Accelerated Stochastic Power Iteration

Peng Xu[1] **Bryan He**[1] **Christopher De Sa**[2] **Ioannis Mitliagkas**[3] **Christopher Ré**[1]
[1]Stanford University [2]Cornell University [3]University of Montréal

Abstract

Principal component analysis (PCA) is one of the most powerful tools for analyzing matrices in machine learning. In this paper, we study methods to accelerate power iteration in the stochastic setting by adding a momentum term. While in the deterministic setting, power iteration with momentum has optimal iteration complexity, we show that naively adding momentum to a stochastic method does not always result in acceleration. We perform a novel, tight variance analysis that reveals a "breaking-point variance" beyond which this acceleration does not occur. Combining this insight with modern variance reduction techniques yields a simple version of power iteration with momentum that achieves the optimal iteration complexities in both the online and offline setting. Our methods are embarrassingly parallel and can produce wall-clock-time speedups. Our approach is very general and applies to many non-convex optimization problems that can now be accelerated using the same technique.

1 Introduction

Principal Component Analysis (PCA) is a fundamental tool for data processing and visualization in machine learning and statistics [Hot33; Jol02]. PCA captures variable interactions in a high-dimensional dataset by identifying the directions of highest variance: the *principal components*. Modern machine learning problems have become too big for full-pass PCA methods, leading practitioners to *stochastic methods*: algorithms that only ingest a random subset of the available data at every iteration. Stochastic methods have been proposed for the *offline*, or *finite-sample* setting, in which the algorithm is given random access to a finite set of samples, and therefore could perform a full-pass periodically [Sha15]. Other methods target the *online* setting, in which the samples are randomly drawn from a distribution, and full passes are not possible [MCJ13; Bou+15; Jai+16]. Information theoretic bounds [AZL16b] show that the *sample complexity*, which is defined as the number of samples necessary to recover the principal component, is at least $\mathcal{O}(1/\Delta^2)$ in the online setting, where Δ is the eigen-gap. Elegant variants of the power method have been shown to match this lower bound [Jai+16; AZL16b]. However, sample complexity is not a great proxy for run time.

Iteration complexity—the number of outer loop iterations required, when the inner loop is embarrassingly parallel—provides an asymptotic measure of an algorithm's performance on a highly parallel computer. We would like to match the Lanczos algorithm's optimal convergence rate of $\mathcal{O}(1/\sqrt{\Delta})$ iterations from the full-pass setting. Unfortunately, the Lanczos algorithm cannot operate in a stochastic setting and none of the simple stochastic power iteration variants achieve this accelerated iteration complexity. Recently, carefully tuned numerical methods based on approximate matrix inversion [Gar+16; AZL16a] have achieved an accelerated rate in the stochastic setting. However, these methods are significantly more complex than stochastic power iteration and are largely theoretical in nature. This context motivates the question: *is it possible to achieve the optimal sample and iteration complexity with a method as simple as power iteration?*

In this paper, we propose a class of simple PCA algorithms based on the power method that (1) operate in the stochastic setting, (2) have a sample complexity with an asymptotically optimal dependence on the eigen-gap, and (3) have an iteration complexity with an asymptotically optimal dependence on the eigen-gap (i.e. one that matches the worst-case rate for the Lanczos method). As background for our method, we first note that a simple modification of the power iteration, *power iteration with momentum*, achieves the

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

optimal accelerated convergence rate $\mathcal{O}(1/\sqrt{\Delta})$ in the deterministic setting. Our proposed algorithms come from the natural idea of designing an efficient, stochastic version of that method.

We demonstrate that simply adding momentum to a stochastic method like Oja's does not always result in acceleration. Although adding momentum with a fixed learning rate accelerates the initial convergence, it also increases the size of the noise ball. This is because momentum accelerates the convergence of the expected iterates, but also increases the variance, which typically dominates, so no overall acceleration is observed (cf. Section 3). Using Chebyshev polynomials to derive an exact expression for the variance of the iterates of our algorithm, we identify the precise relationship between *sample variance* and *acceleration*. Importantly, we identify the exact break-down point beyond which the variance is too much for acceleration to happen.

Based on this analysis, we can design a stochastic momentum power method that is guaranteed to work. We first propose a *mini-batching* technique to ensure acceleration. However, this algorithm requires increasingly large batch sizes to reach small errors. To remedy this, we additionally propose a *variance reduction* technique, which ensures acceleration with a constant batch size. Both of these techniques are used to speed up computation in stochastic optimization and are embarrassingly parallel. This property allows our method to achieve true *wall-clock time acceleration* even in the online setting, something not possible with state-of-the-art results. Hence, we demonstrate that the more complicated techniques based on approximate matrix inversion are not necessary: *simple momentum-based methods are sufficient to accelerate PCA*.

Our tight variance analysis can apply to general stochastic three-term recurrence (cf. Section 4). It is straightforward to show that randomized Kaczmarz algorithms [SV09; GR15] can be accelerated using momentum for solving linear systems. It also enables many non-convex problems, including matrix completion [JNS13], phase retrieval [CLS15] and subspace tracking [BNR10], to be accelerated using a single technique, and suggests that the same might be true for a larger class of non-convex optimization problems.

Our contributions

- We study the relationship between variance and acceleration by finding an exact characterization of variance for a general class of power iteration variants with momentum in Section 3.1.
- Using this bound, we design an algorithm that uses mini-batching to obtain the optimal iteration and sample complexities for the online setting in Section 3.2.
- We design a second algorithm that uses variance reduction to obtain the optimal rate for the offline setting in Section 3.3. Notably, when operating in the offline setting, we are able to use a batch size that is independent of the target accuracy.
- We demonstrate the acceleration of our variance-reduced methods on real-world network datasets in Section 3.4.

2 Power method with momentum

In this section, we provide background knowledge to help introduce our stochastic method. We begin by describing the basic PCA setup and show that a simple momentum scheme accelerates the standard power method. This momentum scheme, and its connection with the Chebyshev polynomial family, serves as the foundation of our stochastic method.

PCA Let $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d$ be n data points. The goal of PCA is to find the top eigenvector of the symmetric positive semidefinite (PSD) matrix $\mathbf{A} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T \in \mathbb{R}^{d\times d}$ (the sample covariance matrix) when the data points are centered at the origin. We assume that the target matrix $\mathbf{A}$ has eigenvalues $1 \geq \lambda_1 > \lambda_2 \geq \cdots \lambda_d \geq 0$ with corresponding normalized eigenvectors $\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_d$. The power method estimates the top eigenvector by repeatedly applying the update step

$$\mathbf{w}_{t+1} = \mathbf{A}\mathbf{w}_t$$

with an initial vector $\mathbf{w}_0 \in \mathbb{R}^d$. After $\mathcal{O}\left(\frac{1}{\Delta}\log\frac{1}{\epsilon}\right)$ steps, the normalized iterate $\mathbf{w}_t/\|\mathbf{w}_t\|$[1] is an ϵ-*accurate estimate* of top principal component. Here ϵ accuracy is measured by the squared sine of the angle between $\mathbf{u}_1$ and $\mathbf{w}_t$, which is $1 - (\mathbf{u}_1^T\mathbf{w}_t)^2/\|\mathbf{w}_t\|^2$.

When λ_1 is close to λ_2 (the eigengap Δ is small), the power method will converge very slowly. To address this, we propose a class of algorithms based on the alternative update step

$$\mathbf{w}_{t+1} = \mathbf{A}\mathbf{w}_t - \beta\mathbf{w}_{t-1}. \qquad (\mathbf{A})$$

We call the extra term, $\beta\mathbf{w}_{t-1}$, the *momentum* term, and β the momentum parameter, in analogy to the heavy ball method [Pol64], which uses the same technique to address poorly conditioned problems in convex optimization. For appropriate settings of β, this *accelerated power method* can converge dramatically faster than the traditional power method; this is not surprising, since the same is true for analogous accelerated methods for convex optimization.

Orthogonal polynomials We now connect the dynamics of the update $(\mathbf{A})$ to the behavior of a family of

[1] The $\|\cdot\|$ in this paper is ℓ_2 norm for vectors and spectral norm for matrices.

Table 1: Asymptotic complexities for variants of the power method to achieve ϵ accuracy, $1-(\mathbf{u}_1^T\mathbf{w})^2 \leq \epsilon$. For momentum methods, we choose the optimal $\beta = \lambda_2^2/4$. Here $\Delta := \lambda_1 - \lambda_2$ is the eigen-gap, σ^2 is the variance of one random sample and r is an a.s. norm bound (see Definition (1)). In $\mathcal{O}$ notation, we omit the factors depending on failure probability δ. Jain et al. [Jai+16] and Shamir [Sha15] give the best known results for stochastic PCA without and with variance reduction respectively. However, neither of these results achieve the optimal iteration complexity. Furthermore, they are not tight in terms of the variance of the problem (i.e. when σ^2 is small, the bounds are loose).

Setting	Algorithm	Number of Iterations	Batch Size	Reference
Deterministic	Power	$\mathcal{O}\left(\frac{1}{\Delta}\cdot\log\left(\frac{1}{\epsilon}\right)\right)$	n	[GVL12]
	Lanczos	$\mathcal{O}\left(\frac{1}{\sqrt{\Delta}}\cdot\log\left(\frac{1}{\epsilon}\right)\right)$	n	[GVL12]
	Power+M	$\mathcal{O}\left(\frac{1}{\sqrt{\Delta}}\cdot\log\left(\frac{1}{\epsilon}\right)\right)$	n	**This paper**
Online	Oja's	$\mathcal{O}\left(\frac{\sigma^2}{\Delta^2}\cdot\frac{1}{\epsilon}+\frac{1}{\sqrt{\epsilon}}\right)$	$\mathcal{O}(1)$	[Jai+16]
	Mini-batch Power+M	$\mathcal{O}\left(\frac{1}{\sqrt{\Delta}}\cdot\log\left(\frac{1}{\epsilon}\right)\right)$	$\mathcal{O}\left(\frac{\sqrt{d}\sigma^2}{\Delta^{3/2}}\cdot\frac{1}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)$	**This paper**
Offline	VR-PCA	$\mathcal{O}\left(\frac{r^2}{\Delta^2}\cdot\log\left(\frac{1}{\epsilon}\right)\right)$	$\mathcal{O}(1)$	[Sha15]
	VR Power+M	$\mathcal{O}\left(\frac{1}{\sqrt{\Delta}}\cdot\log\left(\frac{1}{\epsilon}\right)\right)$	$\mathcal{O}\left(\frac{\sqrt{d}\sigma^2}{\Delta^{3/2}}\right)$	**This paper**

orthogonal polynomials, which allows us to use well-known results about orthogonal polynomials to analyze the algorithm's convergence. Consider the polynomial sequence $p_t(x)$, defined as

$$p_{t+1}(x) = xp_t(x) - \beta p_{t-1}(x), p_0 = 1, p_1 = x/2. \quad (\mathbf{P})$$

According to Favard's theorem [Chi11], this recurrence forms an orthogonal polynomial family—in fact these are scaled Chebyshev polynomials of the first kind. If we use the update $(\mathbf{A})$ with appropriate initialization, then our iterates will be given by

$$\mathbf{w}_t = p_t(\mathbf{A})\mathbf{w}_0 = \textstyle\sum_{i=1}^d p_t(\lambda_i)\mathbf{u}_i\mathbf{u}_i^T\mathbf{w}_0.$$

We use this expression and properties of the Chebyshev polynomials to explicitly bound the convergence rate of the accelerated power method with Theorem 1 (analysis and proof in Appendix A).

Theorem 1. *Given a PSD matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ with eigenvalues $1 \geq \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n \geq 0$, running update $(\mathbf{A})$ with $\lambda_2 \leq 2\sqrt{\beta} < \lambda_1$ results in estimates with worst-case error*

$$1 - \frac{(\mathbf{u}_1^T\mathbf{w}_t)^2}{\|\mathbf{w}_t\|^2} \leq \frac{4}{\left|\mathbf{w}_0^T\mathbf{u}_1\right|^2}\cdot\left(\frac{2\sqrt{\beta}}{\lambda_1+\sqrt{\lambda_1^2-4\beta}}\right)^{2t}.$$

We can derive the following corollary, which gives the iteration complexity to achieve ϵ error.

Corollary 2. *In the same setting as Theorem 1, update $(\mathbf{A})$ with $\mathbf{w}_0 \in \mathbb{R}^d$ such that $\mathbf{u}_1^T\mathbf{w}_0 \neq 0$, for any $\epsilon \in (0,1)$, after $T = \mathcal{O}\left(\frac{\sqrt{\beta}}{\sqrt{\lambda_1^2-4\beta}}\cdot\log\frac{1}{\epsilon}\right)$ iterations achieves $1 - \frac{(\mathbf{u}_1^T\mathbf{w}_T)^2}{\|\mathbf{w}_T\|^2} \leq \epsilon$.*

Remark. Minimizing $\frac{\sqrt{\beta}}{\sqrt{\lambda_1^2-4\beta}}$ over $[\lambda_2^2/4, \lambda_1^2/4)$ tells us that $\beta = \lambda_2^2/4$ is the optimal setting.

In comparison to power iteration, this algorithm converges at an accelerated rate. In fact, as shown in Table 1, this momentum power method scheme (with the optimal assignment of $\beta = \lambda_2^2/4$) even matches the worst-case rate of the Lanczos method.

Extensions In Appendix B.1, we extend this momentum scheme to achieve acceleration in the setting where we want to recover multiple top eigenvectors of A, rather than just one. In Appendix B.2 we show that this momentum method is numerically stable, whereas the Lanczos method suffers from numerical instability [TBI97; GVL12]. Next, in Appendix B.3 we provide a heuristic for auto-tuning the momentum parameter, which is useful in practice. Finally, in Appendix B.4, we consider a larger orthogonal polynomial family, and we show that given some information about the tail spectrum of the matrix, we can obtain even faster convergence by using a 4-term inhomogeneous recurrence.

3 Stochastic PCA

Motivated by the results in the previous section, we study using momentum to accelerate PCA in the stochastic setting. We consider a streaming PCA setting, where we are given a series of i.i.d. samples, $\tilde{\mathbf{A}}_t$, such that

$$\mathbb{E}[\tilde{\mathbf{A}}_t] = \mathbf{A}, \ \max_t \|\tilde{\mathbf{A}}_t\| \leq r, \ \mathbb{E}[\|\tilde{\mathbf{A}}_t - \mathbf{A}\|^2] = \sigma^2. \quad (1)$$

In the sample covariance setting of Section 2, $\tilde{\mathbf{A}}_t$ can be obtained by selecting $\mathbf{x}_i\mathbf{x}_i^T$, where $\mathbf{x}_i$ is uni-

formly sampled from the dataset. One of the most popular streaming PCA algorithms is Oja's algorithm [Oja82], which repeatedly runs the update[2] $\mathbf{w}_{t+1} = (I + \eta\tilde{\mathbf{A}}_t)\mathbf{w}_t$. A natural way to try to accelerate Oja's algorithm is to directly add a momentum term, which leads to

$$\mathbf{w}_{t+1} = (I + \eta\tilde{\mathbf{A}}_t)\mathbf{w}_t - \beta\mathbf{w}_{t-1}. \tag{2}$$

In expectation, this stochastic recurrence behaves like the deterministic three-term recurrence ($\mathbf{A}$), which can achieve acceleration with proper setting of β. However, we observe empirically that (2) usually does not give acceleration. In Figure 1(a), we see that while adding momentum does accelerate the initial convergence to the noise ball, it also increases the size of the noise ball—and decreasing the step size to compensate for this roughly cancels out the acceleration from momentum. This same counterintuitive phenomenon has independently been observed in Goh [Goh17] for stochastic optimization. The inability of momentum to accelerate Oja's algorithm is perhaps not surprising because the sampling complexity of Oja's algorithm is asymptotically optimal in terms of the eigen-gap [AZL16b].

In Section 4, we will characterize this connection between the noise ball size and momentum in more depth by presenting an exact expression for the variance of the iterates. Our analysis shows that when the sample variance is bounded, momentum can yield an accelerated convergence rate. In this section, we will present two methods that can be used to successfully control the variance: mini-batching and variance reduction. A summary of our methods and convergence rates is presented in Table 1.

3.1 Stochastic power method with momentum

In addition to adding momentum to Oja's algorithm, another natural way to try to accelerate stochastic PCA is to use the deterministic update ($\mathbf{A}$) with random samples $\tilde{\mathbf{A}}_t$ rather than the exact matrix $\mathbf{A}$. Specifically, we analyze the stochastic recurrence

$$\mathbf{w}_{t+1} = \tilde{\mathbf{A}}_t\mathbf{w}_t - \beta\mathbf{w}_{t-1}, \tag{3}$$

where $\tilde{\mathbf{A}}_t$ is an i.i.d. unbiased random estimate of $\mathbf{A}$. We write this more explicitly as Algorithm 1.

When the variance is zero, the dynamics of this algorithm are the same as the dynamics of update ($\mathbf{A}$), so it converges at the accelerated rate given in Theorem 1. Even if the variance is nonzero, but sufficiently small, we can still prove that Algorithm 1 converges at an accelerated rate.

[2]Here we consider a constant step size scheme, in which the iterate will converge to a noise ball. The size of the noise ball depends on the variance.

Algorithm 1 Mini-batch Power Method with Momentum (**Mini-batch Power+M**)

Require: Initial point $\mathbf{w}_0$, Number of Iterations T, Batch size s, Momentum parameter β
  $\mathbf{w}_{-1} \leftarrow \mathbf{0}$,
  for t = 0 **to** T - 1 **do**
    Generate a mini-batch of i.i.d. samples $B = \{\tilde{\mathbf{A}}_{t_1}, \cdots, \tilde{\mathbf{A}}_{t_s}\}$
    Update: $\mathbf{w}_{t+1} \leftarrow (\frac{1}{s}\sum_{i=1}^{s}\tilde{\mathbf{A}}_{t_i})\mathbf{w}_t - \beta\mathbf{w}_{t-1}$
    Normalization: $\mathbf{w}_t \leftarrow \mathbf{w}_t/\|\mathbf{w}_{t+1}\|, \mathbf{w}_{t+1} \leftarrow \mathbf{w}_{t+1}/\|\mathbf{w}_{t+1}\|$.
  end for
  return $\mathbf{w}_T$

Theorem 3. *Suppose we run Algorithm 1 with* $2\sqrt{\beta} \in [\lambda_2, \lambda_1)$. *Let* $\Sigma = \mathbb{E}\left[(\mathbf{A}_t - \mathbf{A}) \otimes (\mathbf{A}_t - \mathbf{A})\right]$[3]. *Suppose that* $\|\mathbf{w}_0\| = 1$ *and* $\left|\mathbf{u}_1^T\mathbf{w}_0\right| \geq 1/2$. *For any* $\delta \in (0,1)$ *and* $\epsilon \in (0,1)$, *if*

$$\begin{aligned} T &= \frac{\sqrt{\beta}}{\sqrt{\lambda_1^2 - 4\beta}}\log\left(\frac{32}{\delta\epsilon}\right), \\ \|\Sigma\| &\leq \frac{(\lambda_1^2 - 4\beta)\delta\epsilon}{256\sqrt{dT}} = \frac{(\lambda_1^2 - 4\beta)^{3/2}\delta\epsilon}{256\sqrt{d}\sqrt{\beta}}\log^{-1}\left(\frac{32}{\delta\epsilon}\right), \end{aligned} \tag{4}$$

then with probability at least $1 - 2\delta$, *we have* $1 - (\mathbf{u}_1^T\mathbf{w}_T)^2 \leq \epsilon$.

When we compare this to the result of Theorem 1, we can see that as long as the variance $\|\Sigma\|$ is sufficiently small, the number of iterations we need to run in the online setting is the same as in the deterministic setting (up to a constant factor that depends on δ). In particular, this is faster than the power method without momentum in the deterministic setting. Of course, in order to get this accelerated rate, we need some way of getting samples that satisfy the variance condition of Theorem 3. Certain low-noise datasets might satisfy this condition, but this is not always the case. In the next two sections, we discuss methods of getting lower-variance samples.

