

ODE-Inspired Analysis for the Biological Version of Oja’s Rule in Solving Streaming PCA

Chi-Ning Chou ^{*†}

Mien Brabeeba Wang ^{*‡}

November 7, 2019

Abstract

Oja’s rule [Oja, Journal of mathematical biology 1982] is a well-known biologically-plausible algorithm using a Hebbian-type synaptic update rule to solve streaming principal component analysis (PCA). Computational neuroscientists have known that this biological version of Oja’s rule converges to the top eigenvector of the covariance matrix of the input in the limit. However, prior to this work, it was open to prove any convergence rate guarantee.

In this work, we give the first convergence rate analysis for the biological version of Oja’s rule in solving streaming PCA. Moreover, our convergence rate matches the information theoretical lower bound up to logarithmic factors and outperforms the state-of-the-art upper bound for streaming PCA. Furthermore, we develop a novel framework inspired by ordinary differential equations (ODE) to analyze general stochastic dynamics. The framework abandons the traditional *step-by-step* analysis and instead analyzes a stochastic dynamic in *one-shot* by giving a closed-form solution to the entire dynamic. The one-shot framework allows us to apply stopping time and martingale techniques to have a flexible and precise control on the dynamic. We believe that this general framework is powerful and should lead to effective yet simple analysis for a large class of problems with stochastic dynamics.

1 Introduction

Human brains process an astronomical amount of visual data constantly. In our eyes, 100 millions photoreceptors in the retina receive gigabytes of information per second [A95, WBD77]. To reduce the curse of dimensionality, as a first step, the brain compresses the activities of 100 million photoreceptors into one million retina ganglion cells in optical nerves [GS12]. Neuroscientists have extensively studied the dimensionality reduction in the retina-optical nerve pathway. In particular, Haft and Hemmen [HV98] demonstrated that Principal Component Analysis (PCA) is a likely candidate for the dimensionality reduction in the retina-optical nerve pathway by showing the consistency between theoretical predictions and experiments in the receptive fields of retina ganglion cells. However, their work only proposed PCA as an potential solution to the pathway and did not provide a dynamic to explain the learning process of PCA.

On the other hand, in the seminal work of Oja [Oja82], he proposed a mathematical model for the biological neural network that solves streaming PCA with several biologically-plausible properties: the network not only updates its synaptic weights locally but also normalizes the strength of synapses. This rule, now known as the biological version of Oja’s rule (*biological Oja’s rule*¹ in abbreviation), has been the subject of extensive theoretical [Oja82, OK85, San89, HKP91, Oja92, Plu95, DK96, Zuf02, YYLT05, Duf13, ACS13] and experimental [CL94, KDT94, HP94, Kar96, CKS96, SLY06, SA06, LTYH09, ACS13] studies aimed at understanding its performance. Despite its popularity, the theoretical understanding of the biological Oja’s

^{*}The two authors contribute equally to this paper.

[†]School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, USA. Supported by NSF awards CCF 1565264 and CNS 1618026. Email: chiningchou@eng.harvard.edu.

[‡]CSAIL, MIT, Cambridge, Massachusetts, USA. Email: brabeeba@mit.edu.

¹Also known as *Oja’s rule* in the literature. However, many works in the machine learning community use the name “Oja’s rule” for *non-biologically-plausible* variants of the original Oja’s rule. Thus, in this paper we emphasize the term “*biological*” to distinguish the two. See Section 1.4 for more discussions on their differences.

rule cannot account for the biologically-realistic time scale in the retina-optical nerve pathway because the state-of-the-art theoretical analysis only provides a guarantee on convergence in the limit [Duf13].

In practice, the time scale of the streaming PCA in the retina-optical nerve pathway is on the order of seconds. For example, when a person is walking from a dark room to a bright room, it only takes a few seconds for the eyes to adapt to the new environment. This suggests that a plausible dynamic for explaining the retina-optical nerve pathway should have little or no dependency on the dimension, *i.e.*, the number of neurons, which in this case is on the order of 100 million. Meanwhile, researchers have observed that the biological Oja’s rule (and its variants) has fast convergence rates [HP94, Kar96, SLY06, SA06, LTYH09] in simulations. Thus, to further our understanding in the retina-optical nerve pathway, it is important to give a theoretical analysis to show that the biological Oja’s rule solves streaming PCA in a biologically-realistic time scale. This is nevertheless a challenging task and has remained elusive for almost 40 years [Oja82].

In this work, we provide the first convergence rate analysis for the biological Oja’s rule in solving streaming PCA.

Theorem 1.1 (informal). *The biological Oja’s rule efficiently solves streaming PCA with (nearly) optimal convergence rate. Specifically, the convergence rate we obtain matches the information theoretical lower bound up to logarithmic factors.*

Furthermore, the convergence rate has no dependency on the dimension when the initial weight vector is close to the top eigenvector or has a logarithmic dependency on the dimension when the initial vector is random. Therefore, the biological Oja’s rule solves streaming PCA in a biologically-realistic time scale.

To show the (nearly) optimal convergence rate of biological Oja’s rule in solving streaming PCA, we develop an ODE-inspired framework to analyze stochastic dynamics. Concretely, instead of the traditional *step-by-step* analysis, our framework analyzes a dynamical system in *one-shot* by giving a closed-form solution for the entire dynamic. The framework borrows ideas from ordinary differential equations (ODE) and stochastic differential equations (SDE) to obtain a closed-form characterization of the dynamic and uses stopping time and martingale techniques to precisely control the dynamic. This framework provides a more elegant and more general analysis compared with the previous step-by-step approaches. We believe that this novel framework can provide simple and effective analysis on other problems with stochastic dynamics.

We organize the rest of the introduction as follows. We first formally define biological Oja’s rule and streaming PCA in Section 1.1 and state the main results and their biological relevance in Section 1.2. In Section 1.3, we provide a technical overview on the proof and the analysis framework. Finally, we conclude the introduction with a survey and comparison of related works in Section 1.4.

1.1 Biological Oja’s rule and streaming PCA

In a biological neural network, two neurons primarily interact with each other via action potentials or instantaneous signals, *a.k.a.*, “spikes”, through *synapses* between them. The strength of a synapse might vary from time to time and is called the *synaptic weight*. The ability of a synaptic weight to strengthen or weaken over time is considered as a source for learning and long term memory in our brains. While generally the update of a synaptic weight could depend on the *spiking patterns* of the end neurons, it is common for neuroscientists to focus on the averaging behaviors of a spiking dynamic. Namely, they simplify the model by only considering the *firing rate*, which is defined as the average number of spikes. This is known as the *rate-based model* [WC72, WC73] and since the biological Oja’s rule was defined on a rate-based model, this setting is going to be the focus of this work.

To understand how the biological Oja’s rule works, consider the following baby example with two neurons x and y . Let $x_t, y_t \in \mathbb{R}$ be the firing rates of neurons x, y at time $t \in \mathbb{N}$ and let $w_t \in \mathbb{R}$ be the synaptic weight from x to y at time t . In a biological neural network, w_t could change over time and the dynamic is defined *locally* on the previous synaptic weight as well as the firing rates of the end neurons. Namely, the synaptic weight from the neuron x to y has the following dynamic

$$w_t = w_{t-1} + \eta_t F_t(w_{t-1}, x_t, y_t)$$

where F_t is an update function and η_t is the *plasticity coefficient*, a.k.a., the *learning rate*. Biologically, the update function should further follow the Hebb postulate, “cells that fire together wire together” [Heb49]. One naive way to implement Hebbian learning is to set the update function as $F_t(w_{t-1}, x_t, y_t) = x_t y_t$. However, the values of w_t can grow unboundedly. The biological Oja’s rule is a self-normalizing Hebbian rule with the following synaptic updates.

$$w_t = w_{t-1} + \eta_t y_t (x_t - y_t w_{t-1}) .$$

Using the above synaptic update rule, Oja [Oja82] configured a network that solves streaming PCA while keeping the norm of the weights stable. Before introducing the network, let us formally define the streaming PCA problem.

Streaming PCA Principal component analysis (PCA) [Pea01, Hot33] is a problem to find the top eigenvector of a covariance matrix of a dataset. Let n be the dimension of the data. In the offline setting, one can compute the covariance matrix in $O(n^2)$ space and use the power method to approximate the top eigenvector. As for its variant, the streaming PCA (a.k.a. the stochastic online PCA, see [CG90] for a survey on the literature), the input data arrives in a stream and the algorithm/dynamic only has limited amount of space, e.g., $O(n)$ space. Streaming PCA is important for biological system because the information inherently arrives in a stream in a living system. On the other hand, it is also much more challenging than offline PCA (see for example [AZL17]). In the following, we formally define the streaming PCA problem.²

Problem 1.2 (Streaming PCA). *Let $n, T \in \mathbb{N}$ and $\mathcal{D}$ be a distribution over the unit sphere of $\mathbb{R}^n$. Suppose the input data $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T \stackrel{i.i.d.}{\sim} \mathcal{D}$ are given one by one in a stream. Let $A = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[\mathbf{x}\mathbf{x}^\top]$ be the covariance matrix and $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$ be the eigenvalues of A . Assume $\lambda_1 > \lambda_2$ and let $\mathbf{v}_1$ be the top eigenvector of A of unit length. Then the goal of the streaming PCA problem is to output $\mathbf{w} \in \mathbb{R}^n$ such that $\frac{\langle \mathbf{w}, \mathbf{v}_1 \rangle^2}{\|\mathbf{w}\|_2^2} \geq 1 - \epsilon$.*

Since the inputs arrive in a stream, usually a streaming PCA algorithm/dynamic would maintain a solution $\mathbf{w}_t \in \mathbb{R}^n$ at each time $t \in \mathbb{N}$. Thus, the goal for a streaming PCA algorithm/dynamic would be achieving $\Pr \left[\frac{\langle \mathbf{w}_T, \mathbf{v}_1 \rangle^2}{\|\mathbf{w}_T\|_2^2} \geq 1 - \epsilon \right] \geq 1 - \delta$ with small T .

