alpha,\beta) &\leq \left(\frac{1}{\alpha-4\beta}\right)^2 \left[\left(\frac{\sqrt{\alpha}}{2} + \frac{\sqrt{\alpha-4\beta}}{2}\right)^{t_1+t_2+2} - \left(\frac{\sqrt{\alpha}}{2} - \frac{\sqrt{\alpha-4\beta}}{2}\right)^{t_1+t_2+2}\right]^2 \\
&= \left(\frac{1}{\alpha-4\beta}\right) q_{t_1+t_2+1}(\alpha,\beta).
\end{aligned}$$

This completes the proof. □