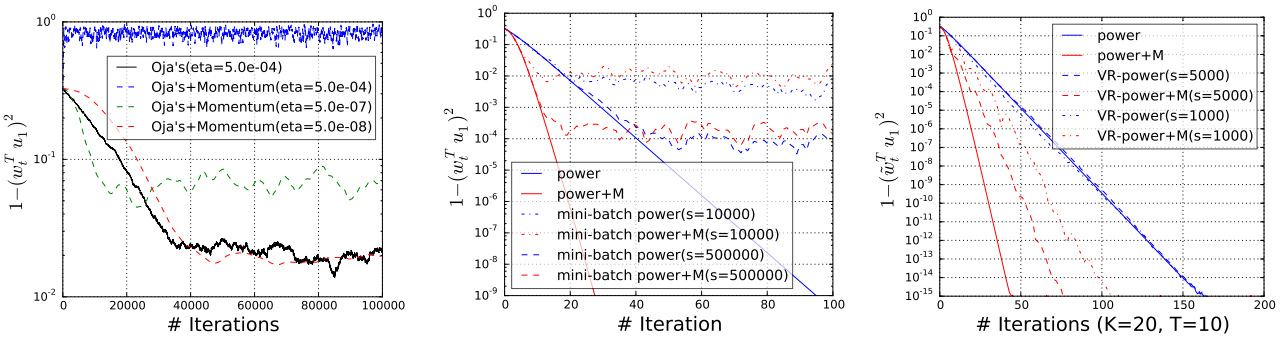
3.2 Controlling variance with mini-batches

In the online PCA setting, a natural way of getting lower-variance samples is to increase the *batch size* (parameter s) used by Algorithm 1. Using the following bound on the variance,

$$\begin{aligned} \|\Sigma\| &= \|\mathbb{E}\left[(\mathbf{A}_t - \mathbf{A}) \otimes (\mathbf{A}_t - \mathbf{A})\right]\| \\ &\leq \mathbb{E}\left[\|(\mathbf{A}_t - \mathbf{A}) \otimes (\mathbf{A}_t - \mathbf{A})\|\right] \\ &= \mathbb{E}\left[\|\mathbf{A}_t - \mathbf{A}\|^2\right] = \frac{\sigma^2}{s}, \end{aligned}$$

we can get an upper bound on the mini-batch size we will need in order to satisfy the variance condition in Theorem 3, which leads to the following corollary.

[3]$\otimes$ denotes the Kronecker product.



(a) Oja's algorithm with Momentum (b) Without Variance Reduction (c) With Variance Reduction

Figure 1: Different PCA algorithms on a synthetic dataset $\mathbf{X} \in \mathbb{R}^{10^6 \times 10}$ where the covariance matrix has eigengap $\Delta = 0.1$. Figure 1(a) shows the performance of Oja's algorithm with momentum. The momentum is set to the optimal $\beta = (1+\eta\lambda_2)^2/4$. Different dashed lines correspond to different step sizes (β changes correspondingly) for momentum methods. Figure 1(b) shows the performance of mini-batch power methods. Increasing the mini-batch size led to a smaller noise ball. Figure 1(c) shows the performance of VR power methods. The epoch length $T = 10$ was estimated according to (7) by setting $\delta = 1\%$ and $c = 1/16$. Stochastic methods report the average performance over 10 runs.

Corollary 4. *Suppose we run Algorithm 1 with* $2\sqrt{\beta} \in [\lambda_2, \lambda_1)$. *Assume that* $\|\mathbf{w}_0\| = 1$ *and* $\left|\mathbf{u}_1^T \mathbf{w}_0\right| \geq 1/2$. *For any* $\delta \in (0,1)$ *and* $\epsilon \in (0,1)$, *if*

$$T = \frac{\sqrt{\beta}}{\sqrt{\lambda_1^2 - 4\beta}} \log\left(\frac{32}{\delta\epsilon}\right),$$
$$s \geq \frac{256\sqrt{d}\sigma^2 T}{(\lambda_1^2 - 4\beta)\delta\epsilon} = \frac{256\sqrt{d}\sqrt{\beta}\sigma^2}{(\lambda_1^2 - 4\beta)^{3/2}\delta\epsilon} \log\left(\frac{32}{\delta\epsilon}\right),$$

then with probability at least $1 - 2\delta$, $1 - (\mathbf{u}_1^T \mathbf{w}_T)^2 \leq \epsilon$.

This means that no matter what the variance of the estimator is, we can still converge at the same rate as the deterministic setting as long as we can compute mini-batches of size s quickly. One practical way of doing this is by using many parallel workers: a mini-batch of size s can be computed in $\mathcal{O}(1)$ time by $\mathcal{O}(s)$ machines working in parallel. If we use a sufficiently large cluster, this means that Algorithm 1 converges in asymptotically less time than any non-momentum power method that uses the cluster for mini-batching, because we converge faster than even the deterministic non-momentum method.

One drawback of this approach is that the required variance decreases as a function of ϵ, so we will need to increase our mini-batch size as the desired error decreases. If we are running in parallel on a cluster of fixed size, this means that we will eventually exhaust the parallel resources of the cluster and be unable to compute the mini-batches in asymptotic $\mathcal{O}(1)$ time. As a result, we now seek methods to reduce the required batch size, and remove its dependence on ϵ.

3.3 Reducing batch size with variance reduction

Another way to generate low-variance samples is the *variance reduction* technique. This technique can be used if we have access to the target matrix $\mathbf{A}$ so that we can occasionally compute an exact matrix-vector product with $\mathbf{A}$. For example, in the offline setting, we can compute $\mathbf{A}\mathbf{w}$ by occasionally doing a complete pass over the data. In PCA, Shamir [Sha15] has applied the standard variance reduction technique that was used in stochastic convex optimization [JZ13], in which the stochastic term in the update is

$$\mathbf{A}\mathbf{w}_t + (\mathbf{A}_t - \mathbf{A})(\mathbf{w}_t - \tilde{\mathbf{w}}), \tag{5}$$

where $\tilde{\mathbf{w}}$ is the (normalized) anchor iterate, for which we know the exact value of $\mathbf{A}\tilde{\mathbf{w}}$. We propose a slightly different variance reduction scheme, where the stochastic term in the update is

$$\begin{aligned} &\left[\mathbf{A} + (\mathbf{A}_t - \mathbf{A})(I - \tilde{\mathbf{w}}\tilde{\mathbf{w}}^T)\right]\mathbf{w}_t \\ &= \mathbf{A}\mathbf{w}_t + (\mathbf{A}_t - \mathbf{A})(I - \tilde{\mathbf{w}}\tilde{\mathbf{w}}^T)\mathbf{w}_t. \end{aligned} \tag{6}$$

It is easy to verify that both (5) and (6) can be computed using only the samples $\mathbf{A}_t$ and the exact value of $\mathbf{A}\tilde{\mathbf{w}}$. In the PCA setting, (6) is more appropriate because progress is measured by the angle between $\mathbf{w}_t$ and $\mathbf{u}_1$, not the ℓ_2 distance as in the convex optimization problem setting: this makes (6) easier to analyze. In addition to being easier to analyze, our proposed update rule (6) produces updates that have generally lower variance because for all unit vectors $\mathbf{w}_t$ and $\tilde{\mathbf{w}}$, $\|\mathbf{w}_t - \tilde{\mathbf{w}}\| \geq \|(I - \tilde{\mathbf{w}}\tilde{\mathbf{w}}^T)\mathbf{w}_t\|$. Using this update step

results in the variance-reduced power method with momentum in Algorithm 2. A number of methods use

Algorithm 2 VR Power Method with Momentum (**VR Power+M**)

Require: Initial point $\mathbf{w}_0$, Number of Iterations T, Batch size s, Momentum parameter β
  $\mathbf{w}_{-1} \leftarrow \mathbf{0}$
  for k = 1 **to** K **do**
    $\tilde{\mathbf{v}} \leftarrow \mathbf{A}\tilde{\mathbf{w}}_k$ (Usually there is no need to materialize $\mathbf{A}$ in practice).
    for t = 1 **to** T **do**
      Generate a mini-batch of i.i.d. samples $B = \{\tilde{\mathbf{A}}_{t_1}, \cdots, \tilde{\mathbf{A}}_{t_s}\}$
      Update: $\alpha \leftarrow \mathbf{w}_t^T \tilde{\mathbf{w}}_k$,
      $\mathbf{w}_{t+1} \leftarrow \frac{1}{s}\sum_{i=1}^{s} \tilde{\mathbf{A}}_{t_i}(\mathbf{w}_t - \alpha\tilde{\mathbf{w}}_k) + \alpha\tilde{\mathbf{v}} - \beta\mathbf{w}_{t-1}$
      Normalization: $\mathbf{w}_t \leftarrow \mathbf{w}_t/\|\mathbf{w}_{t+1}\|$,
      $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_{t+1}/\|\mathbf{w}_{t+1}\|$.
    end for
    $\tilde{\mathbf{w}}_{k+1} \leftarrow \mathbf{w}_T$.
  end for
  return $\mathbf{w}_K$

this kind of SVRG-style variance reduction technique, which converges at a linear rate and is not limited by a noise ball. Our method improves upon that by achieving the *accelerated rate* throughout, and only using a mini-batch size that is constant with respect to ϵ.

Theorem 5. *Suppose we run Algorithm 2 with* $2\sqrt{\beta} \in [\lambda_2, \lambda_1)$ *and a initial unit vector* $\mathbf{w}_0$ *such that* $1 - (\mathbf{u}_1^T\mathbf{w}_0)^2 \leq \frac{1}{2}$. *For any* $\delta, \epsilon \in (0,1)$, *if*

$$T = \frac{\sqrt{\beta}}{\sqrt{\lambda_1^2 - 4\beta}} \log\left(\frac{1}{c\delta}\right), s \geq \frac{32\sqrt{d}\sqrt{\beta}\sigma^2}{c(\lambda_1^2 - 4\beta)\delta} \log\left(\frac{1}{c\delta}\right), \tag{7}$$

then after $K = \mathcal{O}(\log(1/\epsilon))$ *epochs, with probability at least* $1 - \log\left(\frac{1}{\epsilon}\right)\delta$, *we have* $1 - (\mathbf{u}_1^T\tilde{\mathbf{w}}_K)^2 \leq \epsilon$, *where* $c \in (0, 1/16)$ *is a numerical constant.*

By comparing to the results of Theorem 1 and Theorem 5, we notice that the VR power method with momentum achieves the same convergence rate, in terms of the total number of iterations we need to run, as the deterministic setting. In contrast to the non-variance-reduced setting, the mini-batch size we need to use does not depend on the desired error ϵ, which allows us to use a fixed mini-batch size throughout the execution of the algorithm. This means that we can use Algorithm 2 together with a parallel mini-batch-computing cluster of fixed size to compute solutions of arbitrary accuracy at a rate faster than any non-momentum power method could achieve. As shown in Table 1, in terms of number of iterations, the momentum methods achieve accelerated linear convergence with proper mini-batching (our results there follow Corollary 4 and Theorem 5, using the optimal momentum $\beta = \lambda_2^2/4$.).

3.4 Experiments

We first use synthetic experiments (details in Appendix E) to illustrate how the variance affects the momentum methods. Figure 1(b) shows that the stochastic power method maintains the same linear convergence as the deterministic power method before hitting the noise ball. Therefore, the momentum method can accelerate the convergence before hitting the noise ball. Figure 1(c) shows that the variance-reduced power method indeed can achieve an accelerated linear convergence with a much smaller batch size on this same synthetic dataset.

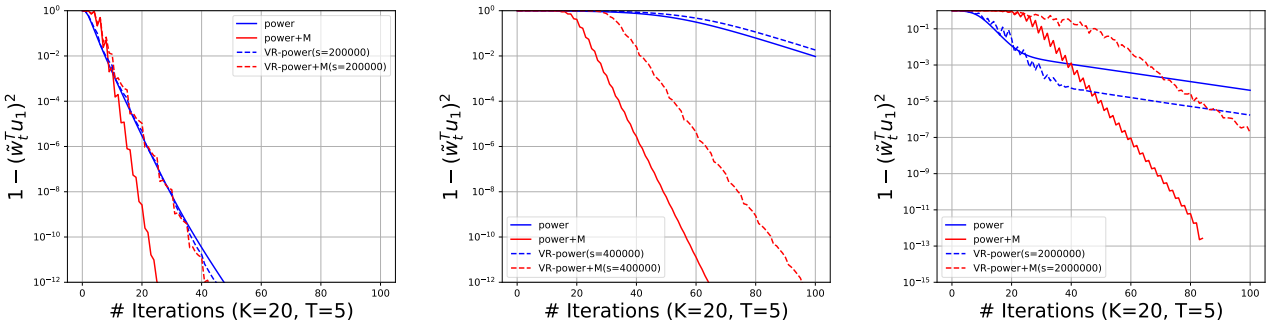
We additionally demonstrate that our variance-reduced method results in accelerated convergence for finding the first eigenvector of the adjacency matrices of several networks. For these experiments, We use the arXiv ASTRO-PH collaboration network [LKF07], the arXiv High-energy physics citation network [LKF05], and the Google web graph [Les+09]. This is the eigenvector centrality task, which is often used in network analysis [Bra05; Bon72]. In the eigenvector centrality task, components of the top eigenvector that are large correspond to nodes that are central to the network. The eigenvector centrality task is closely related the PageRank metric. Figure 2 shows that the variance-reduced power method achieves an accelerated linear convergence on this task. In these experiments, momentum results in larger improvements in convergence when the eigen-gap is small.

4 Convergence analysis

In this section, we sketch the proofs of Theorems 3 and 5. The main idea is to tightly bound the variance of the iterates with properties of the Chebyshev polynomials. Both with mini-batches (Algorithm 1) and variance reduction (Algorithm 2), the dynamics of the stochastic power method with momentum from (3) can be written as $\mathbf{w}_t = \mathbf{F}_t\mathbf{w}_0/\|\mathbf{F}_t\mathbf{w}_0\|$ where $\{\mathbf{F}_t\}$ is a sequence of stochastic matrices in $\mathbb{R}^{d\times d}$ satisfying

$$\mathbf{F}_{t+1} = \mathbf{A}_{t+1}\mathbf{F}_t - \beta\mathbf{F}_{t-1},\ \mathbf{F}_0 = I,\ \mathbf{F}_{-1} = \mathbf{0}. \tag{8}$$

The random matrix $\mathbf{A}_t \in \mathbb{R}^{d\times d}$ has different forms in Algorithm 1 and Algorithm 2. However, in both algorithms, $\mathbf{A}_t$ will be i.i.d. and satisfy $\mathbb{E}[\mathbf{A}_t] = \mathbf{A}$. In fact, this recurrence (8) is general enough to be applied in many other problems, including least-squares regression and the randomized Kaczmarz algorithm [SV09; GR15], as well as some non-convex matrix problems [DSRO15] such as matrix completion [JNS13], phase retrieval [CLS15] and subspace tracking [BNR10]. Since $\mathbf{F}_t$ obeys a linear recurrence, its second moment also follows a linear recurrence (in fact, all its moments do). We decompose



(a) ASTRO-PH collaboration (b) High Energy Physics paper citation (c) Google web graph

Figure 2: Variance-reduced and full-pass PCA algorithms on large network datasets. The dimensions are (a) 18,772, (b) 34,546, and (c) 875,713. The relative eigen-gaps are (a) 0.20, (b) 0.047, and (c) 0.027. Our momentum-based methods result in larger improvements in the convergence rate when the relative eigen-gap is smaller.

this recurrence using Chebyshev polynomials to get a tight bound on the covariance of $\mathbf{F}_t$, which is shown in Lemma 6. This bound is exact in the scalar case.

Lemma 6. *Suppose* $\lambda_1^2 \geq 4\beta$ *and* $\Sigma = \mathbb{E}\left[(\mathbf{A}_t - \mathbf{A}) \otimes (\mathbf{A}_t - \mathbf{A})\right]$. *The norm of the covariance of the matrix* $\mathbf{F}_t$ *is bounded by*

$$\begin{aligned}&\left\|\mathbb{E}\left[\mathbf{F}_t \otimes \mathbf{F}_t\right] - \mathbb{E}\left[\mathbf{F}_t\right] \otimes \mathbb{E}\left[\mathbf{F}_t\right]\right\| \\ &\leq \sum_{n=1}^{t} \|\Sigma\|^n \beta^{t-n} \sum_{\mathbf{k} \in S_{t-n}^{n+1}} \prod_{i=1}^{n+1} U_{k_i}^2\left(\frac{\lambda_1}{2\sqrt{\beta}}\right),\end{aligned}$$

where $U_k(\cdot)$ *is the Chebyshev polynomial of the second kind, and* S_m^n *denotes the set of vectors in* $\mathbb{N}^n$ *with entries that sum to* m, *i.e.*

$$S_m^n = \left\{\mathbf{k} = (k_1, \cdots, k_n) \in \mathbb{N}^n \mid \textstyle\sum_{i=1}^n k_i = m\right\}.$$

For the mini-batch power method without variance reduction (Algorithm 1), the goal is to bound $1 - (\mathbf{u}_1^T \mathbf{w}_t)^2$, which is equivalent to bounding $\sum_{i=2}^d (\mathbf{u}_i^T \mathbf{F}_t \mathbf{w}_0)^2 / (\mathbf{u}_1^T \mathbf{F}_t \mathbf{w}_0)^2$. We use Lemma 6 to get a variance bound for the denominator of this expression, which is

$$\mathbf{Var}\left[\mathbf{u}_1^T \mathbf{F}_t \mathbf{w}_0\right] \leq p_t^2(\lambda_1; \beta) \cdot \frac{8\|\Sigma\| t}{\lambda_1^2 - 4\beta}. \tag{9}$$

With this variance bound and Chebyshev's inequality we get a probabilistic lower bound for $|\mathbf{u}_1^T \mathbf{F}_t \mathbf{w}_0|$. Lemma 6 can also be used to get an upper bound for the numerator, which is

$$\begin{aligned}&\mathbb{E}\left[\sum_{i=2}^{d} (\mathbf{u}_i^T \mathbf{F}_t \mathbf{w}_0)^2\right] \\ &\leq p_t^2(\lambda_1; \beta) \cdot \left(\frac{8\sqrt{d}\|\Sigma\| t}{(\lambda_1^2 - 4\beta)} + \frac{p_t^2(2\sqrt{\beta}; \beta)}{p_t^2(\lambda_1; \beta)}\right)\end{aligned} \tag{10}$$

By Markov's inequality we can get a probabilistic upper bound for $\sum_{i=2}^d (\mathbf{u}_i^T \mathbf{F}_t \mathbf{w}_0)^2$. The result in Theorem 3 now follows by a union bound. The details of the proof appear in Appendix C.1.

Next, we consider the case with variance reduction (Algorithm 2). The analysis contains two steps. The first step is to show a geometric contraction for a single epoch, i.e.

$$1 - (\mathbf{u}_1^T \mathbf{w}_T)^2 \leq \rho \cdot \left(1 - (\mathbf{u}_1^T \mathbf{w}_0)^2\right), \tag{11}$$

with probability at least $1 - \delta$, where $\rho < 1$ is a numerical constant. Afterwards, the second step is to get the final ϵ accuracy of the solution, which trivially requires $\mathcal{O}\left(\log(1/\epsilon)\right)$ epochs. Thus, the analysis boils down to analyzing a single epoch. Notice that in this setting,

$$\mathbf{A}_{t+1} = \mathbf{A} + \left(\tfrac{1}{s}\textstyle\sum_{i=1}^s \tilde{\mathbf{A}}_{t_i} - \mathbf{A}\right)(I - \mathbf{w}_0 \mathbf{w}_0^T), \tag{12}$$

and again $\mathbf{w}_t = \mathbf{F}_t \mathbf{w}_0 / \|\mathbf{F}_t \mathbf{w}_0\|$. Using similar techniques to the mini-batch power method setting, we can prove a variant of Lemma 6 specialized to (12).

Lemma 7. *Suppose* $\lambda_1^2 \geq 4\beta$. *Let* $\mathbf{w}_0 \in \mathbb{R}^d$ *be a unit vector,* $\theta = 1 - (\mathbf{u}_1^T \mathbf{w}_0)^2$, *and*

$$\Sigma = \mathbb{E}\left[\left(\tfrac{1}{s}\textstyle\sum_{i=1}^s \tilde{\mathbf{A}}_{t_i} - \mathbf{A}\right) \otimes \left(\tfrac{1}{s}\textstyle\sum_{i=1}^s \tilde{\mathbf{A}}_{t_i} - \mathbf{A}\right)\right].$$

Then, the norm of the covariance will be bounded by

$$\begin{aligned}&\left\|\mathbb{E}\left[\mathbf{F}_t \mathbf{w}_0 \otimes \mathbf{F}_t \mathbf{w}_0\right] - \mathbb{E}\left[\mathbf{F}_t \mathbf{w}_0\right] \otimes \mathbb{E}\left[\mathbf{F}_t \mathbf{w}_0\right]\right\| \\ &\leq 4\theta \cdot \sum_{n=1}^{t} \|\Sigma\|^n \beta^{t-n} \sum_{\mathbf{k} \in S_{t-n}^{n+1}} \prod_{i=1}^{n+1} U_{k_i}^2\left(\frac{\lambda_1}{2\sqrt{\beta}}\right).\end{aligned}$$

Comparing to the result in Lemma 6, this lemma shows that the covariance is also controlled by the angle between $\mathbf{u}_1$ and $\mathbf{w}_0$ which is the anchor point in each

epoch. Since the anchor point $\tilde{\mathbf{w}}_k$ is approaching $\mathbf{u}_1$, the norm of the covariance is shrinking across epochs—this allows us to prove (11). From here, the proof of Theorem 5 is similar to non-VR case, and the details are in Appendix C.2.

Remark 1. As summarized in Table 1, we achieved the optimal, accelerated iteration complexity, by allowing computation to be wide (massively parallel) instead of deep (many sequential steps). This ability for massive parallelization comes at a cost of an extra $\sqrt{d}$ factor in total computation. It is an interesting open question whether this extra computation is fundamental and unavoidable for massively parallel methods.