Biological Oja’s rule in solving streaming PCA Oja [Oja82] proposed a streaming PCA algorithm using n input neurons and one output neuron. The firing rates of the input neurons at time t are denoted by a vector $\mathbf{x}_t \in \mathbb{R}^n$ and the firing rate of the output neuron is denoted by a scalar $y_t \in \mathbb{R}$. The synaptic weights at time t from the input neurons to the output neuron are denoted by a vector $\mathbf{w}_t \in \mathbb{R}^n$. Note that the weight vector will be the output and ideally it will converge to the top eigenvector $\mathbf{v}_1$.

The input stream $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T$ arrives in the form of firing rates of the input neurons. The firing rate of the output neuron is simply the inner product of the synaptic weight vector and the firing rate vector of the input neurons, i.e., $y_t = \mathbf{x}_t^\top \mathbf{w}_{t-1}$. Now, from the biological Oja’s rule, the dynamic of the synaptic weight vector is described by the following equation.

Definition 1.3 (Biological Oja’s rule). *For any initial vector $\mathbf{w}_0 \in \mathbb{R}^n$ such that $\|\mathbf{w}_0\|_2 = 1$, the dynamic of the biological Oja’s rule is the following. For any $t \in \mathbb{N}$, define*

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \eta_t y_t (\mathbf{x}_t - y_t \mathbf{w}_{t-1}) \tag{1.4}$$

where $y_t = \mathbf{x}_t^\top \mathbf{w}_{t-1}$ and $\mathbf{x}_t$ is the input at time t . See also in Figure 1 for a pictorial definition of biological Oja’s rule in solving streaming PCA.

Follow from the definition, the biological Oja’s rule is automatically *biologically-plausible* in the following sense. First, the synaptic update rule is *local*. Namely, each synapse only depends on the previous synaptic

²In related works, some (e.g., [AZL17]) measure the error using $1 - \langle \mathbf{w}, \mathbf{v}_1 \rangle^2$, some (e.g., [Sha16]) use $1 - \mathbf{w}^\top A \mathbf{w} / \|A\|$, and some (e.g., [JJK+16]) use $\sin^2(\mathbf{w}, \mathbf{v}_1)$. We remark that all of these error measures (including ours) are the same up to a constant multiplicative factor.

Also, some works emphasize other convergence notions such as the gap-free convergence [Sha16]. Though we do not explicitly study the convergence of biological Oja’s rule under these notions, we believe that our results could be easily extended to other convergence notions with comparable convergence rate and leave this for future work.

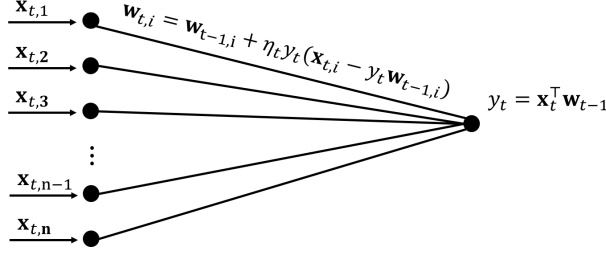


Figure 1: A neural network that uses biological Oja’s rule to solve streaming PCA. The firing rate vector $\mathbf{x}_t$ is the input and the weight vector $\mathbf{w}_t$ is the output at time t .

weight and the firing rates of the two end neurons. Second, with some simple calculations (*e.g.*, [Lemma 5.1](#)), biological Oja’s rule achieves the *synaptic scaling guarantee* [\[AN00\]](#), *i.e.*, $\mathbf{w}_{t,i}$ being bounded for all $t \in \mathbb{N}$ and $i \in [n]$. Thus, one can then interpret the convergence results of this work as showing further biological-plausibilities of the biological Oja’s rule in the retina-optical nerve pathway. See [Section 1.2](#) for more discussions.

Oja’s derivation for the biological Oja’s rule Before going into more technical contents, it would be helpful to take a look at the original derivation for the biological Oja’s rule. Initially, Oja wanted to use the following update rule with normalization³ to solve the streaming PCA problem.

$$\mathbf{w}_t = \frac{(I + \eta_t \mathbf{x}_t \mathbf{x}_t^\top) \mathbf{w}_{t-1}}{\|(I + \eta_t \mathbf{x}_t \mathbf{x}_t^\top) \mathbf{w}_{t-1}\|_2}. \quad (1.5)$$

However, the normalization term $\|(I + \eta_t \mathbf{x}_t \mathbf{x}_t^\top) \mathbf{w}_{t-1}\|_2^{-1}$ is *global*⁴ and does not seem to have a biologically-plausible implementation. To bypass this issue, Oja applied *Taylor’s expansion* on the normalization term and truncated the second order terms of η_t . This exactly results in the biological Oja’s rule (*i.e.*, [Equation 1.4](#)). See [Appendix A](#) for more details on the derivation.

Also, to see why intuitively biological Oja’s rule could solve streaming PCA, one can check that any eigenvector $\mathbf{v}$ of A of unit length with eigenvalue λ is a fixed point of the biological Oja’s rule in expectation. Specifically, the expectation of the update term $y_t(\mathbf{x}_t - y_t \mathbf{w}_{t-1})$ with $\mathbf{w}_{t-1} = \mathbf{v}$ is the following.

$$\mathbb{E} [\mathbf{x}_t^\top \mathbf{v} \mathbf{x}_t - (\mathbf{x}_t^\top \mathbf{v})^2 \mathbf{v}] = A\mathbf{v} - \mathbf{v}^\top A\mathbf{v} \mathbf{v} = \lambda \mathbf{v} - \lambda \|\mathbf{v}\|_2^2 \mathbf{v} = 0.$$

The first equality follows from for all $i, j \in [n]$, $\mathbb{E}[\mathbf{x}_{t,i} \mathbf{x}_{t,j}] = \lambda_i \cdot \mathbf{1}_{i=j}$, and the second equality follows from $A\mathbf{v} = \lambda \mathbf{v}$. By checking the Hessian at the top eigenvector $\mathbf{v}_1$, one can even see that $\mathbf{v}_1$ is a *stable* fixed point.

Previous works: Convergence in the limit results There were many previous works on analyzing the convergence of biological Oja’s rule in solving streaming PCA [\[Oja82, OK85, San89, HKP91, Oja92, Plu95, DK96, Zuf02, YYLT05, Duf13\]](#). However, their works only proved guarantee on convergence in the limit. For example, Duflo [\[Duf13\]](#) showed that $\mathbf{w}_t$ converges to the top eigenvector of A in the limit under some constraints on the learning rates.

Theorem 1.6 ([\[Duf13\]](#), informal). *Let $\mathbf{w}_0$ be a random unit vector in $\mathbb{R}^n$. If $\eta_t \leq \frac{1}{2}$ for all $t \in \mathbb{N}$, $\sum_{t=0}^{\infty} \eta_t = \infty$, and $\sum_{t=0}^{\infty} \eta_t^2 < \infty$, then $\lim_{t \rightarrow \infty} \langle \mathbf{w}_t, \mathbf{v}_1 \rangle^2 = 1$ almost surely.*

The proofs of these previous analyses are usually based on tools from dynamical system such as the Kushner-Clark method or Lyapunov theory. Note that these proof techniques are not sufficient for providing convergence rate guarantee.

³This update rule is doing a variant of power method with normalization. It is widely used in the machine learning community to solve streaming PCA. See [Section 1.4](#) for more discussion.

⁴It is global because computing the ℓ_2 norm requires the information from *every* neurons.

Prior to this work, there had been no efficiency guarantee for the biological Oja’s rule. The main technical barrier is due to the non-linear terms in the update rule which introduces correlations in the traditional step-by-step analysis and thus naive analysis would not work. We explain the difficulty further in [Section 1.3](#) and [Appendix C](#). Given this situation, natural questions on the frontier would then be:

Question: What is the convergence rate of biological Oja’s rule in solving streaming PCA? Is the convergence rate biologically-realistic?

1.2 Our results

In this paper, we answer the above questions by giving the first convergence rate guarantee for the biological Oja’s rule in solving streaming PCA. Furthermore, the convergence rate matches the information-theoretic lower bound for streaming PCA up to logarithmic factors. In terms of the techniques, we develop an ODE-inspired framework to analyze stochastic dynamics. We believe this general framework of using tools and insights from ODE and SDE in analyzing stochastic dynamics is elegant and powerful. We provide more details and intuitions on the ODE-inspired framework in the section on the technical overview (see [Section 1.3](#)). Also, as a byproduct, our convergence rate guarantee for biological Oja’s rule outperforms the state-of-the-art upper bound for streaming PCA (using other variants of Oja’s rule).

There are two common convergence notions in the streaming PCA literature. The *global convergence* requires the algorithm/dynamic to start from a random initial vector while the *local convergence* allows the algorithm/dynamic to start from an initial vector that is highly correlated to the top eigenvector of the covariance matrix. Now, we are ready to state our main theorem as follows.