Remark 2. Note that Theorems 3 and 5 only state the local convergence complexity. However, it is worth mentioning that there is no technical challenge in obtaining the state of warm initialization in the theorems. For example, starting from random uniform initialization, it will take $\mathcal{O}(d/\Delta^2)$ iterations for Alecton [DSRO15] to get into a constant error ball. This is independent of ϵ and negligible comparing to the local complexity.

5 Related work

PCA A recent spike in research activity has focused on improving a number of computational and statistical aspects of PCA, including tighter sample complexity analysis [Jai+16; Li+16], global convergence [DSRO15; AZL16b; BDF13], memory efficiency [MCJ13] and doing online regret analysis [Bou+15]. Some work has also focused on tightening the analysis of power iteration and Krylov methods to provide gap-independent results using polynomial-based analysis techniques [MM15]. However, that work does not consider the stochastic setting. Some works that study Oja's algorithm [Oja82] or stochastic power methods in the stochastic setting focus on the analysis of a gap-free convergence rate for the *distinct PCA formulation of maximizing explained variance* (as opposed to recovering the strongest direction) [Sha16; AZL16b]. Others provide better dependence on the dimension of the problem [Jai+16]. Garber et al. [Gar+16] and Allen-Zhu and Li [AZL16c] use faster linear system solvers to speed up PCA algorithms such that the convergence rate has the square root dependence on the eigengap in the offline setting. However their methods require solving a series of linear systems, which is not trivially parallelizable. Also none of these results give a convergence analysis that is asymptotically tight in terms of variance. Another line of work has focused on variance control for PCA in the stochastic setting [Sha15] to get a different kind of acceleration. Since this is an independent source of improvement, these methods can be further accelerated using our momentum scheme. In addition [XLS15] study the stochastic power methods for kernel PCA in the random feature space, where our momentum scheme is also applicable.

Stochastic acceleration Momentum is a common acceleration technique in convex optimization [Pol64; Nes83], and has been widely adopted as the de-facto optimization method for non-convex objectives in deep learning [Sut+13]. Provably accelerated stochastic methods have previously been found for convex problems [Cot+11; Jai+17]. However, similar results for non-convex problems remain elusive, despite empirical evidence that momentum results in acceleration for some non-convex problems [Sut+13; KB14].

Orthogonal Polynomials The Chebyshev polynomial family is a sequence of orthogonal polynomials [Chi11] that has been used for analyzing accelerated methods. For example, Chebyshev polynomials have been studied to accelerate the solvers of linear systems [GV61; GVL12] and to accelerate convex optimization [SdB16]. Trefethen and Bau III [TBI97] use Chebyshev polynomials to show that the Lanczos method is quadratically faster than the standard power iteration. The Lanczos method is conventionally considered the accelerated version of power method with momentum [HP14].

6 Conclusion

This paper introduced a very simple accelerated PCA algorithm that works in the stochastic setting. As a foundation, we presented the power method with momentum, an accelerated scheme in the deterministic setting. We proved that the power method with momentum obtains quadratic acceleration like in the convex optimization setting. Then, for the stochastic setting, we introduced and analyzed the stochastic power method with momentum. By leveraging the Chebyshev polynomials, we derived a convergence rate that is asymptotically tight in terms of the variance. Using a tight variance analysis, we demonstrated how the momentum scheme behaves in a stochastic system, which can lead to a better understanding of how momentum interacts with variance in stochastic optimization problems [Goh17]. Specifically, with mini-batching, the stochastic power method with momentum can achieve accelerated convergence to the noise ball. Alternatively, using variance reduction, accelerated convergence at a linear rate can be achieved with a much smaller batch size.

Acknowledgments

We thank Aaron Sidford for helpful discussion and feedback on this work.

We gratefully acknowledge the support of the Defense Advanced Research Projects Agency (DARPA) SIMPLEX program under No. N66001-15-C-4043, D3M program under No. FA8750-17-2-0095, the National Science Foundation (NSF) CAREER Award under No. IIS- 1353606, the Office of Naval Research (ONR) under awards No. N000141210041 and No. N000141310129, a Sloan Research Fellowship, the Moore Foundation, an Okawa Research Grant, Toshiba, and Intel. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA, NSF, ONR, or the U.S. government.

References

[AZL16a] Zeyuan Allen-Zhu and Yuanzhi Li. "Doubly Accelerated Methods for Faster CCA and Generalized Eigendecomposition". In: *arXiv preprint arXiv:1607.06017* (2016).

[AZL16b] Zeyuan Allen-Zhu and Yuanzhi Li. "First Efficient Convergence for Streaming k-PCA: a Global, Gap-Free, and Near-Optimal Rate". In: *arXiv preprint arXiv:1607.07837* (2016).

[AZL16c] Zeyuan Allen-Zhu and Yuanzhi Li. "LazySVD: Even faster SVD decomposition yet without agonizing pain". In: *Advances in Neural Information Processing Systems*. 2016, pp. 974–982.

[BDF13] Akshay Balsubramani, Sanjoy Dasgupta, and Yoav Freund. "The fast convergence of incremental pca". In: *Advances in Neural Information Processing Systems*. 2013, pp. 3174–3182.

[BNR10] Laura Balzano, Robert Nowak, and Benjamin Recht. "Online identification and tracking of subspaces from highly incomplete information". In: *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*. IEEE. 2010, pp. 704–711.

[Bon72] Phillip Bonacich. "Factoring and weighting approaches to status scores and clique identification". In: *Journal of Mathematical Sociology* 2.1 (1972), pp. 113–120.

[Bou+15] Christos Boutsidis et al. "Online principal components analysis". In: *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM. 2015, pp. 887–901.

[Bra05] Ulrik Brandes. *Network analysis: methodological foundations*. Vol. 3418. Springer Science & Business Media, 2005.

[Chi11] Theodore S Chihara. *An introduction to orthogonal polynomials*. Courier Corporation, 2011.

[CLS15] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. "Phase retrieval via Wirtinger flow: Theory and algorithms". In: *IEEE Transactions on Information Theory* 61.4 (2015), pp. 1985–2007.

[Cot+11] Andrew Cotter et al. "Better mini-batch algorithms via accelerated gradient methods". In: *Advances in neural information processing systems*. 2011, pp. 1647–1655.

[DSRO15] Christopher De Sa, Christopher Ré, and Kunle Olukotun. "Global Convergence of Stochastic Gradient Descent for Some Non-convex Matrix Problems". In: *International Conference on Machine Learning*. 2015, pp. 2332–2341.

[Gar+16] Dan Garber et al. "Faster eigenvector computation via shift-and-invert preconditioning". In: *International Conference on Machine Learning*. 2016, pp. 2626–2634.

[Goh17] Gabriel Goh. "Why Momentum Really Works". In: *Distill* (2017). DOI: 10.23915/distill.00006. URL: http://distill.pub/2017/momentum.

[GR15] Robert M Gower and Peter Richtárik. "Randomized iterative methods for linear systems". In: *SIAM Journal on Matrix Analysis and Applications* 36.4 (2015), pp. 1660–1690.

[GV61] Gene H Golub and Richard S Varga. "Chebyshev semi-iterative methods, successive overrelaxation iterative methods, and second order Richardson iterative methods". In: *Numerische Mathematik* 3.1 (1961), pp. 147–156.

[GVL12] Gene H Golub and Charles F Van Loan. *Matrix computations*. Vol. 3. JHU Press, 2012.

[Hot33] Harold Hotelling. "Analysis of a complex of statistical variables into principal components." In: *Journal of educational psychology* 24.6 (1933), p. 417.

[HP14] Moritz Hardt and Eric Price. "The noisy power method: A meta algorithm with applications". In: *Advances in Neural Information Processing Systems*. 2014, pp. 2861–2869.

[Jai+16] Prateek Jain et al. "Matching Matrix Bernstein with Little Memory: Near-Optimal Finite Sample Guarantees for Oja's Algorithm". In: *arXiv preprint arXiv:1602.06929* (2016).

[Jai+17] Prateek Jain et al. "Accelerating Stochastic Gradient Descent". In: *arXiv preprint arXiv:1704.08227* (2017).

[JNS13] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. "Low-rank matrix completion using alternating minimization". In: *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*. ACM. 2013, pp. 665–674.

[Jol02] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[JZ13] Rie Johnson and Tong Zhang. "Accelerating stochastic gradient descent using predictive variance reduction". In: *Advances in Neural Information Processing Systems*. 2013, pp. 315–323.

[KB14] Diederik Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[Les+09] Jure Leskovec et al. "Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters". In: *Internet Mathematics* 6.1 (2009), pp. 29–123.

[Li+16] Chris J Li et al. "Near-Optimal Stochastic Approximation for Online Principal Component Estimation". In: *arXiv preprint arXiv:1603.05305* (2016).

[LKF05] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. "Graphs over time: densification laws, shrinking diameters and possible explanations". In: *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM. 2005, pp. 177–187.

[LKF07] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. "Graph evolution: Densification and shrinking diameters". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1.1 (2007), p. 2.

[MCJ13] Ioannis Mitliagkas, Constantine Caramanis, and Prateek Jain. "Memory limited, streaming PCA". In: *Advances in Neural Information Processing Systems*. 2013, pp. 2886–2894.

[MM15] Cameron Musco and Christopher Musco. "Randomized block krylov methods for stronger and faster approximate singular value decomposition". In: *Advances in Neural Information Processing Systems*. 2015, pp. 1396–1404.

[Nes83] Yurii Nesterov. "A method of solving a convex programming problem with convergence rate O (1/k2)". In: *Soviet Mathematics Doklady*. Vol. 27. 2. 1983, pp. 372–376.

[Oja82] Erkki Oja. "Simplified neuron model as a principal component analyzer". In: *Journal of mathematical biology* 15.3 (1982), pp. 267–273.

[Pol64] Boris T Polyak. "Some methods of speeding up the convergence of iteration methods". In: *USSR Computational Mathematics and Mathematical Physics* 4.5 (1964), pp. 1–17.

[SdB16] Damien Scieur, Alexandre d'Aspremont, and Francis Bach. "Regularized Nonlinear Acceleration". In: *Advances In Neural Information Processing Systems*. 2016, pp. 712–720.

[Sha15] Ohad Shamir. "A stochastic PCA and SVD algorithm with an exponential convergence rate". In: *Proc. of the 32st Int. Conf. Machine Learning (ICML 2015)*. 2015, pp. 144–152.

[Sha16] Ohad Shamir. "Convergence of stochastic gradient descent for PCA". In: *International Conference on Machine Learning*. 2016, pp. 257–265.

[Sut+13] Ilya Sutskever et al. "On the importance of initialization and momentum in deep learning". In: *Proceedings of the 30th international conference on machine learning (ICML-13)*. 2013, pp. 1139–1147.

[SV09] Thomas Strohmer and Roman Vershynin. "A randomized Kaczmarz algorithm with exponential convergence". In: *Journal of Fourier Analysis and Applications* 15.2 (2009), pp. 262–278.

[TBI97] Lloyd N Trefethen and David Bau III. *Numerical linear algebra*. Vol. 50. Siam, 1997.

[XLS15] Bo Xie, Yingyu Liang, and Le Song. "Scale up nonlinear component analysis with doubly stochastic gradients". In: *Advances in Neural Information Processing Systems*. 2015, pp. 2341–2349.

A Momentum PCA and Orthogonal Polynomials

In this section, we prove Theorem 1 and give the intuition that the momentum can provide acceleration from both geometric and algebraic perspectives.

First, we restate the update (**A**) for power iteration with momentum,

$$\mathbf{w}_{t+1} = \mathbf{A}\mathbf{w}_t - \beta \mathbf{w}_{t-1}. \tag{\textbf{A}}$$

and the correponding orthogonal polynomial sequence (**P**),

$$p_{t+1}(x) = xp_t(x) - \beta p_{t-1}(x), p_0 = 1, p_1 = x/2. \tag{\textbf{P}}$$

According to Lemma 20, we have the expression of $p_t(x)$,

$$p_t(x) = \begin{cases} \frac{1}{2}\left[\left(\frac{x-\sqrt{x^2-4\beta}}{2}\right)^t + \left(\frac{x+\sqrt{x^2-4\beta}}{2}\right)^t\right], & |x| > 2\sqrt{\beta}, \\ (\sqrt{\beta})^t \cos\left(t \arccos(\frac{x}{2\sqrt{\beta}})\right), & |x| \le 2\sqrt{\beta}. \end{cases}$$

A.1 Proof of Theorem 1

Here we prove a slightly more general result.

Theorem 8. *Given a PSD matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ with eigenvalues $\lambda_1 > \lambda_2 \ge \cdots \ge \lambda_d$ with normalized eigenvectors $\mathbf{u}_1, \cdots, \mathbf{u}_d$, we run the power iteration with momentum update $\mathbf{A}$ with a unit vector $\mathbf{w}_0 \in \mathbb{R}^d$, the we have*

$$1 - \frac{(\mathbf{u}_1^T\mathbf{w}_t)^2}{\|\mathbf{w}_t\|^2} \le \frac{1-(\mathbf{u}_1^T\mathbf{w}_0)^2}{(\mathbf{u}_1^T\mathbf{w}_0)^2} \cdot \begin{cases} 4\left(\frac{2\sqrt{\beta}}{\lambda_1+\sqrt{\lambda_1^2-4\beta}}\right)^{2t}, & \lambda_2 < 2\sqrt{\beta} \\ 4\left(\frac{\lambda_2+\sqrt{\lambda_2^2-4\beta}}{\lambda_1+\sqrt{\lambda_1^2-4\beta}}\right)^{2t}, & \lambda_2 \ge 2\sqrt{\beta} \end{cases}$$

Proof. Denote $d_i = \mathbf{w}_0^T\mathbf{u}_i$, and $\delta^{(t)} = \max_{i=2,\ldots,n} \frac{p_t^2(\lambda_i)}{p_t^2(\lambda_1)}$, then

$$\begin{aligned} 1 - \frac{(\mathbf{u}_1^T\mathbf{w}_t)^2}{\|\mathbf{w}_t\|^2} &= 1 - \frac{(\mathbf{u}_1^T p_t(A)\mathbf{w}_0)^2}{\mathbf{w}_0^T p_t(A)^2\mathbf{w}_0} = 1 - \frac{d_1^2 p_t^2(\lambda_1)}{\sum_{i=1}^d d_i^2 p_t^2(\lambda_i)} = \frac{\sum_{i=2}^n d_i^2 p_t^2(\lambda_i)}{\sum_{i=1}^n d_i^2 p_t^2(\lambda_i)} \\ &= \frac{\sum_{i=2}^n d_i^2 p_t^2(\lambda_i)/p_t^2(\lambda_1)}{d_1^2 + \sum_{i=2}^n d_i^2 p_t^2(\lambda_i)/p_t^2(\lambda_1)} \\ &\le \frac{\sum_{i=2}^2 d_i^2}{d_1^2}\delta^{(t)} \end{aligned}$$

Let's bound $\delta^{(t)}$. Denote k as the smallest index such that $\lambda_k > 2\sqrt{\beta}$. Since $\lambda_1 > 2\sqrt{\beta}$, then $k \ge 1$. Now use Lemma 20, we get

$$\begin{aligned} &|p_t(\lambda_i)| = \frac{1}{2}\left[\left(\frac{\lambda_i - \sqrt{\lambda_i^2-4\beta}}{2}\right)^t + \left(\frac{\lambda_i+\sqrt{\lambda_i^2-4\beta}}{2}\right)^t\right], i \le k, \\ &|p_t(\lambda_i)| \le (\sqrt{\beta})^t, i > k \end{aligned}$$

First, let's consider $2 \le i \le k$.

$$\left|\frac{p_t(\lambda_i)}{p_t(\lambda_1)}\right| = \frac{\left(\frac{\lambda_i-\sqrt{\lambda_i^2-4\beta}}{2}\right)^t + \left(\frac{\lambda_i+\sqrt{\lambda_i^2-4\beta}}{2}\right)^t}{\left(\frac{\lambda_1-\sqrt{\lambda_1^2-4\beta}}{2}\right)^t + \left(\frac{\lambda_1+\sqrt{\lambda_1^2-4\beta}}{2}\right)^t} \le 2\left(\frac{\left(\lambda_i+\sqrt{\lambda_i^2-4\beta}\right)^t}{\left(\lambda_1+\sqrt{\lambda_1^2-4\beta}\right)^t}\right)$$

Now consider $i > k$,

$$\left|\frac{p_t(\lambda_i)}{p_t(\lambda_1)}\right| = \frac{2\left(\sqrt{\beta}\right)^t}{\left(\frac{\lambda_1-\sqrt{\lambda_1^2-4\beta}}{2}\right)^t + \left(\frac{\lambda_1+\sqrt{\lambda_1^2-4\beta}}{2}\right)^t} \leq \frac{2\left(\sqrt{\beta}\right)^t}{\left(\frac{\lambda_1+\sqrt{\lambda_1^2-4\beta}}{2}\right)^t} = 2\left(\frac{2\sqrt{\beta}}{\lambda_1+\sqrt{\lambda_1^2-4\beta}}\right)^t.$$

Therefore plug in the bound for $\delta^{(t)}$ and we get the desired result. □

A.2 Effect of Momentum

In this section, we explain why acceleration happens from both a geometric and algebraic perspective of the orthogonal polynomial recurrence. First, we show the geometric behavior of the orthogonal polynomial sequence. We see that momentum results in a "calm" region, where the orthogonal polynomial sequence grows very slowly and an "explosive" region, where the polynomials grow exponentially fast. We then show how the momentum controls the size of "calm" region. Second, we consider an algebraically equivalent form of the three-term recurrence in terms of an augmented matrix. We see that power iteration with momentum is equivalent to standard power iteration on an augmented matrix and quantitatively how the momentum leads to a "better-conditioned" problem. From either perspective, we get a better understanding about how our methods work.

Regions of the Polynomial Recurrence Now, we demonstrate the effect of momentum on different eigenvalues. In Figure 3, we show the values of the polynomial recurrence, which characterizes the growth of different eigenvalues for varying β.

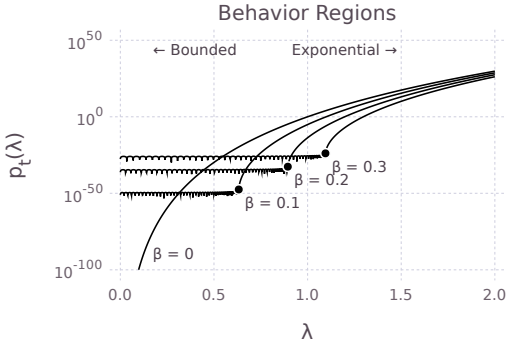


Figure 3: Behavior of polynomial recurrence $\mathbf{P}$ for several values of β. The recurrence is run for $t = 100$ steps.

For power iteration, where $\beta = 0$, $p_t(\lambda) = \lambda^t$. While the recurrence reduces mass on small eigenvalues quickly, eigenvalues near the largest eigenvalue will decay relatively slowly, yielding slow convergence.

As β is increased, a "knee" appears in $p_t(\lambda)$. For values of λ smaller than the knee, $p_t(\lambda)$ remains small, which implies that these eigenvalues decay quickly. For values of λ greater than the knee, $p_t(\lambda)$ grows rapidly, which means that these eigenvalues will remain. By selecting a β value that puts that knee close to λ_2, our recurrence quickly eliminates mass on all but the largest eigenvector.

Well-Conditioned Augmented Matrix

Consider the recurrence

$$\begin{pmatrix}\tilde{\mathbf{w}}_{t+1}\\ \tilde{\mathbf{w}}_t\end{pmatrix} = \begin{pmatrix}\mathbf{A} & -\beta I\\ I & 0\end{pmatrix}\begin{pmatrix}\tilde{\mathbf{w}}_t\\ \tilde{\mathbf{w}}_{t-1}\end{pmatrix}. \tag{A1}$$

Notice that this is simply power iteration on an augmented matrix. It is straightforward to see that the power iteration with momentum is exactly equivalent to standard power iteration on this augmented matrix, i.e. $\{\tilde{\mathbf{w}}_t\}$ from (**A1**) and $\{\mathbf{w}_t\}$ from (**A**) are the same. As a result, we can take advantage of known power iteration properties when studying our method. In the following proposition, we derive the eigenvalues of the augmented matrix.

Proposition 9. *Suppose a matrix* $\mathbf{A}$ *has eigenvalue- eigenvector pairs* $(\lambda_i, \mathbf{u}_i)_{i=1}^n$*, then the augmented matrix*

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & -\beta I \\ I & 0 \end{pmatrix}$$

has eigenvalue-eigenvector pairs

$$\left(\frac{\lambda_i \pm \sqrt{\lambda_i^2 - 4\beta}}{2}, \begin{pmatrix} \frac{\lambda_i \pm \sqrt{\lambda_i^2 - 4\beta}}{2} \mathbf{u}_i \\ \mathbf{u}_i \end{pmatrix} \right)_{i=1}^{n}.$$

In particular, when $\lambda_2 \leq 2\sqrt{\beta} < \lambda_1$, the relative eigen-gap of this augmented matrix is $1 - \frac{2\sqrt{\beta}}{\lambda_1 + \sqrt{\lambda_1^2 - 4\beta}}$. And the standard power iteration on $\mathbf{M}$ has the convergence rate $\mathcal{O}\left(\left(\frac{2\sqrt{\beta}}{\lambda_1 + \sqrt{\lambda_1^2 - 4\beta}} \right)^{2t} \right)$, which matches the result in Theorem 1.

Now we present the proof of Proposition 9 below.

Proof. For any eigenvalue, eigenvector pair $(\lambda, \mathbf{u})$ of $\mathbf{A}$, let μ be a solution of $\mu^2 - \lambda\mu + \beta = 0$. Suppose that we define

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & -\beta I \\ I & 0 \end{pmatrix}, \quad \mathbf{v} = \begin{pmatrix} \mu\mathbf{u} \\ \mathbf{u} \end{pmatrix}.$$

Then,

$$\mathbf{M}\mathbf{v} = \begin{pmatrix} \mathbf{A} & -\beta I \\ I & 0 \end{pmatrix} \begin{pmatrix} \mu\mathbf{u} \\ \mathbf{u} \end{pmatrix} = \begin{pmatrix} \mu\mathbf{A}\mathbf{u} - \beta\mathbf{u} \\ \mu\mathbf{u} \end{pmatrix} = \begin{pmatrix} \lambda\mu\mathbf{u} - \beta\mathbf{u} \\ \mu\mathbf{u} \end{pmatrix} = \begin{pmatrix} \mu^2\mathbf{u} \\ \mu\mathbf{u} \end{pmatrix} = \mu \begin{pmatrix} \mu\mathbf{u} \\ \mathbf{u} \end{pmatrix} = \mu\mathbf{v}.$$

Thus, $\mathbf{v}$ is an eigenvector of $\mathbf{M}$ with corresponding eigenvalue μ. Doing this for all eigenvectors of $\mathbf{A}$ will produce a complete eigendecomposition of $\mathbf{M}$. □

B Extensions

In this section, we consider several extension based on power method with momentum presented in Section 2. In Section B.1, we will generalize our methods to multiple components case, i.e. finding the top k eigenvalues/eigenvectors and show that it is numerically stable in Section B.2. In Section B.3, we provide some simple heuristics to tune the momentum parameter. In Section B.4, we extend our momentum method into an inhomogeous polynomials recurrence and show that it is optimal in expection with respect to the tail distribution of the tail spectrum of the target matrix $\mathbf{A}$. All the proofs for this section are in Section B.6.

B.1 Block Update for Multiple Components

In this section, we use a block version of our method to compute multiple principal components. In this case, the initial state is a matrix $\mathbf{W}_0 \in \mathbb{R}^{d \times k}$, rather than a single vector. The orthogonal polynomal sequence ($\mathbf{P}$) natually corresponds to the update scheme

$$\mathbf{W}_{t+1} = \mathbf{A}\mathbf{W}_t - \beta\mathbf{W}_{t-1}. \tag{A'}$$

To obtain the convergence result, we use the standard definition from Golub and Van Loan [GVL12] to measure the distance between spaces.

Definition 1. *Given two spaces* $S_1, S_2 \subseteq \mathbb{R}^d$*, the distance between* S_1, S_2 *is defined as*

$$\mathrm{dist}(S_1, S_2) = \|\mathbf{P}_1 - \mathbf{P}_2\|_2,$$

where $\mathbf{P}_i$ *is the orthogonal projection onto* S_i*. Furthermore, when* S_1, S_2 *are matrices, we overload the definition as* $\mathrm{dist}(S_1, S_2) = \mathrm{dist}(range(S_1), range(S_2))$*, where* $range(\cdot)$ *denotes the range space.*

The following lemma shows that we can analyze the convergence rate of any update scheme by studying the growth rate of the corresponding orthogonal polynomial.

Lemma 10. *Given a PSD matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, its top k ($1 \le k < d$) eigenvectors $\mathbf{U}_k \in \mathbb{R}^{d\times k}$, and a matrix $W_0 \in \mathbb{R}^{d\times k}$ such that $d_0 = \operatorname{dist}(\mathbf{U}_k, \mathbf{W}_0) \neq 1$, for any polynomial $p(\cdot)$, we have*

$$\operatorname{dist}(p(\mathbf{A})\mathbf{W}_0, \mathbf{U}_k) \le \frac{d_0}{\sqrt{1-d_0^2}} \cdot \max_{\substack{i=1,\dots,k;\\ j=k+1,\dots,n}} \left|\frac{p(\lambda_j)}{p(\lambda_i)}\right|.$$

The following theorem gives the rate at which the space spanned by the first j columns of $\mathbf{W}_t$ approach the space spanned by the top j eigenvectors.

Theorem 11. *Let $\mathbf{W}_t^{(:j)}$ denote the first j columns of $\mathbf{W}_t$ for $1 \le j \le k$. Given a PSD matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, its top k ($1 \le k < d$) eigenvectors $U_k \in \mathbb{R}^{d\times k}$, a matrix $\mathbf{W}_0 \in \mathbb{R}^{d\times k}$ such that $d_0 = \operatorname{dist}(\mathbf{U}_k, \mathbf{W}_0) \neq 1$, and β such that $2\sqrt{\beta} < \lambda_k$, the update scheme ($\mathbf{P}$) results in the top j-eigenspace converging at a rate of*

$$\operatorname{dist}(\mathbf{W}_t^{(:j)}, \mathbf{U}_j) \le \frac{\operatorname{dist}(\mathbf{W}_0^{(:j)}, \mathbf{U}_j)}{\sqrt{1-\operatorname{dist}(\mathbf{W}_0^{(:j)}, U_j)^2}} \cdot 2\left(\frac{\lambda_{j+1}+\sqrt{\lambda_{j+1}^2-4\beta}}{\lambda_j+\sqrt{\lambda_j^2-4\beta}}\right)^t, j = 1,\dots,k-1$$

$$\operatorname{dist}(\mathbf{W}_t^{(:k)}, \mathbf{U}_k) \le \frac{d_0}{\sqrt{1-d_0^2}} \cdot \begin{cases} 2\left(\frac{2\sqrt{\beta}}{\lambda_k+\sqrt{\lambda_k^2-4\beta}}\right)^t, & \lambda_{k+1} < 2\sqrt{\beta} \\ 2\left(\frac{\lambda_{k+1}+\sqrt{\lambda_{k+1}^2-4\beta}}{\lambda_k+\sqrt{\lambda_k^2-4\beta}}\right)^t, & \lambda_{k+1} \ge 2\sqrt{\beta} \end{cases}.$$

B.2 Stable Implementation of Momentum Methods

In this section, we provide a numerically stable implementation of our momentum method for the multi-component case. This implementation can also be applied in the single component case. Consider the update scheme **A'**. Similar to the unnormalized simultaneous iteration (which essentially is the block version of the power method) ([TBI97, Lecture 28]), as $t \to \infty$, all columns of W_t converge to the multiples of the same dominant eigenvectors of A due to the round-off errors. A common technique to remedy the situation is orthonormalization, which is used in the standard power method. However we cannot simply orthonormalize each $\mathbf{w}_t$ or W_t every iteration because it changes the convergence behavior. Instead, we propose the normalization scheme **A"** to stabilize our method:

$$\begin{aligned} \tilde{\mathbf{W}}_{t+\frac{1}{2}} &= \mathbf{A}\tilde{\mathbf{W}}_t - \beta\tilde{\mathbf{W}}_{t-1}\mathbf{R}_t^{-1}, \\ \tilde{\mathbf{W}}_{t+1} &= \tilde{\mathbf{W}}_{t+\frac{1}{2}}\mathbf{R}_{t+1}^{-1}, \end{aligned} \tag{A"}$$

where $\mathbf{R}_t \in \mathbb{R}^{k\times k}$ is an invertible upper triangular matrix and $\mathbf{R}_1 = I$.

First, Lemma 12 shows that $\tilde{\mathbf{W}}_t$ generated by the normalized update scheme **A"** is the same as $\mathbf{W}_t$ generated by the original update up to a invertible upper triangular matrix factor on the right side. Therefore, the column spaces of $\tilde{\mathbf{W}}_t$ and $\mathbf{W}_t$ are the same, so the normalized update scheme has the same convergence property as the scheme **A'**.

Lemma 12. *Suppose $\{\mathbf{W}_t\}$ and $\{\tilde{\mathbf{W}}_t\}$ are the two sequences generated by (**A'**) and (**A"**) respectively and $\mathbf{W}_0 = \tilde{\mathbf{W}}_0, \mathbf{W}_1 = \tilde{\mathbf{W}}_1$, then $\tilde{\mathbf{W}}_t = \mathbf{W}_t\mathbf{C}_t$ where $\mathbf{C}_t \in \mathbb{R}^{k\times k}$ is an invertible upper triangular matrix for any $t > 0$.*

Now consider the actual implementation of scheme **A"**. One choice of $\mathbf{R}_{t+1}$ is found by using the QR factorization

$$\begin{pmatrix} \tilde{\mathbf{W}}_{t+\frac{1}{2}} \\ \tilde{\mathbf{W}}_t \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{W}}_{t+1} \\ \tilde{\mathbf{W}}_t\mathbf{R}_{t+1}^{-1} \end{pmatrix} \mathbf{R}_{t+1}.$$

In this case, the iteration (**A"**) is indeed *backward stable*. In fact, the update (**A"**) with the choice of $\mathbf{R}_t$ above is equivalent to the normalized simultaneous iteration on the augmented matrix[4] $\hat{\mathbf{A}}$, which has backward stablilty [GVL12]. Also notice that we do not have to materialize the augmented matrix and $\tilde{\mathbf{W}}_{t-1}\mathbf{R}_t^{-1}$ and $\tilde{\mathbf{W}}_{t+\frac{1}{2}}\mathbf{R}_{t+1}^{-1}$ is done implicitly through QR factorization.

[4] In general the normalized simultaneous iteration converges to the Schur vectors of the matrix, not the eigenvectors because the matrix is not Hermitian. However in our particular problem, the normalized simultaneous iteration on the augmented matrix can converge to the eigenvectors of $\mathbf{A}$.

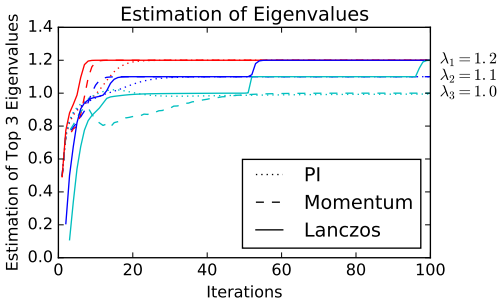


Figure 4: Convergence of standard power iteration, power iteration with momentum, and the Lanczos algorithm to the top eigenvalues of a matrix. Estimation of the first eigenvalue (red), second eigenvalue (blue), and third eigenvalue (cyan) are shown.

We now experimentally demonstrate the efficiency and stability of our method. In Figure 4, we show the estimates of the top three eigenvalues produced by standard power iteration, power iteration with momentum, and the classic Lanczos method. First, notice that the Lanczos iteration is not numerically stable because of the "ghost" eigenvalues problem ([TBI97, Lecture 36]). The estimates of the top three eigenvalues produced by the Lanczos algorithm eventually all converge to the top eigenvalue. In contrast, both standard power iteration and power iteration with momentum successfully find all three eigenvalues. However, standard power iteration takes much longer than power iteration with momentum to converge.

B.3 Tuning Momentum

Our optimal momentum β is determined by λ_2, which will not always be known a priori. We introduce the best heavy ball method to automatically tune β in real time.

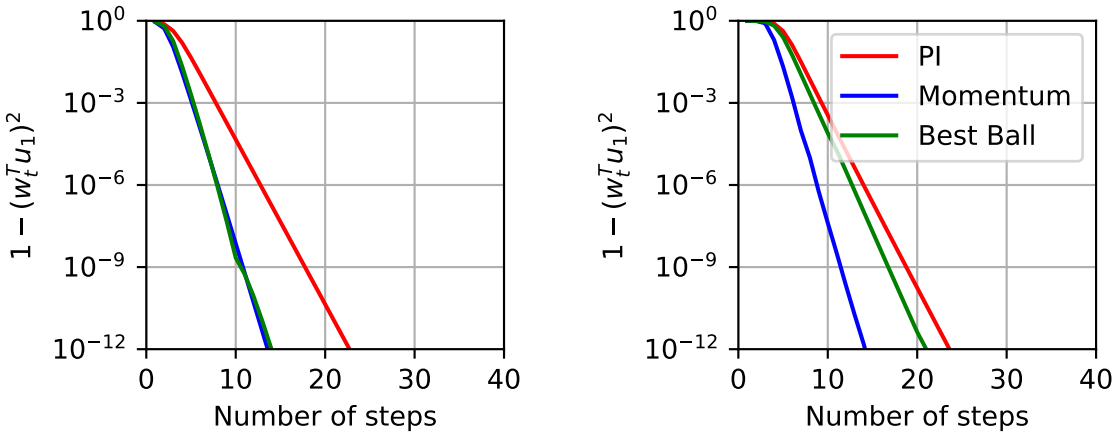
Algorithm 3 Best Heavy Ball

Require: $d \times d$ Matrix $\mathbf{A}$, Number of Iterations T
  $\mathbf{w} \leftarrow$ Random n-dimensional vector
  $\mu \leftarrow (\mathbf{w}^T \mathbf{A} \mathbf{w}) \,/\, (\mathbf{w}^T \mathbf{w})$
  $\beta \leftarrow \mu^2/4$
  for t = 1 **to** T **do**
    Run 10 steps with $2/3\beta, 0.99\beta, \beta, 1.01\beta, 1.5\beta$
    Set β to momentum with largest Rayleigh quotient
  end for
  return $\mathbf{w}$ that gives the largest Rayleigh quotient.

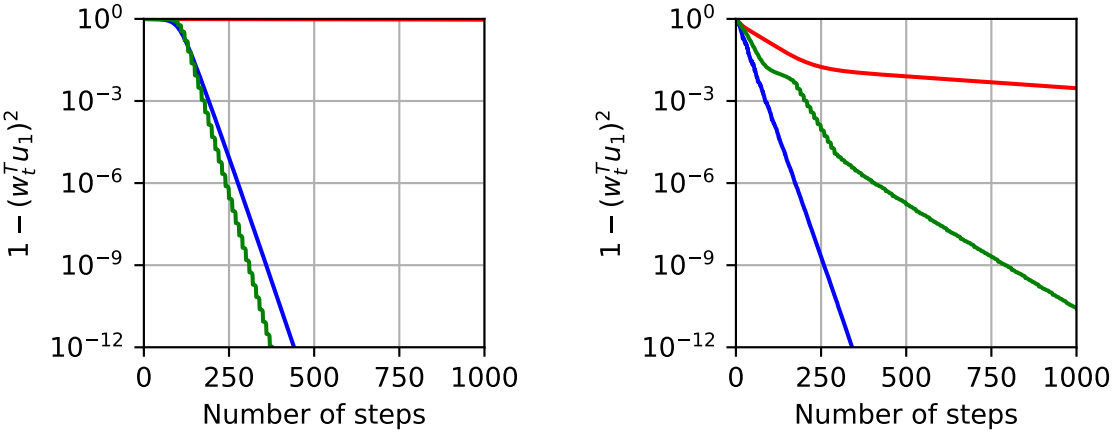
In the heavy ball method, an arbitrary matrix is taken as input, and no information about the matrix is required. A lower bound for the largest eigenvalue is computed by computing the Rayleigh quotient of a random initial vector. This estimate is used to select the initial choice of β. Afterwards, power iteration with momentum is run for 10 steps over a range of values surrounding this choice of β. The performance is measured by the estimation using Rayleigh quotient, i.e., the momentum resulting in the largest Rayleigh quotient[5] is considered the best-performing momentum, and is used as the new center for the search.

Figure 5 compares the performance of power iteration, power iteration with momentum, and the best ball method. Experiments (a) and (b) both have a large eigen-gap, so convergence is fast for all methods. However, even though only a small number of iterations are needed, the best ball method is able to find a suitable value of β, and achieves acceleration. Experiments (c) and (d) have a much smaller eigen-gap, so the acceleration from a well-tuned β is critical for fast convergence. In these experiments, we see that the best ball method is also able to select a β that outperforms power iteration. We also note that the best heavy ball actually outperforms power iteration with momentum in experiment (c), which suggests an inhomogeneous sequence of β's sometimes results in superior performance.

[5]For the multi-component case, we take the sum of all the estimates of top k eigenvalues using Rayleigh quotients.



(a) (b)



(c) (d)

Figure 5: Empirical analysis of the best heavy ball method on four 1000×1000 matrices. The largest eigenvalue of all four matrices is 1. The remaining eigenvalues are: (a) all 0.5. (b) equally spaced from 0 to 0.5. (c) all 0.999. (d) equally spaced from 0 to 0.999.

B.4 Inhomogeneous Polynomial Recurrence

In this section, we present a new algorithm that goes beyond the traditional orthogonal polynomial setting of momentum methods to produce faster convergence of power iteration in some cases. First, we will motivate and derive this method. Suppose that we are trying to run PCA on a matrix A, and that, from experience with other matrices we have encountered in similar settings, we have a rough idea of the spectrum of A. More concretely, suppose that we believe that the largest eigenvalue is λ_1, and the other eigenvalues are independently randomly generated according to some distribution μ (with compact support). As in the momentum case, we want to produce a series of iterates $\mathbf{w}_t$ that approach the dominant eigenvector u_1, and can be written as

$$\mathbf{w}_t = f_t(A)\mathbf{w}_0$$

where f_t is a degree-t polynomial analogous to p_t as defined in (**P**). Our goal is to choose some f_t that can perform better than momentum method, using the extra information we have about the distribution of the spectrum.

The most straightforward way to proceed is to choose the f_t that minimizes the expected error of our estimates over all degree-t polynomials. If we formulate the error of the estimate as

$$\epsilon_t = \frac{\|\mathbf{w}_t\|^2}{(\mathbf{u}_1^T\mathbf{w}_t)^2} - 1 = \sum_{i=2}^{d} \frac{(\mathbf{u}_i^T\mathbf{w}_t)^2}{(\mathbf{u}_1^T\mathbf{w}_t)^2} = \sum_{i=2}^{d} \frac{f_t^2(\lambda_i)(\mathbf{u}_i^T\mathbf{w}_0)^2}{f_t^2(\lambda_1)(\mathbf{u}_1^T\mathbf{w}_0)^2},$$

then

$$\mathbb{E}[\epsilon_t] = \frac{\mathbb{E}_{\lambda\sim\mu}[f_t^2(\lambda)]}{f_t^2(\lambda_1)} \sum_{i=2}^{n} \frac{(\mathbf{u}_i^T\mathbf{w}_0)^2}{(\mathbf{u}_1^T\mathbf{w}_0)^2}.$$

It follows without loss of generality that, to minimize the error, it suffices to solve the optimization problem

$$\begin{array}{ll} \text{minimize} & \mathbb{E}_{\lambda\sim\mu}\left[f_t^2(\lambda)\right] \\ \text{subject to} & f_t(\lambda_1) = 1 \\ & f_t \text{ is a degree-}t \text{ polynomial.} \end{array} \tag{13}$$

This problem statement means that we are interested in finding a update scheme that minimizes the expected power on non-principal components, while keeping a fixed mass on the principal component. We can solve this problem algebraically by decomposing f_t in terms of the family of polynomials $\{q_t\}_{t=0}^{\infty}$ orthogonal with respect to the distribution μ.[6]

This is the unique polynomial family such that q_t is degree-t and

$$\mathbb{E}_{\lambda\sim\mu}\left[q_i(\lambda)q_j(\lambda)\right] = \delta_{i,j}.$$

It turns out that we can solve Equation (13) by representing f_t as a linear combination of orthogonal polynomials from $\{q_t\}_{t=0}^{\infty}$.

Theorem 13. *The degree-t polynomial that solves Equation* (13) *is*

$$f_t^*(\lambda) = \sum_{i=0}^{t} \frac{q_i(\lambda_1)}{\sum_{j=0}^{t} q_i^2(\lambda_1)} q_i(\lambda).$$

Theorem 13 presents the optimal solution as a linear combination of orthogonal polynomials from a particular family. The solution can also be written in the form

$$f_{n+1}^*(x) = f_n^*(x) \cdot \frac{\|\mathbf{r}_n\|^2}{\|\mathbf{r}_{n+1}\|^2} + p_{n+1}(x) \cdot \frac{p_{n+1}(\lambda_1)}{\|\mathbf{r}_{n+1}\|^2}, \tag{14}$$

where $\mathbf{r}_n := \begin{bmatrix} q_1(\lambda_1) & \cdots & q_n(\lambda_1) \end{bmatrix}$. It turns out $f_n^*(x)$ comes from a family of polynomials which has higher-order recurrence. Since the orthogonal polynomial $p_n(x)$ satisfies 3-term recurrence, i.e.

$$p_{n+1}(x) = (\tilde{a}_n x + \tilde{c}_n)p_n(x) - \tilde{b}_n p_{n-1}(x) \tag{15}$$

[6] An orthogonal polynomial family is guaranteed to exist for any distribution with compact support [Chi11].

where $\tilde{a}_n, \tilde{b}_n, \tilde{c}_n \in \mathbb{R}$ depend on the measure μ, then (14) can be simplified into the follow four-term recurrence

$$f_{n+2}(x) = (a_{n+1}x - b_{n+1})f_{n+1}(x) + (c_{n+1}x - d_{n+1})f_n(x) + e_{n+1}f_{n-1}(x). \tag{16}$$

And the derivation can be seen in Appendix B.5.