Theorem 1.7 (Global and local convergence). *With the setting in [Problem 1.2](#) and dynamic in [Definition 1.3](#), let $\text{gap} := \lambda_1 - \lambda_2 > 0$. For any $\epsilon, \delta \in (0, 1)$, we have the following results.*

- (Global convergence) Suppose $\mathbf{w}_0$ is uniformly sampled from the unit sphere in $\mathbb{R}^n$. For any $T \in \mathbb{N}$ and

$$T_0 = \Theta \left(\frac{\lambda_1}{(\epsilon \wedge \delta^2) \cdot \text{gap}^2} \cdot \log^3 \left(T, n, \frac{1}{\text{gap}}, \frac{1}{\epsilon}, \frac{1}{\delta} \right) \right),$$

there exists $\eta = \tilde{O} \left(\frac{(\epsilon \wedge \delta^2) \cdot \text{gap}}{\lambda_1} \right)$ such that if for all $t \in \mathbb{N}$, we let $\eta_t = \eta$, then we have

$$\Pr \left[\forall t \in [T_0, T_0 + T], \frac{\langle \mathbf{w}_t, \mathbf{v}_1 \rangle^2}{\|\mathbf{w}_t\|_2^2} \geq 1 - \epsilon \right] \geq 1 - \delta.$$

- (Local convergence) Suppose $\langle \mathbf{w}_0, \mathbf{v}_1 \rangle^2 = \Omega(1)$. For any $T \in \mathbb{N}$ and

$$T_0 = \Theta \left(\frac{\lambda_1}{\epsilon \cdot \text{gap}^2} \cdot \log^2 \left(T, \frac{1}{\epsilon}, \frac{1}{\delta} \right) \right),$$

there exists $\eta = \tilde{O} \left(\frac{\epsilon \cdot \text{gap}}{\lambda_1 \cdot \log T} \right)$ such that if for all $t \in \mathbb{N}$, we let $\eta_t = \eta$, then we have

$$\Pr \left[\forall t \in [T_0, T_0 + T], \frac{\langle \mathbf{w}_t, \mathbf{v}_1 \rangle^2}{\|\mathbf{w}_t\|_2^2} \geq 1 - \epsilon \right] \geq 1 - \delta.$$

The notation $a \wedge b$ stands for $\min\{a, b\}$ and $\tilde{O}$ hides poly-logarithmic factors in $n, \text{gap}^{-1}, \epsilon^{-1}$, and δ^{-1} .

Biological perspectives Our results provide further theoretical evidences for the biological plausibility of biological Oja’s rule to be a likely candidate of the dimensionality reduction in the retina-optical nerve pathway. Specifically, we show that “biological Oja’s rule is a local Hebbian learning rule with bounded synaptic weights that functions in a biologically-realistic time scale.” In particular, in this work we demonstrate that biological Oja’s rule does not have any dependency on the dimension (*i.e.*, n , the number of neurons) in the local convergence setting while the dependency is logarithmic in the global convergence setting. Moreover, in the local convergence setting, the dependency of the convergence rate on the failure probability δ is inverse-logarithmic in stead of $O(1/\delta)$.

Furthermore, we prove the *for-all-time* guarantee of the biological Oja’s rule as an corollary of the techniques used in the proof for the main theorems. By for-all-time guarantee we refer to the behavior of a dynamic that *always* stays around the optimal solution after convergence. Especially, the dynamic would not temporarily leave the neighborhood of the optimal solution. The for-all-time guarantee is of biological importance because a biological system constantly adapts and functions, and it is not enough for a mechanism to hold for only a brief moment. We state the theorem for the for-all-time guarantee as follows.

Theorem 1.8 (For-all-time guarantee with slowly diminishing rate). *With the setting in Problem 1.2 and dynamic in Definition 1.3, let $\text{gap} := \lambda_1 - \lambda_2 > 0$. For any $\epsilon, \delta \in (0, 1)$, suppose $\frac{\langle \mathbf{w}_0, \mathbf{v}_1 \rangle^2}{\|\mathbf{w}_0\|_2^2} \geq 1 - \epsilon/2$. For any $t \in \mathbb{N}$, let $\eta_t = O\left(\frac{\epsilon \cdot \text{gap}}{\lambda_1 \log(t/\delta)}\right)$, then*

$$\Pr \left[\forall t \in \mathbb{N}, \frac{\langle \mathbf{w}_t, \mathbf{v}_1 \rangle^2}{\|\mathbf{w}_t\|_2^2} \geq 1 - \epsilon \right] \geq 1 - \delta.$$

We should further notice that the learning rate is slowly-diminishing, *i.e.*, $\eta_t = \Theta(1/\log t)$ instead of the commonly used $\eta_t = O(1/t)$, in the for-all-time guarantee (*i.e.*, Theorem 1.8). This suggests the capability of *continual adaptation*, which is crucial in the biological scenario. For example, if a person walks into a new environment, the retina cells need to quickly adapt to the new environment and this cannot be achieved if the learning rate already diminished too fast in the previous environment.

We remark that prior to this work, the for-all-time guarantee with slowly diminishing learning rates was even unknown to any streaming PCA algorithms. The convergence in the limit result for biological Oja’s rule requires $\eta_t = o(1/\sqrt{t})$ [Duf13] and the convergence rate analysis for non-biologically-plausible variants of Oja’s rule requires $\eta_t = \tilde{O}(1/t)$ [JJK⁺16, AZL17, LWLZ18] or $\eta_t = O(1/\sqrt{t})$ [Sha16]. In particular, all previous works satisfy $\sum_t \eta_t^2 < \infty$ while in this work we can achieve for-all-time convergence with much weaker assumptions $\eta_t = \Theta(1/\log t)$ (hence $\sum_t \eta_t^2 = \infty$) for the biological Oja’s rule.

1.3 Technical overview

In this work, we give the first efficiency guarantee for the biological Oja’s rule in solving streaming PCA with an (nearly) optimal convergence rate. In this subsection, we highlight three technical insights of our analysis which lead us to a clean understanding in how the biological Oja’s rule solves streaming PCA. In short, our high-level strategy is to first consider the *continuous* version of the Oja’s rule where the learning rate η is set to be infinitesimal. In the continuous setting, the dynamic can be fully understood by tools from the theory of ordinary differential equations (ODE) or stochastic differential equations (SDE). With the inspiration from the continuous analysis, we are able to identify the right tools (*e.g.*, linearization at two different centers, etc.) to tackle the discrete dynamic.

Before we start, let us recall the problem setting and the goal. For simplicity, here we consider the *diagonal case* where the covariance matrix A is a diagonal matrix, *i.e.*, $A = \text{diag}(\lambda_1 \dots, \lambda_n)$ with $\lambda_1 > \lambda_2 \geq \lambda_3 \geq \dots \geq \lambda_n \geq 0$. Thus the top eigenvector of A is $\mathbf{e}_1$, *i.e.*, the indicator vector for the first coordinate, and the goal becomes showing that $\mathbf{w}_{t,1}^2$ efficiently converges to 1 when $t \rightarrow \infty$. A reduction from the general case to the diagonal case is provided in Section 5.1.

Insight 1: Inspiration from the continuous dynamics The first insight is to analyze the biological Oja’s rule in a way inspired by its continuous analog. The advantage to consider the continuous dynamics is that not only it captures the inherent dynamics but also we can apply the theory of ODE and SDE to obtain *closed-form* solutions. Thus, the continuous dynamic would serve as a hint on how to derive a tight and closed-form analysis for the discrete dynamic.

Interestingly, the continuous SDE of the biological Oja’s rule degenerates into a simple deterministic ODE almost surely (see Section 3 for a derivation). Specifically, for any $t \geq 0$, we have

$$\frac{d\mathbf{w}_{t,1}}{dt} \geq (\lambda_1 - \lambda_2)\mathbf{w}_{t,1}(1 - \mathbf{w}_{t,1}^2) \quad \text{and} \quad \|\mathbf{w}_t\|_2 = 1 \quad (1.9)$$

almost surely. Furthermore, observe that the continuous Oja’s rule is non-decreasing and has three fixed points 0 and ± 1 for $\mathbf{w}_{t,1}$ while the first is unstable and the later two are stable. Namely, in the continuous dynamic, $\mathbf{w}_t$ will eventually converge to $\pm \mathbf{e}_1$, *i.e.*, the top eigenvector of A .

Note that in a discrete stochastic dynamic, there are two sources of the noises: (i) the intrinsic stochasticity from its continuous analog and (ii) the noise due to discretization. Thus, Equation 1.9 suggests that the noise in the biological Oja's rule only comes from discretization since the continuous Oja's rule is deterministic.

In addition to the limiting behavior, one can also read out finer structures of the continuous dynamic from Equation 1.9 by solving the differential equation using standard tools from dynamical system. The right hand side (RHS) of the inequality in Equation 1.9 is non-linear which usually does not have a clean solution. A natural idea from dynamical system would then be *linearizing* the differential equation around fixed points and applying the *exact* solution for a linear ordinary differential equation. Moreover, as there are three fixed points in Equation 1.9, one can linearize the differential equation with center being either 0 or ± 1 . For simplicity, we focus on the two fixed points 0 and 1 while -1 can be analyzed similarly due to symmetry.

For example, we can linearize at 0 by lower bounding the RHS of Equation 1.9 by $\epsilon(\lambda_1 - \lambda_2)\mathbf{w}_{t,1}$ for any $\mathbf{w}_{t,1} \in [0, \sqrt{1-\epsilon}]$ (see Figure 2a). Similarly, we can linearize at 1 by using $\mathbf{w}_{0,1}(\lambda_1 - \lambda_2)(1 - \mathbf{w}_{t,1})$ for any $\mathbf{w}_{t,1} \in [\mathbf{w}_{0,1}, 1]$ (see Figure 2b). Another choice would be *linearizing at both 0 and 1*. Concretely, we linearize at 0 for $\mathbf{w}_{t,1} \in [0, 2/3]$ and linearize at 1 for $\mathbf{w}_{t,1} \in [2/3, 1]$ (see Figure 2c).