In (14), the update scheme depends on λ_1, in practice we usually don't know the exact value of λ_1. However we can replace λ_1 with an underestimate $\tilde{\lambda}_1$ such that $\tilde{\lambda}_1 \leq \lambda_1$ and $\tilde{\lambda}_1 > \lambda_2$. The actual algorithm based on the scheme (14) is presented in Algorithm 4.

Algorithm 4 Inhomogeneous Recurrence Algorithm

Require: $d \times d$ Matrix $\mathbf{A}$, Number of Iterations T, Underestimate of $\tilde{\lambda}_1$ ($\lambda_2 \leq \tilde{\lambda}_1 < \lambda_1$)

Initial values: $\mathbf{w}_0 = \mathbf{p}_0, \mathbf{w}_1 = \mathbf{p}_1 \in \mathbb{R}^d$, $r_1, \tilde{p}_0, \tilde{p}_1 \in \mathbb{R}_+$

for t = 1 **to** T **do**

$\mathbf{p}_{t+1} \leftarrow (\tilde{a}_t \cdot \mathbf{A} + \tilde{c}_t)\mathbf{p}_t - \tilde{b}_t\mathbf{p}_{t-1}$

$\tilde{p}_{t+1} \leftarrow (\tilde{a}_t \cdot \tilde{\lambda}_1 + \tilde{c}_t)\tilde{p}_t - \tilde{b}_t\tilde{p}_{t-1}$

$r_{t+1} \leftarrow r_t + \tilde{p}_{t+1}^2$

$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t \cdot r_t/r_{t+1} + \mathbf{p}_{t+1} \cdot \tilde{p}_{t+1}/r_{t+1}$

Normalization:

$$\mathbf{w}_{t+1} \leftarrow \frac{\mathbf{w}_{t+1}}{\|\mathbf{w}_{t+1}\|}, \mathbf{w}_t \leftarrow \frac{\mathbf{w}_t}{\|\mathbf{w}_{t+1}\|}, \mathbf{w}_{t-1} \leftarrow \frac{\mathbf{w}_{t-1}}{\|\mathbf{w}_{t+1}\|},$$
$$\mathbf{p}_{t+1} \leftarrow \frac{\mathbf{p}_{t+1}}{\|\mathbf{w}_{t+1}\|}, \mathbf{p}_t \leftarrow \frac{\mathbf{p}_t}{\|\mathbf{w}_{t+1}\|}, \mathbf{p}_{t-1} \leftarrow \frac{\mathbf{p}_{t-1}}{\|\mathbf{w}_{t+1}\|}$$

end for

return $\mathbf{w}_T$ as the estimation of the largest eigenvector.

The implementation of Algorithm 4 is based on the equation (14) and (15). More concretely, ignoring the normalization procedure, we have $\mathbf{p}_t = p_t(A)\mathbf{w}_0$, $\mathbf{w}_t = f_t(A)\mathbf{w}_0$, $\tilde{p}_t = p_t(\tilde{\lambda}_1)$ and $r_t = \|\mathbf{r}_t\|^2$.

Example. Now we give a concrete example to show the inhomogeneous algorithm works better than momentum method. Figure 6 shows the performance of the different update schemes on a 500×500 matrix. The principal eigenvalue is 1.001, and the remaining eigenvalues are uniformly selected from the interval $[-1, 1]$. This measure corresponds to the Legendre polynomial family. In this example, we see that the loss of the optimal update scheme is essentially always lower than the loss of either power iteration or constant momentum. This indicates that more complex recurrences are required for obtaining ideal performance.

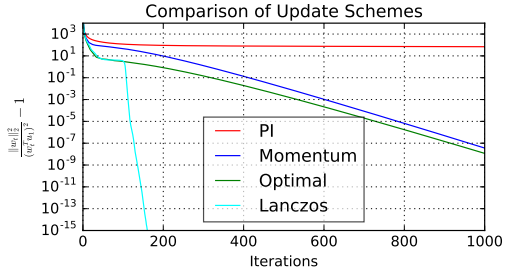


Figure 6: Example comparing the performance of power iteration, constant momentum, and the optimal update scheme.

B.5 Derivation of 4-term Recurrence (16)

First, we restate the inhomogeneous recurrence (14),

$$f_{n+1}(x) = f_n(x) \cdot \frac{\|\mathbf{r}_n\|^2}{\|\mathbf{r}_{n+1}\|^2} + p_{n+1}(x) \cdot \frac{p_{n+1}(\lambda_1)}{\|\mathbf{r}_{n+1}\|^2}.$$

Therefore we have

$$p_{n+1}(x) = \frac{\|\mathbf{r}_{n+1}\|^2}{p_{n+1}(\lambda_1)} \cdot \left(f_{n+1}(x) - f_n(x) \cdot \frac{\|\mathbf{r}_n\|^2}{\|\mathbf{r}_{n+1}\|^2} \right).$$

We assume that the orthogonal polynomial $p_n(x)$ has the following the three-term recurrence,

$$p_{n+1}(x) = \tilde{a}_n x p_n(x) - \tilde{b}_n p_{n-1}(x).$$

Now let's consider $f_{n+2}(x)$,

$$\begin{aligned}
f_{n+2}(x) &= f_{n+1}(x) \cdot \frac{\|\mathbf{r}_{n+1}\|^2}{\|\mathbf{r}_{n+2}\|^2} + p_{n+2}(x) \cdot \frac{p_{n+2}(\lambda_1)}{\|\mathbf{r}_{n+2}\|^2} \\
&= f_{n+1}(x) \cdot \frac{\|\mathbf{r}_{n+1}\|^2}{\|\mathbf{r}_{n+2}\|^2} + \frac{p_{n+2}(\lambda_1)}{\|\mathbf{r}_{n+2}\|^2} \cdot \left(\tilde{a}_{n+1} x p_{n+1}(x) - \tilde{b}_{n+1} p_n(x) \right) \\
&= f_{n+1}(x) \cdot \frac{\|\mathbf{r}_{n+1}\|^2}{\|\mathbf{r}_{n+2}\|^2} + \frac{p_{n+2}(\lambda_1)}{\|\mathbf{r}_{n+2}\|^2} \cdot \left(\tilde{a}_{n+1} x \frac{\|\mathbf{r}_{n+1}\|^2}{p_{n+1}(\lambda_1)} \cdot \left(f_{n+1}(x) - f_n(x) \cdot \frac{\|\mathbf{r}_n\|^2}{\|\mathbf{r}_{n+1}\|^2} \right) \right) \\
&\quad - \frac{p_{n+2}(\lambda_1)}{\|\mathbf{r}_{n+2}\|^2} \cdot \tilde{b}_{n+1} \frac{\|R_n\|^2}{p_n(\lambda_1)} \cdot \left(f_n(x) - f_{n-1}(x) \cdot \frac{\|\mathbf{r}_{n-1}\|^2}{\|R_n\|^2} \right) \\
&= \left(\tilde{a}_{n+1} \frac{p_{n+2}(\lambda_1)\|\mathbf{r}_{n+1}\|^2}{p_{n+1}(\lambda_1)\|\mathbf{r}_{n+2}\|^2} x + \frac{\|\mathbf{r}_{n+1}\|^2}{\|\mathbf{r}_{n+2}\|^2} \right) f_{n+1}(x) \\
&\quad + \left(-\tilde{a}_{n+1} \frac{p_{n+2}(\lambda_1)}{p_{n+1}(\lambda_1)} \frac{\|\mathbf{r}_n\|^2}{\|\mathbf{r}_{n+2}\|^2} x - \tilde{b}_{n+1} \frac{p_{n+2}(\lambda_1)\|\mathbf{r}_n\|^2}{p_n(\lambda_1)\|\mathbf{r}_{n+2}\|^2} \right) f_n(x) \\
&\quad + \left(\tilde{b}_{n+1} \frac{p_{n+2}(\lambda_1)\|\mathbf{r}_{n-1}\|^2}{p_n(\lambda_1)\|\mathbf{r}_{n+2}\|^2} \right) f_n(x)
\end{aligned}$$

Let

$$\begin{aligned}
a_{n+1} &= \tilde{a}_{n+1} \frac{p_{n+2}(\lambda_1)\|\mathbf{r}_{n+1}\|^2}{p_{n+1}(\lambda_1)\|\mathbf{r}_{n+2}\|^2} \\
b_{n+1} &= \frac{\|\mathbf{r}_{n+1}\|^2}{\|\mathbf{r}_{n+2}\|^2} \\
c_{n+1} &= -\tilde{a}_{n+1} \frac{p_{n+2}(\lambda_1)}{p_{n+1}(\lambda_1)} \frac{\|\mathbf{r}_n\|^2}{\|\mathbf{r}_{n+2}\|^2} \\
d_{n+1} &= -\tilde{b}_{n+1} \frac{p_{n+2}(\lambda_1)\|\mathbf{r}_n\|^2}{p_n(\lambda_1)\|\mathbf{r}_{n+2}\|^2} \\
e_{n+1} &= \tilde{b}_{n+1} \frac{p_{n+2}(\lambda_1)\|\mathbf{r}_{n-1}\|^2}{p_n(\lambda_1)\|\mathbf{r}_{n+2}\|^2}
\end{aligned}$$

and we get the 4-term recurrence (16).

B.6 Proofs

Theorem 13. *The degree-t polynomial that solves Equation* (13) *is*

$$f_t^*(\lambda) = \sum_{i=0}^{t} \frac{q_i(\lambda_1)}{\sum_{j=0}^{t} q_i^2(\lambda_1)} q_i(\lambda).$$

Proof. First, we substitute $f_t(\lambda) = \sum_{i=1}^{t} a_{n,i} q_i(\lambda)$ into the optimization problem.

$$\begin{aligned}
\text{minimize} \quad & \mathbb{E}_{\lambda \sim \mu} \left[\left(\textstyle\sum_{i=0}^{t} a_{t,i} q_i(\lambda) \right)^2 \right] \\
\text{subject to} \quad & \textstyle\sum_{i=0}^{t} a_{t,i} q_i(\lambda_1) = 1.
\end{aligned}$$

By taking advantage of the orthogonality of the $q_i(\lambda)$, we have

$$\begin{array}{ll}\text{minimize} & \mathbb{E}_{\lambda\sim\mu}\left[\sum_{i=0}^{t} a_{t,i}^2 q_i^2(\lambda)\right] \\ \text{subject to} & \sum_{i=0}^{t} a_{t,i} q_i(\lambda_1) = 1.\end{array}$$

Then, because $q_i(\lambda)$ is normalized (i.e. $\mathbb{E}_{\lambda\sim\mu}\left[q_i^2(\lambda)\right] = 1$), we have

$$\begin{array}{ll}\text{minimize} & \sum_{i=0}^{t} a_{t,i}^2 \\ \text{subject to} & \sum_{i=0}^{t} a_{t,i} q_i(\lambda_1) = 1.\end{array}$$

This is minimized by

$$a_{t,i} = \frac{p_t(\lambda_1)}{\sum_{j=0}^{t} q_i^2(\lambda_1)}.$$

□

Lemma 10. *Given a PSD matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, its top k ($1 \le k < d$) eigenvectors $\mathbf{U}_k \in \mathbb{R}^{d\times k}$, and a matrix $W_0 \in \mathbb{R}^{d\times k}$ such that $d_0 = \operatorname{dist}(\mathbf{U}_k, \mathbf{W}_0) \neq 1$, for any polynomial $p(\cdot)$, we have*

$$\operatorname{dist}(p(\mathbf{A})\mathbf{W}_0, \mathbf{U}_k) \le \frac{d_0}{\sqrt{1-d_0^2}} \cdot \max_{\substack{i=1,\dots,k;\\ j=k+1,\dots,n}} \left|\frac{p(\lambda_j)}{p(\lambda_i)}\right|.$$

Proof. Suppose $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^T$ is the eigendecomposition of $\mathbf{A}$.Denote $\mathbf{U}_k \in \mathbb{R}^{n\times k}$ as the first k-columns of $\mathbf{U}$ (i.e. the top k eigenvectors of $\mathbf{A}$) and $\mathbf{U}_{-k} \in \mathbb{R}^{n\times(n-k)}$ as the last $n-j$ columns of $\mathbf{U}$ (i.e. the smallest $n-k$ eigenvectors of $\mathbf{A}$). Correspondingly, denote $\Lambda_k \in \mathbb{R}^{k\times k}$ as the top left $k\times k$ block of Λ and $\Lambda_{-k} \in \mathbb{R}^{(n-k)\times(n-k)}$ as the bottom right $(n-k)\times(n-k)$ block of Λ.

Suppose $p(\mathbf{A})\mathbf{W}_0 = \mathbf{Q}\mathbf{R}$ is the QR factorization of $p(\mathbf{A})\mathbf{W}_0$ and $\mathbf{W}_0 = \mathbf{Q}_0\mathbf{R}_0$ is the QR factorization of $\mathbf{W}_0$. Then,

$$\begin{aligned}\mathbf{Q}\mathbf{R} &= p(\mathbf{A})\mathbf{W}_0 \\ &= p(\mathbf{A})\mathbf{Q}_0\mathbf{R}_0 \\ &= \mathbf{U}p(\Lambda)\mathbf{U}^T\mathbf{Q}_0\mathbf{R}_0\end{aligned}$$

Therefore we have

$$\begin{aligned}\mathbf{U}_k^T\mathbf{Q}\mathbf{R} &= p(\Lambda_k)\mathbf{U}_k^T\mathbf{Q}_0\mathbf{R}_0, \\ \mathbf{U}_{-k}^T\mathbf{Q}\mathbf{R} &= p(\Lambda_{-k})\mathbf{U}_{-k}^T\mathbf{Q}_0\mathbf{R}_0.\end{aligned}$$

It is not difficult to see that [GVL12, Theorem 2.5.1, 2.5.2]

$$\begin{aligned}d_0 = \operatorname{dist}(\mathbf{W}_0, \mathbf{U}_k) &= \operatorname{dist}(\mathbf{Q}_0, \mathbf{U}_k) = \|\mathbf{U}_{-k}^T\mathbf{Q}_0\| \\ \sigma_{\min}(\mathbf{U}_k^T\mathbf{Q}_0)^2 + \sigma_{\max}(\mathbf{U}_{-k}^T\mathbf{Q}_0)^2 &= 1.\end{aligned}$$

Now let's compute the distance between $p(\mathbf{A})\mathbf{W}_0$ and $\mathbf{U}_k$,

$$\begin{aligned}\operatorname{dist}(p(\mathbf{A})\mathbf{W}_0, \mathbf{U}_k) &= \|\mathbf{U}_{-k}^T\mathbf{Q}\| \\ &= \|p(\Lambda_{-k})\mathbf{U}_{-k}^T\mathbf{Q}_0\mathbf{R}_0\mathbf{R}^{-1}\| \\ &= \|p(\Lambda_{-k})\mathbf{U}_{-k}^T\mathbf{Q}_0\mathbf{R}_0(p(\Lambda_k)\mathbf{U}_k^T\mathbf{Q}_0\mathbf{R}_0)^{-1}\mathbf{U}_k^T\mathbf{Q}\| \\ &\le \|p(\Lambda_{-k})\|_2\|\mathbf{U}_{-k}^T\mathbf{Q}_0\|\|(p(\Lambda_k))^{-1}\|\|(\mathbf{U}_k^T\mathbf{Q}_0)^{-1}\|\|\mathbf{U}_k^T\mathbf{Q}\| \\ &\le \frac{d_0}{\sqrt{1-d_0^2}} \cdot \max_{\substack{i=1,\dots,k;\\ j=k+1,\dots,n}} \left|\frac{p(\lambda_j)}{p(\lambda_i)}\right|.\end{aligned}$$

□

Theorem 11. *Let $\mathbf{W}_t^{(:j)}$ denote the first j columns of $\mathbf{W}_t$ for $1 \le j \le k$. Given a PSD matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, its top k ($1 \le k < d$) eigenvectors $U_k \in \mathbb{R}^{d\times k}$, a matrix $\mathbf{W}_0 \in \mathbb{R}^{d\times k}$ such that $d_0 = \operatorname{dist}(\mathbf{U}_k, \mathbf{W}_0) \neq 1$, and β such that $2\sqrt{\beta} < \lambda_k$, the update scheme (**P**) results in the top j-eigenspace converging at a rate of*

$$\operatorname{dist}(\mathbf{W}_t^{(:j)}, \mathbf{U}_j) \le \frac{\operatorname{dist}(\mathbf{W}_0^{(:j)}, \mathbf{U}_j)}{\sqrt{1 - \operatorname{dist}(\mathbf{W}_0^{(:j)}, U_j)^2}} \cdot 2\left(\frac{\lambda_{j+1} + \sqrt{\lambda_{j+1}^2 - 4\beta}}{\lambda_j + \sqrt{\lambda_j^2 - 4\beta}}\right)^t, j = 1, \dots, k-1$$

$$\operatorname{dist}(\mathbf{W}_t^{(:k)}, \mathbf{U}_k) \le \frac{d_0}{\sqrt{1-d_0^2}} \cdot \begin{cases} 2\left(\frac{2\sqrt{\beta}}{\lambda_k + \sqrt{\lambda_k^2 - 4\beta}}\right)^t, & \lambda_{k+1} < 2\sqrt{\beta} \\ 2\left(\frac{\lambda_{k+1} + \sqrt{\lambda_{k+1}^2 - 4\beta}}{\lambda_k + \sqrt{\lambda_k^2 - 4\beta}}\right)^t, & \lambda_{k+1} \ge 2\sqrt{\beta} \end{cases}.$$

Proof. First notice that we have $\mathbf{W}_t = p_t(\mathbf{A})\mathbf{W}_0$ for any $t \ge 0$. In fact, we have $\mathbf{W}_t^{(:j)} = p_t(\mathbf{A})\mathbf{W}_0^{(:j)}$ for any $1 \le j \le k$. Hence we can directly apply Lemma 10 and get,

$$\operatorname{dist}(\mathbf{W}_t^{(:j)}, \mathbf{U}_j) \le \frac{\operatorname{dist}(\mathbf{W}_0^{(:j)}, \mathbf{U}_j)}{\sqrt{1 - \operatorname{dist}(\mathbf{W}_0^{(:j)}, \mathbf{U}_j)^2}} \cdot \max_{\substack{i=1,\dots,j; \\ i'=j+1,\dots,n}} \left|\frac{p_t(\lambda_{i'})}{p_t(\lambda_i)}\right|.$$

Now since $2\sqrt{\beta} < \lambda_k$, according to Lemma 20,

$$p_t(\lambda_i) = \begin{cases} \frac{1}{2}\left[\left(\frac{\lambda_i - \sqrt{\lambda_i^2 - 4\beta}}{2}\right)^t + \left(\frac{\lambda_i + \sqrt{\lambda_i^2 - 4\beta}}{2}\right)^t\right], & \lambda_i \ge 2\sqrt{\beta} \\ (\sqrt{\beta})^t \cos\left(t \arccos(\frac{\lambda_i}{2\sqrt{\beta}})\right) & , \lambda_i \le 2\sqrt{\beta} \end{cases}$$

So, plug the polynomials in,

$$\operatorname{dist}(\mathbf{W}_t^{(:j)}, \mathbf{U}_j) \le \frac{\operatorname{dist}(\mathbf{W}_0^{(:j)}, \mathbf{U}_j)}{\sqrt{1 - \operatorname{dist}(\mathbf{W}_0^{(:j)}, \mathbf{U}_j)^2}} \cdot 2\left(\frac{\lambda_{j+1} + \sqrt{\lambda_{j+1}^2 - 4\beta}}{\lambda_j + \sqrt{\lambda_j^2 - 4\beta}}\right)^t, j = 1, \dots, k-1.$$

$$\operatorname{dist}(\mathbf{W}_t^{(:k)}, \mathbf{U}_k) \le \frac{d_0}{\sqrt{1-d_0^2}} \cdot \begin{cases} 2\left(\frac{2\sqrt{\beta}}{\lambda_k + \sqrt{\lambda_k^2 - 4\beta}}\right)^t, & \lambda_{k+1} < 2\sqrt{\beta} \\ 2\left(\frac{\lambda_{k+1} + \sqrt{\lambda_{k+1}^2 - 4\beta}}{\lambda_k + \sqrt{\lambda_k^2 - 4\beta}}\right)^t, & \lambda_{k+1} \ge 2\sqrt{\beta} \end{cases}.$$

□

Lemma 12. *Suppose $\{\mathbf{W}_t\}$ and $\{\tilde{\mathbf{W}}_t\}$ are the two sequences generated by (**A'**) and (**A"**) respectively and $\mathbf{W}_0 = \tilde{\mathbf{W}}_0, \mathbf{W}_1 = \tilde{\mathbf{W}}_1$, then $\tilde{\mathbf{W}}_t = \mathbf{W}_t\mathbf{C}_t$ where $\mathbf{C}_t \in \mathbb{R}^{k\times k}$ is an invertible upper triangular matrix for any $t > 0$.*

Proof. We prove that $\tilde{\mathbf{W}}_t = \mathbf{W}_t\mathbf{C}_t$ where $\mathbf{C}_t = \mathbf{R}_t \cdot \mathbf{R}_{t-1} \cdots \mathbf{R}_0$ by induction. Base case: $\tilde{\mathbf{W}}_0 = \mathbf{W}_0, \tilde{\mathbf{W}}_1 = \mathbf{W}_1$. Assume $\tilde{\mathbf{W}}_i = \mathbf{W}_i\mathbf{C}_i$ holds for any $i \le t$ and consider $\tilde{\mathbf{W}}_{t+1}$ and $\mathbf{W}_{t+1}$,

$$\begin{aligned} \tilde{\mathbf{W}}_{t+1} &= \left(\mathbf{A}\tilde{\mathbf{W}}_t - \beta\tilde{\mathbf{W}}_{t-1}\mathbf{R}_t^{-1}\right)\mathbf{R}_{t+1}^{-1} \\ &= \left(\mathbf{A}\mathbf{W}_t\mathbf{C}_t^{-1} - \beta\tilde{\mathbf{W}}_{t-1}\mathbf{C}_t^{-1}\mathbf{R}_t^{-1}\right)\mathbf{R}_{t+1}^{-1} \\ &= (\mathbf{A}\mathbf{W}_t\mathbf{C}_t^{-1} - \beta\mathbf{W}_{t-1}\mathbf{C}_t^{-1})\mathbf{R}_{t+1}^{-1} \\ &= (\mathbf{A}\mathbf{W}_t - \beta\mathbf{W}_{t-1})\mathbf{C}_t^{-1}\mathbf{R}_{t+1}^{-1} \\ &= \mathbf{W}_{t+1}\mathbf{C}_{t+1}^{-1}. \end{aligned}$$

Therefore, $\mathbf{W}_t\mathbf{C}_t = \tilde{\mathbf{W}}_t$ holds for any $t \ge 0$. □

C Convergence Analysis for Stochastic Power methods with Momentum

In this section we show the detailed analysis for stochastic power memthods with momentum presented in Section 3. Here is the notation we will use for this section.