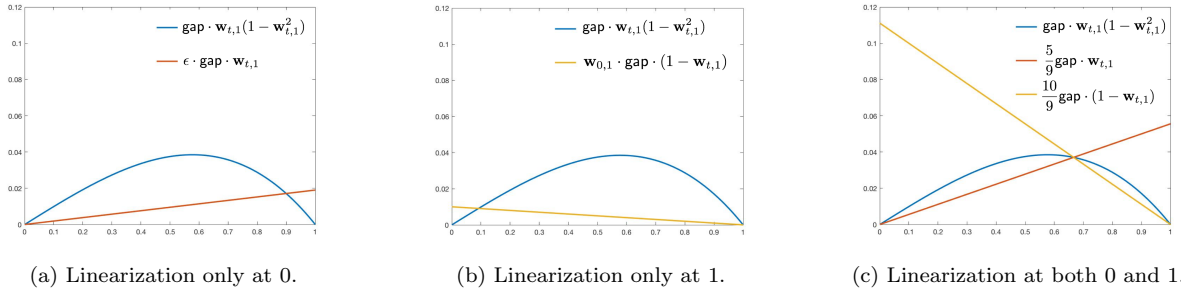


Figure 2: In (a), we only linearize at 0 and use $\epsilon \cdot \text{gap} \cdot \mathbf{w}_{t,1}$ to lower bound Equation 1.9 for $\mathbf{w}_{t,1} \in [0, \sqrt{1-\epsilon}]$. In (b), we only linearize at 1 and use $(\mathbf{w}_{0,1} \cdot \text{gap} \cdot (1 - \mathbf{w}_{t,1}))$ for $\mathbf{w}_{t,1} \in [\mathbf{w}_{0,1}, 1]$. On the other hand, in (c), we linearize at both 0 and 1. For $\mathbf{w}_{t,1} \in [0, \frac{2}{3}]$, we use $\frac{5}{9} \text{gap} \cdot \mathbf{w}_{t,1}$ while for $\mathbf{w}_{t,1} \in [\frac{2}{3}, 1]$ we use $\frac{10}{9} \text{gap} \cdot (1 - \mathbf{w}_{t,1})$. One can see that the lower bounds in (c) are much tighter than that in (a) and (b) in the sense that the slopes are of order $\Omega(\text{gap})$ instead of $O(\epsilon \cdot \text{gap})$ or $O(\mathbf{w}_{0,1} \cdot \text{gap})$.

The main difference between linearizing only at a single fixed point and linearizing at two fixed points is the *slope* in the linearization. Note that the slopes of the linearizations in Figure 2a and Figure 2b are $\epsilon(\lambda_1 - \lambda_2)$ and $\mathbf{w}_{0,1}(\lambda_1 - \lambda_2)$ respectively while the slope is of the order $\Omega(\lambda_1 - \lambda_2)$ in Figure 2c. As the slope corresponds to the *speed* of the convergence, the extra ϵ or $\mathbf{w}_{0,1}$ in the slope of linearization at a single fixed point would result in an extra ϵ^{-1} or $\mathbf{w}_{0,1}^{-1}$ in the convergence rate. See Figure 2 for a pictorial explanation.

Another key inspiration from the continuous dynamic is the *ODE trick* which provides a close form characterization of the dynamic in terms of the drifting term captured by the continuous dynamic and the noise term originated from the linearization and discretization. The ODE trick is inspired by the solution to a linear ordinary differential equation (linear ODE). Consider the following simple linear ODE

$$\frac{dy(t)}{dt} = ay(t) + b(t)$$

for some constant a and function $b(t)$. To put into the context, one can think of a as the drifting term and $b(t)$ as the noise term in the continuous Oja's rule due to the linearization⁵. By the standard tool for solving linear ODE, the solution of $y(t)$ at $t = T$ is

$$y(T) = e^{aT} \cdot \left(y(0) + \int_0^T e^{-at} b(t) dt \right). \quad (1.10)$$

⁵In the biological Oja's rule, the *discretization* also contributes in the noise term.

From the above equation, one can see that the solution of a linear ODE extracts the drifting term into a *multiplier* e^{aT} and decouples the initial condition $y(0)$ with the noise term $\int_0^T e^{-at}b(t)dt$. As a consequence, once we can show that the noise term is much smaller than the initial value, then $y(T)$ is dominated by the drifting term $e^{aT}y(0)$ and thus we are able to analyze the progress of $y(T)$.

To sum up, the continuous dynamic informs us to linearize the biological Oja's rule at different centers in different phases of the analysis. Further, the ODE trick provides us a closed-form approximation to the dynamic. We are then able to analyze the biological Oja's rule in *one-shot* rather than doing the traditional step-by-step analysis.

Insight 2: One-shot analysis instead of step-by-step analysis The second insight of this work is performing an *one-shot analysis* instead of the traditional step-by-step analysis (e.g., [AZL17]).

Traditional step-by-step analysis To see the difference, let us illustrate how would the step-by-step analysis on the biological Oja's rule work as follows. Denote the natural filtration as $\{\mathcal{F}_t\}$ where $\mathcal{F}_t$ is the σ -algebra generated by $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_t$. For any $t \in \mathbb{N}$, we have

$$\begin{aligned}\mathbb{E}[\mathbf{w}_{t,1}] &= \mathbb{E}\left[\mathbb{E}\left[\mathbf{w}_{t-1,1} + \eta_t(\mathbf{x}_t^\top \mathbf{w}_{t-1})\mathbf{x}_{t,1} - \eta_t(\mathbf{x}_t^\top \mathbf{w}_{t-1})^2 \mathbf{w}_{t-1,1} \mid \mathcal{F}_{t-1}\right]\right] \\ &= \mathbb{E}\left[\mathbf{w}_{t-1,1} + \eta_t \lambda_1 \mathbf{w}_{t-1,1} - \eta_t \left(\sum_{i=1}^n \lambda_i \mathbf{w}_{t-1,i}^2\right) \mathbf{w}_{t-1,1}\right]\end{aligned}$$

where the second equation is due to the fact that for any $i, j \in [n]$, $\mathbb{E}[\mathbf{x}_{t,i}\mathbf{x}_{t,j} \mid \mathcal{F}_{t-1}] = A_{ij} = \lambda_i \cdot \mathbf{1}_{i=j}$ and for any $i \in [n]$, $\mathbb{E}[\mathbf{w}_{t-1,i} \mid \mathcal{F}_{t-1}] = \mathbf{w}_{t-1,i}$. In a step-by-step analysis, one then argues that the expectation $\mathbb{E}[\mathbf{w}_{t,1}]$ would be improved from $\mathbb{E}[\mathbf{w}_{t-1,1}]$ by a certain factor. Then, an induction on each step followed by showing concentration would give some convergence rate guarantee. However, there are two difficulties in getting optimal convergence rate (these difficulties usually also appear in the step-by-step analysis for other problems).

- First, there are some non-linear terms of $\mathbf{w}_{t-1,1}$ in the update noise. This usually requires some hacks tailored to the specific problem to enable the analysis.
- Second, the improvement factor at each step can depend on $\mathbf{w}_{t-1}$ and at worst case, the dynamic can show no improvement or even deteriorate. Taking expectation loses precise controls of the values of $\mathbf{w}_{t-1}$. This makes naive martingale analysis difficult to work and probably requires more ad hoc tricks.

For instance, the first difficulty is exactly what [AZL17] encountered in their analysis for a variant of the biological Oja's rule. They resolved the first difficulty by decomposing the non-linear term in the dynamic into a *multi-dimensional chain* and carefully bounding the chain with strong assumptions on learning rates to enable martingale analysis. They used extremely delicate and complicated techniques tailored to the dynamic to achieve optimal convergence rate. The biological Oja's rule, in addition to having the first difficulty, also has the second difficulty (see Appendix C for more discussions). Therefore, applying the traditional step-by-step analysis on the biological Oja's rule will encounter great obstacles.

Our one-shot analysis In this work, we use an *one-shot* analysis to avoid the complication of a step-by-step analysis. Namely, instead of looking at the process iteratively, we study the entire dynamic at once. Two key ingredients are needed to implement such an one-shot analysis: (i) a closed-form characterization of the dynamic and (ii) stopping time techniques. As discussed in the previous discussion, the continuous dynamic of the biological Oja's rule inspires us to get a closed-form lower bound for $\mathbf{w}_{t,1}$ by the *ODE trick*. Concretely, as a simplified example⁶, we have

$$\mathbf{w}_{T,1} = H^T \cdot \left(\mathbf{w}_{0,1} + \sum_{t=1}^T \frac{N_t}{H^t} \right) \quad (1.11)$$

⁶In general, the multiplier term also varies with respect to time t .

where $H > 1$ is the multiplier term and $\{N_t\}$ is the noise term which forms a martingale on the natural filtration. See [Corollary 6.4](#) and [Corollary 7.3](#) for a precise formulation of H and $\{N_t\}$ in our analysis. Intuitively, one should think of $H^T \mathbf{w}_{0,1}$ as the *drifting term* and the other part as the *noise term*. The goal of the ODE trick in the discrete dynamic is to show that the drifting term dominates the noise term.

To show that the noise in [Equation 1.11](#) is small, Azuma's inequality (see [Lemma 2.4](#)) would be a natural tool to start with. However, the *bounded difference* condition in Azuma's inequality would immediately cause an issue: the noise at time t is correlated with $\mathbf{w}_{t-1,1}$ and thus one cannot get a small bounded difference almost surely. For example, suppose the bounded difference of $\{N_t\}$ at time t is at most $\mathbf{w}_{t-1,1}^2$. Since we do not yet know the behavior of $\mathbf{w}_{t-1,1}$, we can only upper bound the bounded difference of $\{N_t\}$ in the worst case⁷ by $1 + o(1)$. In the meantime, both $\mathbf{w}_{t,1}^2$ and the noise are expected to be very small in the early stage of the dynamic with high probability.

To circumvent this obstacle, we consider the *stopped process* of the original martingale in which the bounded difference is under control. For example, consider the above situation where the noise term $\{N_t\}$ is a martingale and a stopping time τ for the event $\{\mathbf{w}_{\tau,1}^2 \geq 0.1\}$. The stopped process, denoted by $\{N_{t \wedge \tau}\}$ where $t \wedge \tau = \min\{t, \tau\}$, is a process that simulates $\{N_t\}$ and *stops* at the first time t^* such that $\mathbf{w}_{t^*,1}^2 \geq 0.1$. It is known that a stopped process of a martingale is also a martingale. Furthermore, the bounded difference of the stopped process $\{N_{t \wedge \tau}\}$ would be 0.1 almost surely by the choice of τ . It turns out that this improvement in the bounded difference condition drastically increases the quality of Azuma's inequality and gives the desiring concentration for the stopped process.