Notation: $T_t(z)$ is the t-th degree Chebyshev polynomial of the first kind, which satisfies the recurrence,

$$T_{t+1}(z) = 2zT_t(z) - T_{t-1}(z), T_1 = z, U_0 = 1.$$

$U_t(z)$ is the t-th degree Chebyshev polynomial of the second kind, which satisfies the recurrence,

$$U_{t+1}(z) = 2zU_t(z) - U_{t-1}(z), U_1 = 2z, U_0 = 1.$$

$p_t(z)$ is the t-th degree orthogonal polynomial which satisfies the recurrence,

$$p_{t+1}(z) = zp_t(z) - \beta p_{t-1}(z), p_1 = z, p_0 = 1.$$

S_m^n denotes the set of vectors in $\mathbb{N}^n$ with entries that sum to m, i.e.

$$S_m^n = \{\mathbf{k} = (k_1, \cdots, k_n) \in \mathbb{N}^n | \sum_{i=1}^n k_i = m\}.$$

$\otimes$ denotes the Kronecker product.

C.1 Convergence analysis for Algorithm 1

Consider the following stochastic matrix sequence $\{\mathbf{F}_t\}$, which satisfies $\mathbf{F}_0 = I, F_{-1} = \mathbf{0}$, and

$$\mathbf{F}_{t+1} = \mathbf{A}_{t+1}\mathbf{F}_t - \beta\mathbf{F}_{t-1}, \forall t \geq 0. \tag{17}$$

Here $\mathbf{A}_t \in \mathbb{R}^{d\times d}$ is i.i.d. stochastic matrix, with $\mathbb{E}[\mathbf{A}_t] = \mathbf{A}$ and $\mathbb{E}[(\mathbf{A}_t - \mathbf{A}) \otimes (\mathbf{A}_t - \mathbf{A})] = \Sigma$.

Lemma 6. *Suppose $\lambda_1^2 \geq 4\beta$ and $\Sigma = \mathbb{E}[(\mathbf{A}_t - \mathbf{A}) \otimes (\mathbf{A}_t - \mathbf{A})]$. The norm of the covariance of the matrix $\mathbf{F}_t$ is bounded by*

$$\begin{aligned}&\|\mathbb{E}[\mathbf{F}_t \otimes \mathbf{F}_t] - \mathbb{E}[\mathbf{F}_t] \otimes \mathbb{E}[\mathbf{F}_t]\| \\ \leq &\sum_{n=1}^{t} \|\Sigma\|^n \beta^{t-n} \sum_{\mathbf{k}\in S_{t-n}^{n+1}} \prod_{i=1}^{n+1} U_{k_i}^2\left(\frac{\lambda_1}{2\sqrt{\beta}}\right),\end{aligned}$$

where $U_k(\cdot)$ is the Chebyshev polynomial of the second kind, and S_m^n denotes the set of vectors in $\mathbb{N}^n$ with entries that sum to m, i.e.

$$S_m^n = \{\mathbf{k} = (k_1, \cdots, k_n) \in \mathbb{N}^n \mid \textstyle\sum_{i=1}^n k_i = m\}.$$

Proof. First, let

$$\mathbf{M}_t = \begin{bmatrix} \mathbf{A}_t & -\beta I \\ I & 0 \end{bmatrix}, \mathbf{M} = \begin{bmatrix} \mathbf{A} & -\beta I \\ I & 0 \end{bmatrix}$$

and

$$\mathbf{E}_1 = \begin{bmatrix} I \\ 0 \end{bmatrix}.$$

and then we have

$$\mathbf{F}_t = \mathbf{E}_1^T \cdot \mathbf{M}_t \cdots \mathbf{M}_1 \cdot \mathbf{E}_1.$$

Therefore we have the second moment,

$$\begin{aligned}\mathbb{E}[\mathbf{F}_t \otimes \mathbf{F}_t] &= \mathbb{E}\left[\left(\mathbf{E}_1^T \cdot \mathbf{M}_t \cdot \mathbf{M}_{t-1} \cdots \mathbf{M}_1 \cdot \mathbf{E}_1\right)^2\right] \\ &= \mathbb{E}\left[\left(\mathbf{E}_1^T \cdot \mathbf{M}_t \cdot \mathbf{M}_{t-1} \cdots \mathbf{M}_1 \cdot \mathbf{E}_1\right)^{\otimes 2}\right] \\ &= \mathbb{E}\left[(\mathbf{E}_1^T)^{\otimes 2} \cdot \mathbf{M}_t^{\otimes 2} \cdot \mathbf{M}_{t-1}^{\otimes 2} \cdots \mathbf{M}_1^{\otimes 2} \cdot \mathbf{E}_1^{\otimes 2}\right] \\ &= (\mathbf{E}_1 \otimes \mathbf{E}_1)^T \cdot \mathbb{E}\left[\mathbf{M}_t^{\otimes 2}\right] \cdot \mathbb{E}\left[\mathbf{M}_{t-1}^{\otimes 2}\right] \cdots \mathbb{E}\left[\mathbf{M}_1^{\otimes 2}\right] \cdot (\mathbf{E}_1 \otimes \mathbf{E}_1)\end{aligned}$$

Since the $\mathbf{M}_i$ are i.i.d. as before, all the expected values in the last expression above will be the same.

$$\begin{aligned}
\mathbb{E}\left[\mathbf{M}_t^{\otimes 2}\right] &= \mathbb{E}\left[\begin{bmatrix} \mathbf{A}_t & -\beta \\ I & 0 \end{bmatrix}^{\otimes 2}\right] \\
&= \mathbb{E}\left[\left(\begin{bmatrix} \mathbf{A} & -\beta \\ I & 0 \end{bmatrix} + \begin{bmatrix} \mathbf{A}_t - \mathbf{A} & 0 \\ 0 & 0 \end{bmatrix}\right)^{\otimes 2}\right] \\
&= \mathbb{E}\left[\left(\mathbf{M} + \mathbf{E}_1(\mathbf{A}_t - \mathbf{A})\mathbf{E}_1^T\right)^{\otimes 2}\right] \\
&= \mathbf{M}\otimes\mathbf{M} + (\mathbf{E}_1\otimes\mathbf{E}_1)\mathbb{E}\left[(\mathbf{A}_t - \mathbf{A})\otimes(\mathbf{A}_t - \mathbf{A})\right](\mathbf{E}_1\otimes\mathbf{E}_1)^T \\
&= \mathbf{M}\otimes\mathbf{M} + (\mathbf{E}_1\otimes\mathbf{E}_1)\Sigma(\mathbf{E}_1\otimes\mathbf{E}_1)^T.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\mathbf{F}_t\otimes\mathbf{F}_t\right] &= (\mathbf{E}_1\otimes\mathbf{E}_1)^T\left(\mathbf{M}\otimes\mathbf{M} + (\mathbf{E}_1\otimes\mathbf{E}_1)\Sigma(\mathbf{E}_1\otimes\mathbf{E}_1)^T\right)^t(\mathbf{E}_1\otimes\mathbf{E}_1) \\
&= \sum_{n=0}^{t}\sum_{\mathbf{k}\in S_{t-n}^{n+1}}\prod_{i=2}^{n+1}\left(p_{k_i}^{\otimes 2}(\mathbf{A};\beta)\cdot\Sigma\right)\cdot p_{k_1}^{\otimes 2}(\mathbf{A};\beta).
\end{aligned}$$

The last equality follows from the binomial expansion of matrices (Fact 21).

And further we have

$$\mathbb{E}\left[\mathbf{F}_t\otimes\mathbf{F}_t\right] - \mathbb{E}\left[\mathbf{F}_t\right]\otimes\mathbb{E}\left[\mathbf{F}_t\right] = \sum_{n=1}^{t}\sum_{\mathbf{k}\in S_{t-n}^{n+1}}\prod_{i=2}^{n+1}\left(p_{k_i}^{\otimes 2}(\mathbf{A};\beta)\cdot\Sigma\right)\cdot p_{k_1}^{\otimes 2}(\mathbf{A};\beta).$$

Taking the norm, and knowing $0 \preceq \mathbf{A} \preceq \lambda_1 I$,

$$\begin{aligned}
\left\|\mathbb{E}\left[\mathbf{F}_t\otimes\mathbf{F}_t\right] - \mathbb{E}\left[\mathbf{F}_t\right]\otimes\mathbb{E}\left[\mathbf{F}_t\right]\right\| &\le \sum_{n=1}^{t}\sum_{\mathbf{k}\in S_{t-n}^{n+1}}\prod_{i=2}^{n+1}\left(\left\|p_{k_i}^{\otimes 2}(\mathbf{A};\beta)\right\|\cdot\|\Sigma\|\right)\cdot\left\|p_{k_1}^{\otimes 2}(\mathbf{A};\beta)\right\| \\
&= \sum_{n=1}^{t}\|\Sigma\|^n\sum_{\mathbf{k}\in S_{t-n}^{n+1}}\prod_{i=1}^{n+1}\left\|p_{k_i}(\mathbf{A};\beta)\right\|^2 \\
&\le \sum_{n=1}^{t}\|\Sigma\|^n\sum_{\mathbf{k}\in S_{t-n}^{n+1}}\prod_{i=1}^{n+1}p_{k_i}^2(\lambda_1;\beta) \\
&= \sum_{n=1}^{t}\|\Sigma\|^n\beta^{t-n}\sum_{\mathbf{k}\in S_{t-n}^{n+1}}\prod_{i=1}^{n+1}U_{k_i}^2\left(\frac{\lambda_1}{2\sqrt{\beta}}\right).
\end{aligned}$$

The last equality follows from the fact that $p_t(x) = (\sqrt{\beta})^t\cdot U_t(\frac{x}{2\sqrt{\beta}})$. This is what we wanted to show. □

Remark. It is straightforward to see that this analysis can be applied to the case $\beta = 0$ which is the power iteration case.

Corollary 14. *Under the same condidtion in Lemma 6, we have*

$$\left\|\mathbb{E}\left[\mathbf{F}_t\otimes\mathbf{F}_t\right] - \mathbb{E}\left[\mathbf{F}_t\right]\otimes\mathbb{E}\left[\mathbf{F}_t\right]\right\| \le p_t^2\left(\lambda_1;\beta\right)\left(\exp\left(\frac{4\|\Sigma\|t}{\lambda_1^2 - 4\beta}\right) - 1\right).$$

Further if

$$\|\Sigma\| \le \frac{\lambda_1^2 - 4\beta}{4t}, \tag{18}$$

we have

$$\left\|\mathbb{E}\left[\mathbf{F}_t\otimes\mathbf{F}_t\right] - \mathbb{E}\left[\mathbf{F}_t\right]\otimes\mathbb{E}\left[\mathbf{F}_t\right]\right\| \le p_t^2\left(\lambda_1;\beta\right)\cdot\frac{8\|\Sigma\|t}{\lambda_1^2 - 4\beta}.$$

Proof. First, according to Lemma 6 and 22, we have

$$\begin{aligned}
\|\mathbb{E}\left[\mathbf{F}_t \otimes \mathbf{F}_t\right] - \mathbb{E}\left[\mathbf{F}_t\right] \otimes \mathbb{E}\left[\mathbf{F}_t\right]\| &\leq \sum_{n=1}^{t} \|\Sigma\|^n \beta^{t-n} \sum_{\mathbf{k} \in S_{t-n}^{n+1}} \prod_{i=1}^{n+1} U_{k_i}^2 \left(\frac{\lambda_1}{2\sqrt{\beta}}\right) \\
&\leq \sum_{n=1}^{t} \|\Sigma\|^n \beta^{t-n} \sum_{\mathbf{k} \in S_{t-n}^{n+1}} U_{\sum_{i=1}^{n+1} k_i + n}^2 \left(\frac{\lambda_1}{2\sqrt{\beta}}\right) \cdot \frac{1}{\left(\left(\frac{\lambda_1^2}{4\beta}\right) - 1\right)^n} \\
&= \beta^t U_t^2 \left(\frac{\lambda_1}{2\sqrt{\beta}}\right) \sum_{n=1}^{t} \binom{t}{t-n} \frac{4^n \|\Sigma\|^n}{(\lambda_1^2 - 4\beta)^n} \\
&= p_t^2(\lambda_1) \cdot \left(\left(\frac{4\|\Sigma\|}{\lambda_1^2 - 4\beta} + 1\right)^t - 1\right) \\
&\leq p_t^2(\lambda_1) \cdot \left(\exp\left(\frac{4\|\Sigma\| t}{\lambda_1^2 - 4\beta}\right) - 1\right)
\end{aligned}$$

If $\|\Sigma\| \leq \frac{\lambda_1^2 - 4\beta}{4t}$, by the fact that $e^x \leq 1 + 2x$ for any $x \in (0, 1)$, then we have

$$\|\mathbb{E}\left[\mathbf{F}_t \otimes \mathbf{F}_t\right] - \mathbb{E}\left[\mathbf{F}_t\right] \otimes \mathbb{E}\left[\mathbf{F}_t\right]\| \leq p_t^2\left(\lambda_1; \beta\right) \cdot \frac{8\|\Sigma\| t}{\lambda_1^2 - 4\beta}.$$

which is the desired result. □

Corollary 15. *For any $\mathbf{w}_0 \in \mathbb{R}^d$ such that $\|\mathbf{w}_0\| = 1$, we have*

$$\|\mathbb{E}\left[\mathbf{F}_t\mathbf{w}_0 \otimes \mathbf{F}_t\mathbf{w}_0\right] - \mathbb{E}\left[\mathbf{F}_t\mathbf{w}_0\right] \otimes \mathbb{E}\left[\mathbf{F}_t\mathbf{w}_0\right]\| \leq p_t^2(\lambda_1; \beta) \cdot \frac{8\|\Sigma\| t}{\lambda_1^2 - 4\beta}.$$

Proof. Using the mixed-product property of Kronecker product, we have

$$\begin{aligned}
\|\mathbb{E}\left[\mathbf{F}_t\mathbf{w}_0 \otimes \mathbf{F}_t\mathbf{w}_0\right] - \mathbb{E}\left[\mathbf{F}_t\mathbf{w}_0\right] \otimes \mathbb{E}\left[\mathbf{F}_t\mathbf{w}_0\right]\| &= \|\mathbb{E}\left[\mathbf{F}_t \otimes \mathbf{F}_t\right] \cdot (\mathbf{w}_0 \otimes \mathbf{w}_0) - (\mathbb{E}\left[\mathbf{F}_t\right] \otimes \mathbb{E}\left[\mathbf{F}_t\right]) \cdot (\mathbf{w}_0 \otimes \mathbf{w}_0)\| \\
&\leq \|\mathbb{E}\left[\mathbf{F}_t \otimes \mathbf{F}_t\right] - \mathbb{E}\left[\mathbf{F}_t\right] \otimes \mathbb{E}\left[\mathbf{F}_t\right]\| \cdot \|\mathbf{w}_0 \otimes \mathbf{w}_0\| \\
&= \|\mathbb{E}\left[\mathbf{F}_t \otimes \mathbf{F}_t\right] - \mathbb{E}\left[\mathbf{F}_t\right] \otimes \mathbb{E}\left[\mathbf{F}_t\right]\| \cdot \|\mathbf{w}_0\|^2 \\
&\leq p_t^2(\lambda_1; \beta) \cdot \frac{8\|\Sigma\| t}{\lambda_1^2 - 4\beta}.
\end{aligned}$$

□

Corollary 16. *For any $\mathbf{u}, \mathbf{w}_0 \in \mathbb{R}^d$ such that $\|\mathbf{u}\| = 1, \|\mathbf{w}_0\| = 1$, we have*

$$\mathbf{Var}\left[\mathbf{u}^T\mathbf{F}_t\mathbf{w}_0\right] \leq p_t^2(\lambda_1; \beta) \cdot \frac{8\|\Sigma\| t}{\lambda_1^2 - 4\beta}.$$

Proof.

$$\begin{aligned}
\mathbf{Var}\left[\mathbf{u}^T\mathbf{F}_t\mathbf{w}_0\right] &= \mathbb{E}\left[(\mathbf{u}^T\mathbf{F}_t\mathbf{w}_0)^2\right] - (\mathbb{E}\left[\mathbf{u}^T\mathbf{F}_t\mathbf{w}_0\right])^2 \\
&= \mathbb{E}\left[(\mathbf{u}^T\mathbf{F}_t\mathbf{w}_0) \otimes (\mathbf{u}^T\mathbf{F}_t\mathbf{w}_0)\right] - \mathbb{E}\left[(\mathbf{u}^T\mathbf{F}_t\mathbf{w}_0)\right] \otimes \mathbb{E}\left[(\mathbf{u}^T\mathbf{F}_t\mathbf{w}_0)\right] \\
&= (\mathbf{u} \otimes \mathbf{u})^T \cdot (\mathbb{E}\left[\mathbf{F}_t\mathbf{w}_0 \otimes \mathbf{F}_t\mathbf{w}_0\right] - \mathbb{E}\left[\mathbf{F}_t\mathbf{w}_0\right] \otimes \mathbb{E}\left[\mathbf{F}_t\mathbf{w}_0\right]) \\
&\leq \|\mathbf{u} \otimes \mathbf{u}\| \cdot \|\mathbb{E}\left[\mathbf{F}_t\mathbf{w}_0 \otimes \mathbf{F}_t\mathbf{w}_0\right] - \mathbb{E}\left[\mathbf{F}_t\mathbf{w}_0\right] \otimes \mathbb{E}\left[\mathbf{F}_t\mathbf{w}_0\right]\| \\
&\leq p_t^2(\lambda_1; \beta) \cdot \frac{8\|\Sigma\| t}{\lambda_1^2 - 4\beta}.
\end{aligned}$$

The last inequality follows from Corollary 15. □

Corollary 17. *Suppose* $\mathbf{u}_2, \cdots, \mathbf{u}_d$ *are the last* $d-1$ *eigenvectors of* $\mathbf{A}$ *and* $2\sqrt{\beta} \in [\lambda_2, \lambda_1)$. *For any fixed* $\mathbf{w}_0 \in \mathbb{R}^d$ *such that* $\|\mathbf{w}_0\| = 1$, $\delta \in (0,1)$, *with probability at least* $1-\delta$, *we have*

$$\sum_{i=2}^{d} (\mathbf{u}_i^T \mathbf{F}_t \mathbf{w}_0)^2 \leq p_t^2(\lambda_1;\beta) \cdot \left(\frac{8\sqrt{d}\|\Sigma\| t}{\delta(\lambda_1^2 - 4\beta)} + \frac{p_t^2(2\sqrt{\beta};\beta)}{\delta p_t^2(\lambda_1;\beta)} \right)$$

Proof. First, we consider the second moment

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=2}^{d} (\mathbf{u}_i^T \mathbf{F}_t \mathbf{w}_0)^2\right] &= \sum_{i=2}^{d} \mathbb{E}\left[(\mathbf{u}_i^T \mathbf{F}_t \mathbf{w}_0)^2\right] \\
&= \sum_{i=2}^{d} \left[\mathbb{E}\left[(\mathbf{u}_i \mathbf{F}_t \mathbf{w}_0)^{\otimes 2}\right] - \mathbb{E}\left[\mathbf{u}_i^T \mathbf{F}_t \mathbf{w}_0\right]^{\otimes 2}\right] + \sum_{i=2}^{d} \mathbb{E}\left[\mathbf{u}_i \mathbf{F}_t \mathbf{w}_0\right]^2 \\
&= (\sum_{i=2}^{d} \mathbf{u}_i^{\otimes 2})^T \cdot \left(\mathbb{E}\left[\mathbf{F}_t^{\otimes 2}\right] - \mathbb{E}\left[\mathbf{F}_t\right]^{\otimes 2}\right) \cdot \mathbf{w}_0^{\otimes 2} + \sum_{i=2}^{d} p_t^2(\lambda_i;\beta)(\mathbf{u}_i^T \mathbf{w}_0)^2 \\
&\leq \left\| \sum_{i=2}^{d} \mathbf{u}_i^{\otimes 2} \right\| \cdot \left\| \mathbb{E}\left[\mathbf{F}_t^{\otimes 2}\right] - \mathbb{E}\left[\mathbf{F}_t\right]^{\otimes 2} \right\| \cdot \left\| \mathbf{w}_0^{\otimes 2} \right\| + p_t^2(2\sqrt{\beta};\beta) \\
&\leq \sqrt{d} \cdot p_t^2(\lambda_1;\beta) \frac{8\|\Sigma\| t}{\lambda_1^2 - 4\beta} + p_t^2(2\sqrt{\beta};\beta) \\
&= p_t^2(\lambda_1;\beta) \cdot \left(\frac{8\sqrt{d}\|\Sigma\| t}{(\lambda_1^2 - 4\beta)} + \frac{p_t^2(2\sqrt{\beta};\beta)}{p_t^2(\lambda_1;\beta)} \right)
\end{aligned}$$

The last inequality follows from the fact $\left\| \sum_{i=2}^{d} \mathbf{u}_i^{\otimes 2} \right\| = \sqrt{d-1}$. For any $\delta \in (0,1)$, by Markov's inequality we can get the desired result. □

Theorem 3. *Suppose we run Algorithm 1 with* $2\sqrt{\beta} \in [\lambda_2, \lambda_1)$. *Let* $\Sigma = \mathbb{E}\left[(\mathbf{A}_t - \mathbf{A}) \otimes (\mathbf{A}_t - \mathbf{A})\right]$[7]. *Suppose that* $\|\mathbf{w}_0\| = 1$ *and* $\left|\mathbf{u}_1^T \mathbf{w}_0\right| \geq 1/2$. *For any* $\delta \in (0,1)$ *and* $\epsilon \in (0,1)$, *if*

$$\begin{aligned}
T &= \frac{\sqrt{\beta}}{\sqrt{\lambda_1^2 - 4\beta}} \log\left(\frac{32}{\delta\epsilon}\right), \\
\|\Sigma\| &\leq \frac{(\lambda_1^2 - 4\beta)\delta\epsilon}{256\sqrt{d}T} = \frac{(\lambda_1^2 - 4\beta)^{3/2}\delta\epsilon}{256\sqrt{d}\sqrt{\beta}} \log^{-1}\left(\frac{32}{\delta\epsilon}\right),
\end{aligned} \tag{4}$$

then with probability at least $1-2\delta$, *we have* $1 - (\mathbf{u}_1^T \mathbf{w}_T)^2 \leq \epsilon$.

Proof. In order to bound $1 - (\mathbf{u}_1^T \mathbf{w}_t)^2$, it is equivalent to bound $1 - \frac{(\mathbf{u}_1^T \mathbf{F}_t \mathbf{w}_0)^2}{\|\mathbf{F}_t \mathbf{w}_0\|^2}$.

$$1 - \frac{(\mathbf{u}_1^T \mathbf{F}_t \mathbf{w}_0)^2}{\|\mathbf{F}_t \mathbf{w}_0\|^2} \leq \frac{\sum_{i=2}^{d} (\mathbf{u}_i^T \mathbf{F}_t \mathbf{w}_0)^2}{(\mathbf{u}_1^T \mathbf{F}_t \mathbf{w}_0)^2}.$$

Notice that

$$\mathbb{E}\left[\mathbf{u}_i^T \mathbf{F}_t \mathbf{w}_0\right] = p_t(\lambda_i;\beta) \mathbf{u}_i^T \mathbf{w}_0.$$

According to Corollary 16, by Chebyshev's inequality, for any $\delta \in (0,1)$,we have

$$\Pr\left(\left|\mathbf{u}_1^T \mathbf{F}_t \mathbf{w}_0 - p_t(\lambda_1;\beta) \mathbf{u}_1^T \mathbf{w}_0\right| \geq \frac{1}{\sqrt{\delta}} \cdot p_t\left(\lambda_1;\beta\right) \cdot \sqrt{\frac{8\|\Sigma\| t}{\lambda_1^2 - 4\beta}} \right) \leq \delta.$$

That is,

$$\Pr\left(\left|\mathbf{u}_1^T \mathbf{F}_t \mathbf{w}_0\right| \leq p_t(\lambda_1;\beta) \left(\left|\mathbf{u}_1^T \mathbf{w}_0\right| - \sqrt{\frac{8\|\Sigma\| t}{(\lambda_1^2 - 4\beta)\delta}} \right) \right) \leq \delta.$$

[7] $\otimes$ denotes the Kronecker product.

On the other hand,according to Corollary 17,

$$\Pr\left(\sum_{i=2}^{d}(\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0)^2 \geq p_t^2\left(\lambda_1;\beta\right)\left(\frac{8\sqrt{d}\|\Sigma\|t}{(\lambda_1^2-4\beta)\delta}+\frac{p_t^2(2\sqrt{\beta};\beta)}{\delta p_t^2\left(\lambda_1;\beta\right)}\right)\right)\leq\delta.$$

It follows by a union bound that

$$\Pr\left(\frac{\sum_{i=2}^{d}(\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0)^2}{(\mathbf{u}_1^T\mathbf{F}_t\mathbf{w}_0)^2}\geq\left(\frac{8\sqrt{d}\|\Sigma\|t}{(\lambda_1^2-4\beta)\delta}+\frac{p_t^2(2\sqrt{\beta};\beta)}{\delta p_t^2\left(\lambda_1;\beta\right)}\right)\left(\left|\mathbf{u}_1^T\mathbf{w}_0\right|-\sqrt{\frac{8\|\Sigma\|t}{(\lambda_1^2-4\beta)\delta}}\right)^{-2}\right)\leq 2\delta.$$

For any $\epsilon\in(0,1/16)$, when

$$t=\frac{\sqrt{\beta}}{\sqrt{\lambda_1^2-4\beta}}\log(\frac{1}{\delta\epsilon}),$$

we have

$$\frac{p_t^2(2\sqrt{\beta};\beta)}{\delta p_t^2(\lambda_1;\beta)}\leq\epsilon.$$

When

$$\|\Sigma\|\leq\frac{(\lambda_1^2-4\beta)\delta\epsilon}{8\sqrt{d}t}=\frac{(\lambda_1^2-4\beta)^{3/2}\delta\epsilon}{8\sqrt{d}\sqrt{\beta}}\log^{-1}\left(\frac{1}{\delta\epsilon}\right),$$

we have

$$\frac{8\sqrt{d}\|\Sigma\|t}{\delta(\lambda_1^2-4\beta)}\leq\epsilon.$$

With both conditions, we have with probability at least $1-2\delta$,

$$\frac{\sum_{i=2}^{d}(\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0)^2}{(\mathbf{u}_1^T\mathbf{F}_t\mathbf{w}_0)^2}\leq\frac{2\epsilon}{\left(|\mathbf{u}_1\mathbf{w}_0|-\sqrt{\frac{\epsilon}{\sqrt{d}}}\right)^2}\leq 32\epsilon.$$

Rescale ϵ down by a factor of 32 would lead to the desired result. □

To prove Corollary 4, we can simply use the fact that

$$\|\Sigma\|=\left\|\mathbb{E}\left[(\mathbf{A}_t-\mathbf{A})^{\otimes 2}\right]\right\|\leq\mathbb{E}\left[\left\|(\mathbf{A}_t-\mathbf{A})^{\otimes 2}\right\|\right]=\mathbb{E}\left[\|\mathbf{A}_t-\mathbf{A}\|^2\right]=\frac{\sigma^2}{s},$$

and we immediately get a sufficent condition of batch size to satisfy the variance condition (4) in Theorem 3, and that is

$$s\geq\frac{256\sqrt{d}\sigma^2T}{(\lambda_1^2-4\beta)\delta\epsilon}=\frac{256\sqrt{d}\sqrt{\beta}\sigma^2}{(\lambda_1^2-4\beta)^{3/2}\delta\epsilon}\log\left(\frac{32}{\delta\epsilon}\right).$$

So with Theorem 3, we get the result of Corollary 4.

C.2 Convergence analysis for Algorithm 2

For the convergence analysis for Algorithm 2, we first analyze the convergence for one epoch. For that, Consider the following stochastic matrix sequence $\{\mathbf{F}_t\}$, which satisfies $\mathbf{F}_0=I,\mathbf{F}_{-1}=\mathbf{0}$, and

$$\mathbf{F}_{t+1}=[\mathbf{A}+(\mathbf{A}_{t+1}-\mathbf{A})(I-\mathbf{w}_0\mathbf{w}_0^T)]\mathbf{F}_t-\beta\mathbf{F}_{t-1},\forall t\geq 0. \tag{19}$$

Here $\mathbf{A}_t\in\mathbb{R}^{d\times d}$ is i.i.d. stochastic matrix, with $\mathbb{E}\left[\mathbf{A}_t\right]=\mathbf{A}$ and $\mathbb{E}\left[(\mathbf{A}_t-\mathbf{A})\otimes(\mathbf{A}_t-\mathbf{A})\right]=\Sigma$. And $\mathbf{w}_0\in\mathbb{R}^d$ is a fixed unit vector.

Lemma 7. *Suppose $\lambda_1^2 \geq 4\beta$. Let $\mathbf{w}_0 \in \mathbb{R}^d$ be a unit vector, $\theta = 1 - (\mathbf{u}_1^T \mathbf{w}_0)^2$, and*

$$\Sigma = \mathbb{E}\left[\left(\frac{1}{s}\sum_{i=1}^{s} \tilde{\mathbf{A}}_{t_i} - \mathbf{A}\right) \otimes \left(\frac{1}{s}\sum_{i=1}^{s} \tilde{\mathbf{A}}_{t_i} - \mathbf{A}\right)\right].$$

Then, the norm of the covariance will be bounded by

$$\begin{aligned}&\left\|\mathbb{E}\left[\mathbf{F}_t\mathbf{w}_0 \otimes \mathbf{F}_t\mathbf{w}_0\right] - \mathbb{E}\left[\mathbf{F}_t\mathbf{w}_0\right] \otimes \mathbb{E}\left[\mathbf{F}_t\mathbf{w}_0\right]\right\| \\ \leq\ & 4\theta \cdot \sum_{n=1}^{t} \|\Sigma\|^n \beta^{t-n} \sum_{\mathbf{k}\in S_{t-n}^{n+1}} \prod_{i=1}^{n+1} U_{k_i}^2\left(\frac{\lambda_1}{2\sqrt{\beta}}\right).\end{aligned}$$

Proof. First, let

$$\mathbf{M}_t = \begin{bmatrix} \mathbf{A} + (\mathbf{A}_t - \mathbf{A})(I - \mathbf{w}_0\mathbf{w}_0^T) & -\beta I \\ I & 0 \end{bmatrix}, \mathbf{M} = \begin{bmatrix} \mathbf{A} & -\beta I \\ I & 0 \end{bmatrix}$$

and

$$\mathbf{E}_1 = \begin{bmatrix} I \\ 0 \end{bmatrix}.$$

and then we have

$$\mathbf{F}_t = \mathbf{E}_1^T \cdot \mathbf{M}_t \cdots \mathbf{M}_1 \cdot \mathbf{E}_1.$$

Therefore we have the second moment,

$$\mathbb{E}\left[\mathbf{F}_t \otimes \mathbf{F}_t\right] = (\mathbf{E}_1 \otimes \mathbf{E}_1)^T \cdot \mathbb{E}\left[\mathbf{M}_t^{\otimes 2}\right] \cdot \mathbb{E}\left[\mathbf{M}_{t-1}^{\otimes 2}\right] \cdots \mathbb{E}\left[\mathbf{M}_1^{\otimes 2}\right] \cdot (\mathbf{E}_1 \otimes \mathbf{E}_1)$$

Since the A_i are i.i.d. as before, all the expected values in the last expression above will be the same.

$$\begin{aligned}\mathbb{E}\left[\mathbf{M}_t^{\otimes 2}\right] &= \mathbb{E}\left[\begin{bmatrix} \mathbf{A} + (\mathbf{A}_t - \mathbf{A})(I - \mathbf{w}_0\mathbf{w}_0^T) & -\beta \\ I & 0 \end{bmatrix}^{\otimes 2}\right] \\ &= \mathbb{E}\left[\left(\begin{bmatrix} \mathbf{A} & -\beta \\ I & 0 \end{bmatrix} + \begin{bmatrix} (\mathbf{A}_t - \mathbf{A})(I - \mathbf{w}_0\mathbf{w}_0^T) & 0 \\ 0 & 0 \end{bmatrix}\right)^{\otimes 2}\right] \\ &= \mathbb{E}\left[\left(\mathbf{M} + \mathbf{E}_1(\mathbf{A}_t - \mathbf{A})(I - \mathbf{w}_0\mathbf{w}_0^T)\mathbf{E}_1^T\right)^{\otimes 2}\right] \\ &= \mathbf{M} \otimes \mathbf{M} + (\mathbf{E}_1 \otimes \mathbf{E}_1)\mathbb{E}\left[(\mathbf{A}_t - \mathbf{A}) \otimes (\mathbf{A}_t - \mathbf{A})\right](I - \mathbf{w}_0\mathbf{w}_0^T)^{\otimes 2}(\mathbf{E}_1 \otimes \mathbf{E}_1)^T \\ &= \mathbf{M} \otimes \mathbf{M} + (\mathbf{E}_1 \otimes \mathbf{E}_1)\Sigma(I - \mathbf{w}_0\mathbf{w}_0^T)^{\otimes 2}(\mathbf{E}_1 \otimes \mathbf{E}_1)^T. \\ &= \mathbf{M} \otimes \mathbf{M} + (\mathbf{E}_1 \otimes \mathbf{E}_1)\hat{\Sigma}(\mathbf{E}_1 \otimes \mathbf{E}_1)^T.\end{aligned}$$

where $\hat{\Sigma} = \Sigma(I - \mathbf{w}_0\mathbf{w}_0^T)^{\otimes 2}$. Therefore,

$$\begin{aligned}\mathbb{E}\left[\mathbf{F}_t \otimes \mathbf{F}_t\right] &= (\mathbf{E}_1 \otimes \mathbf{E}_1)^T \left(\mathbf{M} \otimes \mathbf{M} + (\mathbf{E}_1 \otimes \mathbf{E}_1)\hat{\Sigma}(\mathbf{E}_1 \otimes \mathbf{E}_1)^T\right)^t (\mathbf{E}_1 \otimes \mathbf{E}_1) \\ &= \sum_{n=0}^{t} \sum_{\mathbf{k}\in S_{t-n}^{n+1}} \prod_{i=1}^{n} (p_{k_i}^{\otimes 2}(\mathbf{A};\beta) \cdot \hat{\Sigma}) \cdot p_{k_{n+1}}^{\otimes 2}(\mathbf{A};\beta),\end{aligned}$$

and

$$\begin{aligned}&\mathbb{E}\left[\mathbf{F}_t\mathbf{w}_0 \otimes \mathbf{F}_t\mathbf{w}_0\right] - \mathbb{E}\left[\mathbf{F}_t\mathbf{w}_0\right] \otimes \mathbb{E}\left[\mathbf{F}_t\mathbf{w}_0\right] \\ =\ & \sum_{n=1}^{t} \sum_{\mathbf{k}\in S_{t-n}^{n+1}} \prod_{i=1}^{n} (p_{k_i}^{\otimes 2}(\mathbf{A};\beta) \cdot \hat{\Sigma}) \cdot p_{k_{n+1}}^{\otimes 2}(\mathbf{A};\beta)\mathbf{w}_0^{\otimes 2}.\end{aligned}$$

Taking the norm, if $0 \preceq X \preceq \lambda_1 I$,

$$\begin{aligned}
&\|\mathbb{E}\left[\mathbf{F}_t\mathbf{w}_0 \otimes \mathbf{F}_t\mathbf{w}_0\right] - \mathbb{E}\left[\mathbf{F}_t\mathbf{w}_0\right] \otimes \mathbb{E}\left[\mathbf{F}_t\mathbf{w}_0\right]\| \\
&\leq \sum_{n=1}^{t} \sum_{\mathbf{k}\in S_{t-n}^{n+1}} \left\|p_{k_i}^{\otimes 2}(\mathbf{A};\beta)\right\| \cdot \left\|\hat{\Sigma}\right\| \cdot \left\|p_{k_2}^{\otimes 2}(\mathbf{A};\beta)\right\| \cdot \left\|\hat{\Sigma}\right\| \cdots \|\Sigma\| \cdot \left\|(I-\mathbf{w}_0\mathbf{w}_0^T)^{\otimes 2} p_{k_{n+1}}^{\otimes 2}(\mathbf{A};\beta)\mathbf{w}_0^{\otimes 2}\right\| \\
&\leq \sum_{n=1}^{t} \|\Sigma\|^n \sum_{\mathbf{k}\in S_{t-n}^{n+1}} \prod_{i=1}^{n} \|p_{k_i}(\mathbf{A};\beta)\|^2 \left\|(I-\mathbf{w}_0\mathbf{w}_0^T)^{\otimes 2} p_{k_{n+1}}^{\otimes 2}(\mathbf{A};\beta)\mathbf{w}_0^{\otimes 2}\right\| \\
&= \sum_{n=1}^{t} \|\Sigma\|^n \sum_{K\in S_{t-n}^{n+1}} \prod_{i=1}^{n} p_{k_i}^2(\lambda_1;\beta) \left\|(I-\mathbf{w}_0\mathbf{w}_0^T) p_{k_{n+1}}(\mathbf{A};\beta) z\right\|^2 \\
&\leq 4\theta \sum_{n=1}^{t} \|\Sigma\|^n \beta^{t-n} \sum_{\mathbf{k}\in S_{t-n}^{n+1}} \prod_{i=1}^{n+1} U_{K_i}^2\left(\frac{\lambda_1}{2\sqrt{\beta}}\right).
\end{aligned}$$

The last inequality follows from

$$\begin{aligned}
\left\|(I-\mathbf{w}_0\mathbf{w}_0^T)p(\mathbf{A};\beta)\mathbf{w}_0\right\|^2 &\leq 2\left\|(I-\mathbf{w}_0\mathbf{w}_0^T)(I-\mathbf{u}_1\mathbf{u}_1^T)p(\mathbf{A};\beta)\mathbf{w}_0\right\|^2 + 2\left\|(I-\mathbf{w}_0\mathbf{w}_0^T)\mathbf{u}_1\mathbf{u}_1^T p(\mathbf{A};\beta)\mathbf{w}_0\right\|^2 \\
&\leq 2\|p(\mathbf{A};\beta)\|^2\left\|(I-\mathbf{u}_1\mathbf{u}_1^T)\mathbf{w}_0\right\|^2 + 2\left\|(I-\mathbf{w}_0\mathbf{w}_0^T)\mathbf{u}_1\right\|^2\|p(\mathbf{A};\beta)\|^2(\mathbf{u}_1^T\mathbf{w}_0)^2 \\
&\leq 4\theta\|p(\mathbf{A};\beta)\|^2,
\end{aligned}$$

where the last inequality uses Lemma 23.

This is what we wanted to show. □

Remark. Comparing to Corollary 15, which is for the non-SVRG setting, Corollary 7 shows the covariance is controlled by the angle bewteen $\mathbf{u}_1$ and $\mathbf{w}_0$ which leads to shrinking variance across epochs.