There is one last missing step before showing the dominance of $\mathbf{w}_{0,1}$ in [Equation 1.11](#): we have to show that the concentration for the stopped process $\{N_{t \wedge \tau}\}$ can be extended to the original process $\{N_t\}$. We achieve this task by developing a *pull-out lemma* which is able to utilize the structure of the martingale and pull out the stopping time from a concentration inequality.

Insight 3: Maximal martingale inequality and pull-out lemma In general, there is no hope to pull out the stopping time from a concentration inequality for the stopped process without blowing up the failure probability. The naive union bound would give a blow-up of factor T in the failure probability and it is undesirable.

Let $M_t = \sum_{t'=1}^t H^{-t'} N_{t'}$ be the noise term in the ODE trick (*i.e.*, [Equation 1.11](#)) and τ be a stopping time that ensures good bounded difference condition. Note that as $\{N_t\}$ is a martingale, we know that $\{M_{t \wedge \tau}\}$ is also a martingale. There are two key ingredients to pull out the stopping time from $\{M_{t \wedge \tau}\}$, *i.e.*, the stopped process of the noise term.

First, we use the *maximal* concentration inequality (*e.g.*, [Lemma 2.5](#)) which gives the following stronger guarantee than the traditional Azuma's inequality.

$$\Pr \left[\sup_{1 \leq t \leq T} |M_{t \wedge \tau} - M_0| \geq a \right] < \delta \quad (1.12)$$

for some $a > 0$, $T \in \mathbb{N}$, and $\delta \in (0, 1)$. Note that the maximal concentration inequality gives concentration for any $1 \leq t \leq T$ without paying an union bound.

Second, we identify a *chain structure* on the martingale and the stopping time τ we are working with. Concretely, we are able to show that for all $t \in [T]$,

$$\Pr \left[\tau \geq t + 1 \mid \sup_{1 \leq t' \leq t} |M_{t'} - M_0| < a \right] = 1. \quad (1.13)$$

Namely, if the bad event has not happened, then the martingale would not stop immediately. Intuitively, [Equation 1.13](#) holds due to the ODE trick because $\{\sup_{1 \leq t' \leq t} |M_{t'} - M_0| < a\}$ implies the noise term to be small and thus the drifting term dominates. As τ is properly chosen such that the martingale would not stop if the process $\mathbf{w}_t$ followed the drifting term, we know that $\tau \geq t + 1$.

Combining the above two ingredients (*i.e.*, [Equation 1.12](#) and [Equation 1.13](#)), we are able to show in the pull-out lemma that

$$\Pr \left[\sup_{1 \leq t \leq T} |M_t - M_0| \geq a \right] < \delta,$$

⁷This is because we are able to upper bound $\mathbf{w}_{t-1,1}$ by $1 + o(1)$ almost surely. See [Section 5.2](#). Note that there are ways to get better bounded difference condition in the worst case but this is still not sufficient.

i.e., the stopping time has been *pulled out*.

Let us end this subsection with a high-level sketch on the proof for the pull-out lemma. The key idea is to consider another stopping time τ' for the event $\{|M_{\tau'} - M_0| \geq a\}$ and partition the probability space of the error event $\{\sup_{1 \leq t \leq T} |M_t - M_0| \geq a\}$ in to two parts P_1 and P_2 with the following properties. In P_1 , we can show that

$$\Pr \left[\sup_{1 \leq t \leq T} |M_t - M_0| \geq a, P_1 \right] = \Pr \left[\sup_{1 \leq t \leq T} |M_{t \wedge \tau} - M_0| \geq a, P_1 \right].$$

As for P_2 , we use the chain condition in [Equation 1.13](#) to show that the probability of error event is 0 based on a *diagonal argument*. Thus, we have

$$\begin{aligned} \Pr \left[\sup_{1 \leq t \leq T} |M_t - M_0| \geq a \right] &= \Pr \left[\sup_{1 \leq t \leq T} |M_t - M_0| \geq a, P_1 \right] + \Pr \left[\sup_{1 \leq t \leq T} |M_t - M_0| \geq a, P_2 \right] \\ &= \Pr \left[\sup_{1 \leq t \leq T} |M_{t \wedge \tau} - M_0| \geq a, P_1 \right] + 0 \\ &\leq \Pr \left[\sup_{1 \leq t \leq T} |M_{t \wedge \tau} - M_0| \geq a \right] < \delta. \end{aligned}$$

See [Section 7.4](#) and [Figure 3](#) for more details on the chain condition for biological Oja's rule and how to partition the probability space of the error event.

1.4 Related works

Biological Oja's variants and other Hebbian learning rules Computational neuroscientists have proposed several variants of the biological Oja's rule that solve different computational problems. For example, least square learning [[OBL00](#)], streaming PCA with ℓ_1 constraint [[Apa12](#)], and subspace learning [[PHC15](#)]. However, similar to the situation in the biological Oja's rule, people only showed guarantees on the convergence in the limit and did not provide a convergence rate analysis. Some works also attempted to obtain more than one principal component using variants of Oja's rule, but these dynamics are less biologically-plausible because they are either not local [[San89](#)] [[Oja92](#)] or store complicated local variables [[PHC15](#)]. Other forms of Hebbian learning rules has also been developed for applications other than PCA. For example, BCM rule [[BCM82](#)] and covariance rule [[Sej77](#)] are two well-known Hebbian learning rules which were successfully applied to explain the input selectivity in the development of the visual field. A recent work by Lynch and Mallmann-Trenn [[LMT19](#)] used Oja's rule to learn hierarchically structured concepts. They provided theoretical guarantee for the efficiency in a special case where the input is deterministic or nearly deterministic while we can handle arbitrary inputs.

Oja's rule in machine learning Unlike the situation in the biological Oja's rule, a line of recent exciting results [[HP14](#), [DSOR15](#), [BDWY16](#), [Sha16](#), [JJK⁺16](#), [AZL17](#)] showed convergence rate analysis for variants of Oja's rule in the machine learning community. Since the update rules of these works are not biologically-plausible, we call them *ML Oja's rules* to distinguish from the biological Oja's rule.

To see the difference between the biological Oja's rule and the ML Oja's rule, let us take the update rule from [[Sha16](#), [JJK⁺16](#), [AZL17](#)] as an example. Note that the other variants of ML Oja's rule also have the similar fundamental difference to the biological Oja's rule as illustrated by the following example. Let $\mathbf{w}_t \in \mathbb{R}^n$ be the output vector at time $t = 0, 1, \dots, T$, the update rule is

$$\mathbf{w}_t = \prod_{t'=1}^t (1 + \eta_{t'} \mathbf{x}_{t'} \mathbf{x}_{t'}^\top) \mathbf{w}_0$$

and the output is $\mathbf{w}_T / \|\mathbf{w}_T\|_2$. Note that the above update rule is equivalent to [Equation 1.5](#), *i.e.*, applying Taylor's expansion on the ML Oja's rule and truncating the higher-order terms would result in biological Oja's rule.

A natural idea would be trying to *couple* the biological Oja’s rule with the ML Oja’s rule by showing that for all $t \in \mathbb{N}$, the weight vectors from the two dynamics would be close to each other. However, this seems to be more difficult than direct analysis and we leave it as an interesting open problem to investigate whether this is the case. Moreover, the corresponding continuous dynamics suggest an intrinsic difference between the two: the continuous version of the ML Oja’s rule can be tightly characterized by a single linear ODE while that of the biological Oja’s rule requires two linear ODEs in different regimes for tight analysis. See [Section 3](#) and [Appendix C](#) for more details.

To sum up, the biological Oja’s rule and the ML Oja’s rule are similar but the analysis of the later cannot be directly applied to the former. While following the proof idea for the ML Oja’s rule might give some hints on how to analyze the biological Oja’s rule, in this work we develop a completely different framework (as briefly discussed in [Section 1.3](#)). This framework not only gives the first and nearly optimal convergence rate guarantee for the biological Oja’s rule, but also could improve the convergence rate of the ML Oja’s rule with better logarithmic dependencies and we leave it as a future work.

Comparing with other streaming PCA algorithms Streaming PCA is a well-studied and challenging computational problem. Many works [[DSOR15](#), [Sha16](#), [LWLZ18](#), [JJK⁺16](#), [AZL17](#)] provided theoretical guarantees for streaming PCA algorithms. Interestingly, all of the streaming PCA algorithms in these works are some variants of the biological Oja’s rule.

Recall that there are two standard convergence notions: the global convergence where $\mathbf{w}_0$ is an uniformly random unit vector and the local convergence where $\mathbf{w}_0$ is constantly correlated with the top eigenvector. There are 5 parameters of interest: the dimension $n \in \mathbb{N}$, the eigenvalue gap $\text{gap} := \lambda_1 - \lambda_2 \in (0, 1)$, the top eigenvalue $\lambda_1 \in (0, 1)$, the error parameter $\epsilon \in (0, 1)$, and the failure probability $\delta \in (0, 1)$. Ideally, the goal is to achieve the information-theoretic lower bound $\Omega(\lambda_1 \text{gap}^{-2} \epsilon^{-1} \log(\delta^{-1}))$ given by [[AZL17](#)]. Prior to this work, the state-of-the-art for both global and local convergences are achieved by [[AZL17](#)] using ML Oja’s rule (see the second to last row of [Table 1](#)). In this work, as a byproduct, the convergence rate we get for the biological Oja’s rule outperforms [[AZL17](#)] by a logarithmic factor in both settings. See [Table 1](#) for a summary.

Algorithm	Reference	Any Input	Global Convergence		Local Convergence	
			Convergence Rate	Degree in Log Terms [*]	Convergence Rate	Degree in Log Terms [*]
Biological Oja’s Rule	This Work	Y	$\tilde{O}\left(\frac{\lambda_1}{\text{gap}^2} \cdot \frac{1}{\epsilon \wedge \delta^2}\right)$	3	$\tilde{O}\left(\frac{\lambda_1}{\text{gap}^2} \cdot \frac{1}{\epsilon}\right)$	2
ML Oja’s Rule	[DSOR15]	N	$\tilde{O}\left(\frac{n}{\text{gap}^2} \cdot \frac{1}{\epsilon}\right)^\dagger$	- $\ddagger$	$\tilde{O}\left(\frac{n}{\text{gap}^2} \cdot \frac{1}{\epsilon}\right)^\dagger$	- $\ddagger$
	[Sha16]	Y	$\tilde{O}\left(\frac{n}{\text{gap}^2} \cdot \frac{1}{\epsilon}\right)^\dagger$	- $\ddagger$	$\tilde{O}\left(\frac{n}{\text{gap}^2} \cdot \frac{1}{\epsilon}\right)^\dagger$	- $\ddagger$
	[LWLZ18]	N	$\tilde{O}\left(\frac{\lambda_1 n}{\text{gap}^2} \cdot \frac{1}{\epsilon \delta^6}\right)$	- $\ddagger$	$\tilde{O}\left(\frac{\lambda_1 n}{\text{gap}^2} \cdot \frac{1}{\epsilon \delta^4}\right)$	- $\ddagger$
	[JJK⁺16]	Y	$\tilde{O}\left(\frac{\lambda_1}{\text{gap}^2} \cdot \frac{1}{\epsilon \delta^3}\right)$	2	$\tilde{O}\left(\frac{\lambda_1}{\text{gap}^2} \cdot \frac{1}{\epsilon \delta^3}\right)$	2
	[AZL17]	Y	$\tilde{O}\left(\frac{\lambda_1}{\text{gap}^2} \cdot \frac{1}{\epsilon \wedge \delta^2}\right)$	≥ 4	$\tilde{O}\left(\frac{\lambda_1}{\text{gap}^2} \cdot \frac{1}{\epsilon}\right)$	≥ 3
Any Algorithm	[AZL17]	$\Omega\left(\frac{\lambda_1}{\text{gap}^2} \cdot \frac{\log \frac{1}{\delta}}{\epsilon}\right)^\S$				

Table 1: Convergence rate for biological Oja’s rule and ML Oja’s rule in solving streaming PCA. The “Any Input” column indicates that whether the analysis has higher moment conditions on the unknown distribution $\mathcal{D}$. Note that having higher moment conditions would drastically simplify the problem because the non-linear terms in the update rule can then be non-trivially replaced with the first order term.

Algorithms inspired by biological neural networks In recent years, the study of the algorithmic aspect of mathematical models for biological neural networks is an emerging field in theoretical CS. For example, the efficiency of spiking neural networks in solving the *winner-take-all (WTA)* problem [LMP17a, LMP17b, LMP17c, LM18, SCL19], the efficiency of spiking neural networks in storing temporal information [WL19, HP19], assemblies [LMPV18, PV18], and spiking neural networks in solving optimization problems [CCL19, Peh19]. Under this context, this work provides an algorithmic insight in a plasticity learning rule that solves streaming PCA.

2 Preliminaries

2.1 Notations

We use $\mathbb{N} = \{1, 2, \dots\}$ and $\mathbb{N}_{\geq 0} = \{0, 1, \dots\}$. For each $n \in \mathbb{N}$, denote $[n] = \{1, 2, \dots, n\}$ and $[n]_{\geq 0} = \{0, 1, \dots, n\}$. For a vector indexed by time t , *e.g.*, $\mathbf{w}_t$, its i^{th} coordinate is denoted by $\mathbf{w}_{t,i}$. The notation $\tilde{O}$ (similarly, $\tilde{\Omega}$ and $\tilde{\Theta}$) is the same as the big-O notation by ignoring extra poly-logarithmic term. $\mathbf{1}_E$ stands for the indicator function for a probability event E . We sometimes abuse the big O notation by $y = O(x)$ meaning $|y| = O(x)$ and this will be clear in the context. Throughout the paper, λ is used to denote the vector $(\lambda_1, \lambda_2, \dots, \lambda_n)$ where $\lambda_1 > \lambda_2 \geq \dots \geq \lambda_n \geq 0$ are the eigenvalues of the covariance matrix A . $\text{diag}(\lambda)$ denotes the diagonal matrix with λ on the diagonal.

2.2 Probability toolbox

2.2.1 Random unit vector

The following lemma shows that with random initialization, the inner product between $\mathbf{w}_0$ and the top eigenvector of the covariance matrix is of order $\Omega(1/\sqrt{n})$ with high probability.

Lemma 2.1. *For any $\delta \in (0, 1)$ and unit vector $\mathbf{v}_0 \in \mathbb{R}^n$, let $\mathbf{v}$ be a random unit vector in $\mathbb{R}^n$. Then*

$$\Pr \left[\left| \mathbf{v}_0^\top \mathbf{v} \right|^2 \leq \frac{10\delta^2}{n} \right] < \delta.$$

2.2.2 Random process and concentration inequality

Random process is a central tool in this paper. Let us start with the most general definition on adapted random process.

Definition 2.2 (Adapted random process). *Let $\{X_t\}_{t \in \mathbb{N}_{\geq 0}}$ be a sequence of random variables and $\{\mathcal{F}_t\}_{t \in \mathbb{N}_{\geq 0}}$ be a filtration. We say $\{X_t\}_{t \in \mathbb{N}_{\geq 0}}$ is an adapted random process with respect to $\{\mathcal{F}_t\}_{t \in \mathbb{N}_{\geq 0}}$ if for each $t \in \mathbb{N}_{\geq 0}$, the σ -algebra generated by $X_0, X_1, \dots, X_t$ is contained in $\mathcal{F}_t$.*

In most of the situation, we use $\mathcal{F}_t$ to denote the *natural filtration* of $\{X_t\}_{t \in \mathbb{N}_{\geq 0}}$, namely, $\mathcal{F}_t$ is defined as the σ -algebra generated by $X_0, X_1, \dots, X_t$. One of the most common adapted processes is the martingale.

Definition 2.3 (Martingale). *Let $\{M_t\}_{t \in \mathbb{N}_{\geq 0}}$ be a sequence of random variables and let $\{\mathcal{F}_t\}_{t \in \mathbb{N}}$ be its natural filtration. We say $\{M_t\}_{t \in \mathbb{N}_{\geq 0}}$ is a martingale if for each $t \in \mathbb{N}$, $\mathbb{E}[M_{t+1} \mid \mathcal{F}_t] = M_t$.*

Note that for any adapted random process $\{X_t\}_{t \in \mathbb{N}_{\geq 0}}$, one can always turn it into a martingale by defining $M_0 = X_0$ and for any $t \in \mathbb{N}$, let $M_t = X_t - \mathbb{E}[X_t \mid \mathcal{F}_{t-1}]$. When the difference of a martingale can be bounded almost surely, the *Azuma's inequality* provides an useful concentration inequality with exponential tail.

*Let $f(\log n, \log(1/\epsilon), \log(1/\delta), \log(1/\text{gap}))$ be the polynomial of the logarithmic dependencies in the convergence rate. We compare the maximum degree of f among different analyses. Note that this measure makes sense when $n, 1/\epsilon, 1/\delta, 1/\text{gap}$ are polynomially related.

[†]Both [DSOR15] and [Sha16] cannot handle arbitrary failure probability so we ignore their δ dependency on the table.

[‡]In [DSOR15, Sha16, LWLZ18], their convergence rates are far from the information-theoretic lower bound. So we do not trace down their logarithmic dependencies.

[§]In [AZL17], they only stated $\Omega(\frac{\lambda_1}{\text{gap}} \cdot \frac{1}{\epsilon})$ lower bound. We observe that their lower bound can be improved by a $\log(1/\delta)$ factor using the fact that distinguishing a fair coin from a biased coin with probability at least δ requires $\Omega(\log(1/\delta))$ samples.

Lemma 2.4 (Azuma's inequality [Azu67]). *Let $\{M_t\}_{t \in \mathbb{N}_{\geq 0}}$ be a martingale. Let $T \in \mathbb{N}$ and $a, c \geq 0$ be some constants. Suppose for each $t = 1, 2, \dots, T$, $|M_t - M_{t-1}| \leq c$ almost surely, then we have*

$$\Pr[|M_T - M_0| \geq a] < \exp\left(-\Omega\left(\frac{a^2}{c^2 T}\right)\right).$$

The following maximal Azuma's inequality shows that one can even get union bound for free with the help of Doob's inequality.

Lemma 2.5 (Maximal Azuma's inequality [HMRAR13, Section 3]). *Let $\{M_t\}_{t \in \mathbb{N}_{\geq 0}}$ be a martingale. Let $T \in \mathbb{N}$ and $a, c \geq 0$ be some constants. Suppose for each $t = 1, 2, \dots, T$, $|M_t - M_{t-1}| \leq c$ almost surely, then we have*

$$\Pr\left[\sup_{0 \leq t \leq T} |M_t - M_0| \geq a\right] < \exp\left(-\Omega\left(\frac{a^2}{c^2 T}\right)\right).$$

The Azuma's inequality can be strengthened by considering the conditional variance. This is known as the Freedman's inequality.

Lemma 2.6 (Freedman's inequality [Fre75]). *Let $\{M_t\}_{t \in \mathbb{N}_{\geq 0}}$ be a martingale. Let $T \in \mathbb{N}$ and $a, c, \sigma_t \geq 0$ be some constants for all $t \in [T]$. Suppose for each $t = 1, 2, \dots, T$, $|M_t - M_{t-1}| \leq c$ almost surely and $\text{Var}[M_t \mid \mathcal{F}_{t-1}] \leq \sigma_t^2$, then we have*

$$\Pr\left[\sup_{0 \leq t \leq T} |M_t - M_0| \geq a\right] < \exp\left(-\Omega\left(\frac{a^2}{\sum_{t=1}^T \sigma_t^2 + ca}\right)\right).$$

Finally, we state a corollary of Freedman's inequality for adapted random process with small conditional expectation.