Corollary 18. *Under the same condition of Lemma 7, we have*

$$\|\mathbb{E}\left[\mathbf{F}_t\mathbf{z} \otimes \mathbf{F}_t\mathbf{z}\right] - \mathbb{E}\left[\mathbf{F}_t\mathbf{z}\right] \otimes \mathbb{E}\left[\mathbf{F}_t\mathbf{z}\right]\| \leq 4\theta \cdot p_t^2(\lambda_1;\beta)\left(\exp\left(\frac{4\|\Sigma\|t}{\lambda_1^2-4\beta}\right)-1\right)$$

Further, if $4\|\Sigma\|t \leq \lambda_1^2 - 4\beta$, *we have*

$$\|\mathbb{E}\left[\mathbf{F}_t\mathbf{z} \otimes \mathbf{F}_t\mathbf{z}\right] - \mathbb{E}\left[\mathbf{F}_t\mathbf{z}\right] \otimes \mathbb{E}\left[\mathbf{F}_t\mathbf{z}\right]\| \leq p_t^2(\lambda_1;\beta)\frac{32\theta\|\Sigma\|t}{\lambda_1^2-4\beta}.$$

Proof. The proof is the same as the one of Corollary 14. □

Lemma 19. *Suppose we run Algorithm 2 for one epoch of length t with initial unit vector $\mathbf{w}_0$. Assume that $\theta = 1-(\mathbf{u}_1^T\mathbf{w}_0)^2 < 1/2$ is small. Under the same condition of Lemma 7, for any $\delta \in (0,1)$, when*

$$\begin{aligned}
t &= \frac{\sqrt{\beta}}{\sqrt{\lambda_1^2-4\beta}}\log\left(\frac{1}{\delta c}\right) \\
\|\Sigma\| &\leq \frac{(\lambda_1^2-4\beta)^{3/2}\delta c}{32\sqrt{d}\sqrt{\beta}}\log^{-1}\left(\frac{1}{\delta c}\right).
\end{aligned}$$

then with probability at least $1-2\delta$, we have

$$1-\frac{(\mathbf{u}_1^T\mathbf{w}_t)^2}{\|\mathbf{w}_t\|^2} \leq \frac{1}{9}\cdot(1-(\mathbf{u}_1^T\mathbf{w}_0)^2).$$

where $c \in (0, 1/16)$ is some small constant.

Proof. First,

$$\begin{aligned}\mathbf{Var}\left[\mathbf{u}_1^T\mathbf{F}_t\mathbf{w}_0\right] &= (\mathbf{u}_1\otimes\mathbf{u}_1)^T(\mathbb{E}\left[\mathbf{F}_t\mathbf{w}_0\otimes\mathbf{F}_t\mathbf{w}_0\right]-\mathbb{E}\left[\mathbf{F}_t\mathbf{w}_0\right]\otimes\mathbb{E}\left[\mathbf{F}_t\mathbf{w}_0\right])\\ &\le \|\mathbf{u}_1\otimes\mathbf{u}_1\|_2\|\mathbb{E}\left[\mathbf{F}_t\mathbf{w}_0\otimes\mathbf{F}_t\mathbf{w}_0\right]-\mathbb{E}\left[\mathbf{F}_t\mathbf{w}_0\right]\otimes\mathbb{E}\left[\mathbf{F}_t\mathbf{w}_0\right]\|_2\\ &\le 4\theta\cdot p_t^2(\lambda_1;\beta)\frac{32\theta\|\Sigma\|t}{\lambda_1^2-4\beta}\end{aligned}$$

by Chebyshev's inequality, for any $\delta>0$, we have

$$\Pr\left(\left|\mathbf{u}_1^T\mathbf{F}_t\mathbf{w}_0-p_t(\lambda_1;\beta)\mathbf{u}_1^T\mathbf{w}_0\right|\ge\frac{1}{\sqrt{\delta}}\cdot p_t\left(\lambda_1;\beta\right)\cdot\sqrt{\frac{32\theta\|\Sigma\|t}{(\lambda_1^2-4\beta)}}\right)\le\delta.$$

That is,

$$\Pr\left(\left|\mathbf{u}_1^T\mathbf{F}_t\mathbf{w}_0\right|\le p_t(\lambda_1;\beta)\left(\left|\mathbf{u}_1^T\mathbf{w}_0\right|-\sqrt{\frac{32\theta\|\Sigma\|t}{(\lambda_1^2-4\beta)\delta}}\right)\right)\le\delta.$$

On the other hand,

$$\begin{aligned}\sum_{i=2}^d\mathbb{E}\left[(\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0)^2\right] &= \sum_{i=2}^d\left(\mathbb{E}\left[(\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0)^2\right]-\mathbb{E}\left[\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0\right]^2+\mathbb{E}\left[\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0\right]^2\right)\\ &= \sum_{i=2}^d\left[\mathbb{E}\left[(\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0)^{\otimes 2}\right]-\mathbb{E}\left[\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0\right]^{\otimes 2}\right]+\sum_{i=2}^d p_t^2(\lambda_i;\beta)(\mathbf{u}_i^T\mathbf{w}_0)^2\\ &= (\sum_{i=2}^d\mathbf{u}_i^{\otimes 2})^T\cdot\left[\mathbb{E}\left[(\mathbf{F}_t\mathbf{w}_0)^{\otimes 2}\right]-\mathbb{E}\left[\mathbf{F}_t\mathbf{w}_0\right]^{\otimes 2}\right]+\sum_{i=2}^d p_t^2(\lambda_i;\beta)(\mathbf{u}_i^T\mathbf{w}_0)^2\\ &\le \sqrt{d}\cdot\left(p_t^2\left(\lambda_1;\beta\right)\cdot\frac{32\theta\|\Sigma\|t}{(\lambda_1^2-4\beta)}\right)+\theta p_t^2(2\sqrt{\beta};\beta)\\ &\le \sqrt{d}\cdot p_t^2\left(\lambda_1;\beta\right)\cdot p_t^2\left(\lambda_1;\beta\right)\cdot\frac{32\theta\|\Sigma\|t}{(\lambda_1^2-4\beta)}+\theta p_t^2(2\sqrt{\beta};\beta)\\ &\le \theta\cdot p_t^2\left(\lambda_1;\beta\right)\left(\frac{32\sqrt{d}\|\Sigma\|t}{\lambda_1^2-4\beta}+\frac{p_t^2(2\sqrt{\beta};\beta)}{p_t^2\left(\lambda_1;\beta\right)}\right).\end{aligned}$$

Therefore, by Markov's inequality,

$$\Pr\left(\sum_{i=2}^n(\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0)^2\ge\theta\cdot p_t^2\left(\lambda_1;\beta\right)\left(\frac{32\sqrt{d}\|\Sigma\|t}{(\lambda_1^2-4\beta)\delta}+\frac{p_t^2(2\sqrt{\beta};\beta)}{\delta p_t^2\left(\lambda_1;\beta\right)}\right)\right)\le\delta.$$

It follows by a union bound that

$$\Pr\left(\frac{\sum_{i=2}^d(\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0)^2}{(\mathbf{u}_1^T\mathbf{F}_t\mathbf{w}_0)^2}\ge\theta\left(\frac{32\sqrt{d}\|\Sigma\|t}{(\lambda_1^2-4\beta)\delta}+\frac{p_t^2(2\sqrt{\beta};\beta)}{\delta p_t^2\left(\lambda_1;\beta\right)}\right)\left(\left|\mathbf{u}_1^T\mathbf{w}_0\right|-\sqrt{\frac{32\theta\|\Sigma\|t}{(\lambda_1^2-4\beta)\delta}}\right)^{-2}\right)\le 2\delta.$$

Since $\left|\mathbf{u}_1^T\mathbf{w}_0\right|^2\ge 1-\theta$, then

$$\Pr\left(\frac{\sum_{i=2}^d(\mathbf{u}_i^T\mathbf{F}_t\mathbf{w}_0)^2}{(\mathbf{u}_1^T\mathbf{F}_t\mathbf{w}_0)^2}\ge\theta\cdot\left(\frac{32\sqrt{d}\|\Sigma\|t}{(\lambda_1^2-4\beta)\delta}+\frac{p_t^2(2\sqrt{\beta};\beta)}{\delta p_t^2\left(\lambda_1;\beta\right)}\right)\left(\sqrt{1-\theta}-\sqrt{\frac{32\theta\|\Sigma\|t}{(\lambda_1^2-4\beta)\delta}}\right)^{-2}\right)\le 2\delta.$$

For any $c\in(0,1/16)$, when

$$t=\frac{\sqrt{\beta}}{\sqrt{\lambda_1^2-4\beta}}\log(\frac{1}{\delta c}),$$

we have

$$\frac{p_t^2(2\sqrt{\beta};\beta)}{\delta p_t^2(\lambda_1;\beta)}\le c.$$

When

$$\|\Sigma\| \leq \frac{(\lambda_1^2 - 4\beta)\delta c}{8\sqrt{d}t} = \frac{(\lambda_1^2 - 4\beta)^{3/2}\delta c}{42\sqrt{d}\sqrt{\beta}} \log^{-1}\left(\frac{1}{\delta c}\right),$$

we have

$$\frac{8\sqrt{d}\|\Sigma\|t}{\delta(\lambda_1^2 - 4\beta)} \leq \epsilon.$$

With both conditions, we have with probability at least $1 - 2\delta$,

$$\begin{aligned}\frac{\sum_{i=2}^d (\mathbf{u}_i^T \mathbf{F}_t \mathbf{w}_0)^2}{(\mathbf{u}_1^T \mathbf{F}_t \mathbf{w}_0)^2} &\leq \theta \cdot \frac{2c}{\left(\sqrt{1-\theta} - \sqrt{\frac{c\theta}{\sqrt{d}}}\right)^2} \\ &\leq \theta \cdot \frac{4c}{(1-\sqrt{c})^2} \\ &\leq \frac{1}{9}\theta.\end{aligned}$$

The last two inequalities follow from the fact that $\theta < 1/2$ and $c < 1/16$. Therefore with the conditions above, we have with probability at least $1 - 2\delta$,

$$1 - \frac{(\mathbf{u}_1^T \mathbf{w}_t)^2}{\|\mathbf{w}_t\|^2} \leq \frac{1}{9} \cdot \theta.$$

□

Theorem 5. *Suppose we run Algorithm 2 with $2\sqrt{\beta} \in [\lambda_2, \lambda_1)$ and a initial unit vector $\mathbf{w}_0$ such that $1 - (\mathbf{u}_1^T \mathbf{w}_0)^2 \leq \frac{1}{2}$. For any $\delta, \epsilon \in (0,1)$, if*

$$T = \frac{\sqrt{\beta}}{\sqrt{\lambda_1^2 - 4\beta}} \log\left(\frac{1}{c\delta}\right), s \geq \frac{32\sqrt{d}\sqrt{\beta}\sigma^2}{c(\lambda_1^2 - 4\beta)\delta} \log\left(\frac{1}{c\delta}\right), \tag{7}$$

then after $K = \mathcal{O}(\log(1/\epsilon))$ epochs, with probability at least $1 - \log\left(\frac{1}{\epsilon}\right)\delta$, we have $1 - (\mathbf{u}_1^T \tilde{\mathbf{w}}_K)^2 \leq \epsilon$, where $c \in (0, 1/16)$ is a numerical constant.

Proof. According to Lemma 19, if we have

$$t = \frac{\sqrt{\beta}}{\sqrt{\lambda_1^2 - 4\beta}} \log\left(\frac{1}{\delta c}\right), \quad \|\Sigma\| \leq \frac{(\lambda_1^2 - 4\beta)^{3/2}\delta c}{32\sqrt{d}\sqrt{\beta}} \log^{-1}\left(\frac{1}{\delta c}\right),$$

then with probability at least $1 - 2\delta$,

$$1 - (\mathbf{u}_1^T \tilde{\mathbf{w}}_{k+1})^2 \leq \frac{1}{9}\left(1 - (\mathbf{u}_1^T \tilde{\mathbf{w}}_k)^2\right).$$

holds. In order to achieve ϵ accuracy, we run $K = \frac{\log(1/\epsilon)}{\log 9} = \mathcal{O}(\log(1/\epsilon))$ epochs and the success probability follows from a union bound which is $1 - 2\frac{\log(1/\epsilon)}{\log 9}\delta \geq 1 - \log(1/\epsilon)\delta$.

Now use the fact that

$$\|\Sigma\| = \left\|\mathbb{E}\left[(\mathbf{A}_t - \mathbf{A})^{\otimes 2}\right]\right\| \leq \mathbb{E}\left[\left\|(\mathbf{A}_t - \mathbf{A})^{\otimes 2}\right\|\right] = \mathbb{E}\left[\|\mathbf{A}_t - \mathbf{A}\|^2\right] = \frac{\sigma^2}{s},$$

and we get a sufficient condition on the batch size s, which is

$$s \geq \frac{32\sqrt{d}\sqrt{\beta}\sigma^2}{c(\lambda_1^2 - 4\beta)\delta} \log\left(\frac{1}{\delta}\right).$$

With that, it completes the proof. □

D Technical Lemmas

This section contains the lemmas or statements that were used for the analysis in the appendix.

Lemma 20. *Given the polynomial sequence $\{p_t(x)\}$ defined in (**P**), when $\beta > 0$, we have*

$$p_t(x) = \begin{cases} \frac{1}{2}\left[\left(\frac{x-\sqrt{x^2-4\beta}}{2}\right)^t + \left(\frac{x+\sqrt{x^2-4\beta}}{2}\right)^t\right], & |x| > 2\sqrt{\beta}, \\ (\sqrt{\beta})^t \cos\left(t \arccos(\frac{x}{2\sqrt{\beta}})\right), & |x| \le 2\sqrt{\beta}. \end{cases}$$

Proof. Consider the generating function of $\{p_t(x)\}$, $G(x,z) = \sum_{t=0}^{\infty} p_t(x) z^t, z \in \mathbb{C}$. And

$$\sum_{t=1}^{\infty} p_{t+1} z^{t+1} = \sum_{t=1}^{\infty} x p_t z^{t+1} - \beta \sum_{t=1}^{\infty} p_{t-1} z^{t+1} \tag{20}$$

$$G(x,z) - p_0 - p_1 z = xz(G(x,z) - p_0) - \beta z^2 G(x,z) \tag{21}$$

$$(\beta z^2 - xz + 1) G(x,z) = p_0 + (p_1 - p_0 x) z \tag{22}$$

Since $p_0 = 1, p_1 = x/2$, we have

$$G(x,z) = \frac{1 - xz/2}{\beta z^2 - xz + 1} = \frac{1 - xz/2}{\beta(z - r_1)(z - r_2)},$$

where $r_1, r_2 \in \mathbb{C}$ are two roots of $\beta z^2 - xz + 1$. When $r_1 \neq r_2$, we have

$$\begin{aligned} G(x,z) &= \frac{1 - xz/2}{\beta(r_1 - r_2)} \left(\frac{1}{r_2 - z} - \frac{1}{r_1 - z}\right) \\ &= \frac{1 - xz/2}{\beta(r_1 - r_2)} \sum_{t=0}^{\infty} \left[\left(\frac{1}{r_2}\right)^{t+1} - \left(\frac{1}{r_1}\right)^{t+1}\right] z^t \\ &= \sum_{t=0}^{\infty} \left[\frac{1/r_2 - x/2}{\beta(r_1 - r_2)} \left(\frac{1}{r_2}\right)^t - \frac{1/r_1 - x/2}{\beta(r_1 - r_2)} \left(\frac{1}{r_1}\right)^t\right] z^t. \end{aligned}$$

When $|x| > 2\sqrt{\beta}$, $r_{1,2} = \frac{x \pm \sqrt{x^2 - 4\beta}}{2\beta}$. Therefore,

$$G(x,z) = \frac{1 - xz/2}{\beta(z - r_1)(z - r_2)} = \sum_{t=0}^{\infty} \frac{1}{2} \left[\left(\frac{1}{r_2}\right)^t + \left(\frac{1}{r_1}\right)^t\right] z^t.$$

By complex analysis theory, when $|z| < r_2$, $G(x,z)$ is well-defined. By comparing the coefficient of z^t, we get

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

When $|x| < 2\sqrt{\beta}$, $r_{1,2} = \frac{x \pm \mathbf{i}\sqrt{4\beta - x^2}}{2\beta}$. Then

$$G(x,z) = \sum_{t=0}^{\infty} \frac{1}{2} \left[\left(\frac{1}{r_2}\right)^t + \left(\frac{1}{r_1}\right)^t\right] z^t.$$

Then $z < \frac{1}{\sqrt{\beta}} = |r_{1,2}|$, $G(x,z)$ is well-defined. Then

$$p_t(x) = \frac{1}{2} \left[\left(\frac{x - \mathbf{i}\sqrt{4\beta - x^2}}{2}\right)^t + \left(\frac{x + \mathbf{i}\sqrt{4\beta - x^2}}{2}\right)^t\right].$$

When $|x| = 2\sqrt{\beta}$, $r_{1,2} = \frac{x}{2\beta}$. Suppose $x = 2\sqrt{\beta}$, then

$$G(x,z) = \frac{1}{1-\sqrt{\beta}z} = \sum_{t=0}^{\infty}(\sqrt{\beta})^t z^t.$$

When $|z| \le 1/\sqrt{\beta}$, then $G(x,z)$ is well-defined. Then

$$p_t(2\sqrt{\beta}) = (\sqrt{\beta})^t.$$

Similarly, if $x = -2\sqrt{\beta}$, we have $p_t(-2\sqrt{\beta}) = (-\sqrt{\beta})^t$.

Combine all three cases, we have

$$p_t(x) = \begin{cases} \frac{1}{2}\left[\left(\frac{x-\sqrt{x^2-4\beta}}{2}\right)^t + \left(\frac{x+\sqrt{x^2-4\beta}}{2}\right)^t\right], & |x| > 2\sqrt{\beta}, \\ \frac{1}{2}\left[\left(\frac{x-\mathbf{i}\sqrt{4\beta-x^2}}{2}\right)^t + \left(\frac{x+\mathbf{i}\sqrt{4\beta-x^2}}{2}\right)^t\right], & |x| \le 2\sqrt{\beta}. \end{cases}$$

□

Fact 21 (Binomial Expansion of Matrices)**.** *For any matrix $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$, the binomial expansion $(\mathbf{A}+\mathbf{B})^t$ has the following form,*

$$(\mathbf{A}+\mathbf{B})^t = \sum_{j=0}^{t}\sum_{\mathbf{k}\in S_{t-j}^{j+1}} \mathbf{A}^{k_1}\prod_{i=2}^{j+1}\mathbf{B}\mathbf{A}^{k_i},$$

S_m^n denotes the set of vectors in $\mathbb{N}^n$ with entries that sum to m, i.e.

$$S_m^n = \{\mathbf{k} = (k_1, \cdots, k_n) \in \mathbb{N}^n | \sum_{i=1}^{n} k_i = m\}.$$

Lemma 22. *Given $k_1, \cdots, k_n, k_{n+1} \in \mathbb{N}$, and $z \ge 1$, then we have*

$$\prod_{i=1}^{n+1} U_{k_i}^2(z) \le U^2_{\sum_{i=1}^{n+1} k_i + n}(z) \cdot \frac{1}{(z^2-1)^n}$$

Proof. We prove it by induction. First base case when $n = 0$, this is trivial. Assume $n = t$ this inequality holds, and consider $n = t+1$,

$$\begin{aligned}
\prod_{i=1}^{t+2} U_{k_i}^2(z) &\le U^2_{k_{t+2}}(z) \cdot U^2_{\sum_{i=1}^{t+1}(k_i+1)}(z) \cdot \frac{1}{(z^2-1)^t} \\
&= (T^2_{k_{t+2}+1}(z) - 1) \cdot U^2_{\sum_{i=1}^{t+1} k_i + t}(z) \cdot \frac{1}{(z^2-1)^{t+1}} \\
&\le T^2_{k_{t+2}+1}(z) \cdot U^2_{\sum_{i=1}^{t+1} k_i + t}(z) \cdot \frac{1}{(z^2-1)^{t+1}} \\
&\le U^2_{\sum_{i=1}^{t+2} k_i + t + 1}(z) \cdot \frac{1}{(z^2-1)^{t+1}},
\end{aligned}$$

which completes the induction. □

Lemma 23. *Given any two unit vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$, then we have*

$$1 - (\mathbf{u}^T\mathbf{v})^2 = \left\|(I - \mathbf{u}\mathbf{u}^T)\mathbf{v}\right\|^2 = \left\|(I - \mathbf{v}\mathbf{v}^T)\mathbf{u}\right\|^2.$$

Proof.

$$\begin{aligned}
\left\|(I - \mathbf{u}\mathbf{u}^T)\mathbf{v}\right\|^2 &= \|\mathbf{v}\|^2 - 2(\mathbf{v}^T\mathbf{u})^2 + (\mathbf{u}^T\mathbf{v})^2 \\
&= 1 - (\mathbf{v}^T\mathbf{u})^2.
\end{aligned}$$

Similarly, we have $\left\|(I - \mathbf{v}\mathbf{v}^T)\mathbf{u}\right\|^2 = 1 - (\mathbf{v}^T\mathbf{u})^2$. □

E Data Generation for Figure 1

The synthetic dataset $\mathbf{X} \in \mathbb{R}^{10^6 \times 10}$ was just generated through its singular value decomposition. Specifically we fix a 10 by 10 diagonal matrix $\Sigma = \text{diag}\{1, \sqrt{0.9}, \cdots, \sqrt{0.9}\}$ and generate random orthogonal projection matrix $\mathbf{U} \in \mathbb{R}^{10^6 \times 10}$ and random orthogonal matrix $\mathbf{V} \in \mathbb{R}^{10 \times 10}$. And the dataset $X = \sqrt{\mathbf{n}}\mathbf{U}\Sigma\mathbf{V}^T$ which guarantees that the matrix $\mathbf{A} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$ has eigen-gap 0.1.