Corollary 2.7. *Let $\{M_t\}_{t \in \mathbb{N}_{\geq 0}}$ be a random process. Let $T \in \mathbb{N}$ and $a, c, \sigma_t, \mu_t \geq 0$ be some constants for all $t \in [T]$. Suppose for each $t = 1, 2, \dots, T$, $|M_t - M_{t-1}| \leq c$ almost surely, $\text{Var}[M_t \mid \mathcal{F}_{t-1}] \leq \sigma_t^2$, and $|\mathbb{E}[M_t - M_{t-1} \mid \mathcal{F}_{t-1}]| \leq \mu_t$, then we have*

$$\Pr\left[\sup_{0 \leq t \leq T} |M_t - M_0| \geq a + \max_{1 \leq t \leq T} \sum_{t=1}^T \mu_t\right] < \exp\left(-\Omega\left(\frac{a^2}{\sum_{t=1}^T \sigma_t^2 + ca}\right)\right).$$

2.2.3 Stopping time

One powerful technique for studying martingale is the notion of *stopping time* defined as follows.

Definition 2.8 (Stopping time). *Let $\{X_t\}_{t \in \mathbb{N}_{\geq 0}}$ be an adapted random process associated with filtration $\{\mathcal{F}_t\}_{t \in \mathbb{N}_{\geq 0}}$. An integer-valued random variable τ is a stopping time for $\{X_t\}_{t \in \mathbb{N}_{\geq 0}}$ if for all $t \in \mathbb{N}$, $\{\tau = t\} \in \mathcal{F}_t$.*

Let $\{M_t\}_{t \in \mathbb{N}_{\geq 0}}$ be a martingale, the most common stopping time for $\{M_t\}_{t \in \mathbb{N}_{\geq 0}}$ is of the following form. For any $a \in \mathbb{R}$, let

$$\tau := \min_{M_t \geq a} t.$$

Namely, τ is the first time when the martingale becomes at least a . For convenience, in the rest of the paper, we would define stopping time of this form by saying “ τ is the stopping time for the event $\{M_\tau \geq a\}$ ”.

Given a martingale $\{M_t\}_{t \in \mathbb{N}_{\geq 0}}$ and a stopping time τ , it is then natural to consider the corresponding *stopped process* $\{M_{t \wedge \tau}\}_{t \in \mathbb{N}_{\geq 0}}$ where $t \wedge \tau = \min\{t, \tau\}$ is also a random variable. An useful and powerful fact here is that the stopped process of a martingale is also a martingale. See [Wil91, Theorem 10.9] for a proof for this classic result. In the rest of this paper, we sometimes write $\{M_{t_0 + t \wedge \tau}\}$ to denote martingale starts with time t_0 and τ is used as the stopping time after t_0 (instead of starting from time 0).

2.2.4 Brownian motion

In [Section 3](#), we consider a continuous version of biological Oja's rule by modeling the input stream as a Brownian motion. Here, we provide background that is sufficient for the readers to understand the discussion there.

First, we introduce the 1-dimensional Brownian motion using the following operational definition. In the following, we use $N(\mu, \sigma^2)$ to denote the Gaussian distribution with mean μ and variance σ^2 .

Definition 2.9 (1-dimensional Brownian motion). *Let $\{\beta_t\}_{t \geq 0}$ be a real-valued random process. We say $\{\beta_t\}_{t \geq 0}$ is a 1-dimensional Brownian motion if the following holds.*

- For any $0 \leq t_1 < t_2$, $\beta_{t_2} - \beta_{t_1} \sim N(0, t_2 - t_1)$.
- For any $0 = t_0 \leq t_1 < \dots < t_\ell$, for each $i \in [\ell]$, $\beta_{t_i} - \beta_{t_{i-1}}$ are independent.

With the above definition, it is then natural to consider some variants such as putting n independent copies of 1-dimensional Brownian motion into a vector, *i.e.*, the n -dimensional Brownian motion, or applying linear operations on an n -dimensional Brownian motion, or considering the *calculus* on Brownian motion by looking at $d\beta_t = \lim_{\Delta \rightarrow 0} \beta_{t+\Delta} - \beta_t$. The role of Brownian motion in the study of continuous random process is similar to Gaussian random variance in discrete random process and many properties in the discrete world directly extend to the continuous world. One property of Brownian motion though obviously does not hold in the discrete setting and might be counter-intuitive for people who see this in the first time. This is the *quadratic variation* of Brownian motion as stated below.

Lemma 2.10 (Quadratic variation of Brownian motion). *Let $\{\beta_t\}_{t \geq 0}$ and $\{\beta'_t\}_{t \geq 0}$ be two independent 1-dimensional Brownian motions. The following holds almost surely.*

$$d\beta_t^2 = dt \quad \text{and} \quad d\beta_t d\beta'_t = 0.$$

We omit the proof of [Lemma 2.10](#) here and refer the interested readers to standard textbook such as [\[LG16\]](#) for more details on Brownian motion.

2.3 ODE toolbox

Lemma 2.11 (ODE trick for scalar). *Let $\{X_t\}_{t \in \mathbb{N}_{\geq 0}}$, $\{A_t\}_{t \in \mathbb{N}}$, and $\{H_t\}_{t \in \mathbb{N}}$ be sequences of random variables with the following dynamic*

$$X_t = H_t X_{t-1} + A_t \tag{2.12}$$

for all $t \in \mathbb{N}$. Then for all $t_0, t \in \mathbb{N}_{\geq 0}$ such that $t_0 < t$, we have

$$X_t = \prod_{i=t_0+1}^t H_i \cdot \left(X_{t_0} + \sum_{i=t_0+1}^t \frac{A_i}{\prod_{j=t_0+1}^i H_j} \right).$$

Proof of Lemma 2.11. For each $t_0 < i \leq t$, dividing [Equation 2.12](#) with $\prod_{j=t_0+1}^i H_j$ on both sides, we have

$$\frac{X_i}{\prod_{j=t_0+1}^i H_j} = \frac{X_{i-1}}{\prod_{j=t_0+1}^{i-1} H_j} + \frac{A_i}{\prod_{j=t_0+1}^i H_j}.$$

By telescoping the above equation from $t = t_0 + 1$ to t , we get the desiring expression. $\square$

Lemma 2.13 (ODE trick for vector). *Let $m \in \mathbb{N}$ and $\{X_t\}_{t \in \mathbb{N}_{\geq 0}}$, $\{A_t\}_{t \in \mathbb{N}}$ be sequences of m -dimensional random variables and $\{H_t\}_{t \in \mathbb{N}}$ be a sequence of random $m \times m$ matrices with the following dynamic*

$$X_t = H_t X_{t-1} + A_t \tag{2.14}$$

for all $t \in \mathbb{N}$. Then for all $t_0, t \in \mathbb{N}_{\geq 0}$ such that $t_0 < t$, we have

$$X_t = \prod_{i=t_0+1}^t H_i X_{t_0} + \sum_{i=t_0+1}^t \prod_{j=i+1}^t H_j A_i.$$

Proof of Lemma 2.13. The proof is a direct induction. $\square$

2.4 Approximation toolbox

Here we state some useful inequalities. Since some are quite standard, the proofs are omitted.

Lemma 2.15. *For any $x \in (-0.5, 1)$,*

$$1 + x \leq e^x \leq 1 + x + x^2 \leq 1 + 2x.$$

In fact for all $x \geq 0$, the first inequality holds.

Lemma 2.16. *For any $x \in (0, 0.5)$ and $t \in \mathbb{N}$,*

$$1 + \frac{xt}{2} \leq e^{\frac{xt}{2}} \leq (1 + x)^t \leq e^{xt}.$$

Lemma 2.17. *For any $\epsilon \in (0, 0.2)$, we have*

$$\left(\frac{\epsilon}{8}\right)^{1 - \frac{1}{\log \frac{1}{\epsilon}}} \leq \epsilon.$$

Proof. Rewrite the expression as the follows.

$$\left(\frac{\epsilon}{8}\right)^{1 - \frac{1}{\log \frac{1}{\epsilon}}} = \epsilon \cdot \left(\frac{8}{\epsilon}\right)^{\frac{1}{\log \frac{1}{\epsilon}}} \cdot \frac{1}{8}.$$

It suffices to show that the product of the last two terms is smaller than 1. Consider the logarithm of them, we have

$$\log \left(\left(\frac{8}{\epsilon}\right)^{\frac{1}{\log \frac{1}{\epsilon}}} \cdot \frac{1}{8} \right) = \frac{1}{\log \frac{1}{\epsilon}} \left(3 + \log \frac{1}{\epsilon} \right) - 3 = 3 \left(\frac{1}{\log \frac{1}{\epsilon}} - 1 \right) + 1.$$

When $\epsilon < 0.2$, we have $\log(1/\epsilon) > 2$ and the equation becomes

$$\leq \left(\frac{1}{2} - 1 \right) 3 + 1 < 0$$

as desired. □

3 Analyzing the Continuous Version of Oja's Rule

In this section, we introduce the continuous version of Oja's rule and analyze its convergence rate. The analysis here serves as an inspiration for attacking the discrete dynamic. To model the continuous dynamics, we use *Brownian motion* to capture the continuous stream of inputs. Surprisingly, it turns out that this continuous version of Oja's rule is *deterministic*. Thus, we are able to use the tools from ODE to easily give an exact characterization of how it converges to the top eigenvector of the covariance matrix. As a disclaimer, since the analysis for continuous Oja's rule is mainly for intuition, we would omit some mathematical details and point the interested readers to the corresponding resources.

3.1 Continuous Oja's rule is deterministic

In the rest of the section we are going to focus on the *diagonal case* where the covariance matrix $A = \text{diag}(\lambda)$ and the goal is showing that $\mathbf{w}_{t,1}$ goes to 1. This is sufficient since there is a reduction from the general case to the diagonal case as explained in [Section 5.1](#).

Intuitively, the continuous dynamic is the limiting process of biological Oja's rule with learning rate η going to 0. Formally, for each $i \in [n]$, let $(\beta_t^{(i)})_{t \geq 0}$ be an independent 1-dimensional Brownian motion and let $(B_t)_{t \geq 0}$ be an n -dimensional random process with the i^{th} entry being $B_{t,i} = \sqrt{\lambda_i} \beta_t^{(i)}$ for each $t \geq 0$. Now, the difference of B_t should then be thought of as $\eta \mathbf{x}_t$.

Concretely, to see why $(B_t)_{t \geq 0}$ captures the input behavior of streaming PCA in the continuous setting, let us first discretize $(B_t)_{t \geq 0}$ using constant step size $\Delta > 0$. Now, observe that for each $t \geq 0$, $B_{t+\Delta} - B_t$ is an isotropic Gaussian vector with the variance of the i^{th} entry being $\lambda_i \cdot \Delta$. Namely,

$$\frac{1}{\Delta} \mathbb{E} \left[(B_{t+\Delta} - B_t) (B_{t+\Delta} - B_t)^\top \right] = \text{diag}(\lambda). \quad (3.1)$$

Thus, by discretizing B_t into intervals of length $\Delta > 0$, $\left\{ \frac{1}{\sqrt{\Delta}} (B_{j \cdot \Delta} - B_{(j-1) \cdot \Delta}) \right\}_{j \in \mathbb{N}}$ naturally forms a stream of i.i.d. input⁸ with covariance matrix being A . To put this into the context of biological Oja's rule, one should think of $\eta = \Delta$, $\mathbf{x}_j = \frac{1}{\sqrt{\Delta}} \Delta B_j$, and $y_j = \mathbf{x}_j^\top \mathbf{w}_{j-1}$ for each $j \in \mathbb{N}$ where $\Delta B_j = (B_{j \cdot \Delta} - B_{(j-1) \cdot \Delta})$ ⁹. Then, we get the following dynamic.

$$\begin{aligned} \mathbf{w}_j &= \mathbf{w}_{j-1} + \eta \cdot y_j (\mathbf{x}_j - y_j \mathbf{w}_{j-1}) \\ &= \mathbf{w}_{j-1} + \Delta B_j^\top \mathbf{w}_{j-1} \Delta B_j - [\Delta B_j^\top \mathbf{w}_{j-1}]^2 \mathbf{w}_{j-1}. \end{aligned}$$

The above dynamics becomes continuous once we let $\Delta \rightarrow 0$. Formally, we replace¹⁰ $B_{t+\Delta} - B_t$ with dB_t and index the weight vector by $t \geq 0$, i.e., $(\mathbf{w}_t)_{t \geq 0}$. The above dynamic becomes the following SDE.

$$d\mathbf{w}_t = dB_t^\top \mathbf{w}_t dB_t - (dB_t^\top \mathbf{w}_t)^2 \mathbf{w}_t. \quad (3.2)$$

It might look absurd at first glance (for those who have not seen stochastic calculus before) that there is a quadratic term of dB_t in Equation 3.2. Nevertheless, it is in fact mathematically well-defined and we recommend standard resource such as [LG16] for more details. Intuitively, the quadratic term (which is formally called the *quadratic variation*) of a Brownian motion should be thought of as a *deterministic* quantity. Concretely, let $(\beta_t)_{t \geq 0}$ be a Brownian motion, we have $d\beta_t^2 = dt$ almost surely (see Lemma 2.10). Thus, for the $(B_t)_{t \geq 0}$ defined here, we would have

$$dB_{t,i} dB_{t,j} = \begin{cases} \lambda_i dt & , i = j \\ 0 & , i \neq j \end{cases}$$

for each $i, j \in [n]$. As a consequence, the continuous Oja's rule defined in Equation 3.2 can be rewritten as the following *deterministic* process almost surely.

$$d\mathbf{w}_t = [\text{diag}(\lambda) \mathbf{w}_t - \mathbf{w}_t^\top \text{diag}(\lambda) \mathbf{w}_t \mathbf{w}_t] dt. \quad (3.3)$$

With the continuous Oja's rule being deterministic as in Equation 3.3, it is then not difficult to have a tight analysis on its convergence using tools from ODE as explained in the next subsection.

3.2 One-sided versus two-sided linearization

In this subsection, we analyze Equation 3.3 by linearizing the dynamic at 0 and 1 respectively and get two incomparable convergence rates (Theorem 3.4 and Theorem 3.5).

Theorem 3.4 (Linearization at 0). *Suppose $\mathbf{w}_{0,1} > 0$. For any $\epsilon \in (0, 1)$, when $t \geq \Omega\left(\frac{\log(1/\mathbf{w}_{0,1}^2)}{\epsilon(\lambda_1 - \lambda_2)}\right)$, we have $\mathbf{w}_{t,1}^2 > 1 - \epsilon$.*

Theorem 3.5 (Linearization at 1). *Suppose $\mathbf{w}_{0,1} > 0$. For any $\epsilon \in (0, 1)$, when $t \geq \Omega\left(\frac{\log(1/\epsilon)}{\mathbf{w}_{0,1}(\lambda_1 - \lambda_2)}\right)$, we have $\mathbf{w}_{t,1}^2 > 1 - \epsilon$.*

⁸Though here is a caveat that the length of the input vector might not be 1. Nevertheless, the point of continuous dynamic is not to exactly characterize the limiting behavior of discrete Oja's rule. Instead, the goal here is to capture the intrinsic properties of the biological Oja's rule.

⁹Here we abuse the notation of Δ . When we write ΔB_j , the Δ is regarded as an *operator* instead of the interval length.

¹⁰This replacement might look weird for those who have not seen Brownian motion before. But this is standard and can be found in textbook such as [LG16].

The proofs for [Theorem 3.4](#) and [Theorem 3.5](#) are based on applying Taylor's expansion on [Equation 3.3](#) with center either being 0 or 1. Then, we approximate the dynamics with linear differential equations and use tools from ODE to get an tight analysis. See [Appendix B](#) for the details on the linearizations of continuous Oja's rule.

When starting with a random vector, *i.e.*, $\mathbf{w}_{0,1} = \Omega(1/\sqrt{n})$ with high probability, the above convergence rates become $O\left(\frac{\log n}{\epsilon(\lambda_1 - \lambda_2)}\right)$ and $O\left(\frac{\sqrt{n} \log(1/\epsilon)}{\lambda_1 - \lambda_2}\right)$ respectively. This indicates that linearizing only on one side (either at 0 or at 1) would not give tight analysis. Nevertheless, if we invoke [Theorem 3.4](#) with the error parameter being 0.5, then for some $t_1 = O\left(\frac{\log n}{\lambda_1 - \lambda_2}\right)$, we have $\mathbf{w}_{t_1,1} > 0.5$. Next, we invoke [Theorem 3.5](#) starting from $\mathbf{w}_{t_1}$ and with the error parameter being ϵ , then for some $t_2 = O\left(\frac{\log(1/\epsilon)}{\lambda_1 - \lambda_2}\right)$, we have $\mathbf{w}_{t_1+t_2,1} > 1 - \epsilon$. Putting these together, we have the following theorem combining the linearizations on both sides.

Theorem 3.6 (Linearization at both 0 and 1). *Suppose $\mathbf{w}_{0,1} > 0$. For any $\epsilon \in (0, 1)$, when*

$$t \geq \Omega\left(\frac{\log \frac{1}{\mathbf{w}_{0,1}^2} + \log \frac{1}{\epsilon}}{\lambda_1 - \lambda_2}\right),$$

we have $\mathbf{w}_{t,1}^2 > 1 - \epsilon$.

The above theorem for the convergence rate of the continuous Oja's rule gives three key insights. First, it suggests that one should linearize at 0 in the beginning of the process and switch to linearizing at 1 when $\mathbf{w}_{t,1}$ becomes $\Omega(1)$. Second, after the linearization, using linear ODE to give exact characterization of the dynamic would give tight analysis. Finally, the continuous dynamic is deterministic and will stay around the optimal region for all time after certain point. This suggests that the *for-all-time* guarantee could potentially happen in the original discrete setting.

4 Main Result

Now, let us state the formal version of the main theorem for the biological Oja's rule. In the following, all of the theorems and lemmas are stated with respect to the setting of [Problem 1.2](#) and [Definition 1.3](#). Thus, for simplicity, we would not repeat the setup in their statements.

In [Theorem 4.1](#), we show that it would not take too much time for the biological Oja's rule to achieve both constant error and arbitrary small error. Next, in [Theorem 4.2](#) we show that once $\mathbf{w}_t$ becomes ϵ -close to the top eigenvector $\mathbf{v}_1$, it will stay in the neighborhood of $\mathbf{v}_1$ for a long time. Thus, combining [Theorem 4.1](#) and [Theorem 4.2](#), we can get the standard convergence notions as stated in [Theorem 1.7](#) and [Theorem 1.8](#).

Theorem 4.1. *For any $\epsilon, \delta \in (0, 1)$, we have the following results.*

- (Phase 1: Toward constant error) *Suppose $\mathbf{w}_0$ is uniformly sampled from the unit sphere of $\mathbb{R}^n$. Let τ to be the stopping time for the event $\{\frac{\langle \mathbf{w}_\tau, \mathbf{v}_1 \rangle^2}{\|\mathbf{w}_\tau\|_2^2} \geq 2/3\}$ and let $\eta_t = \eta = O\left(\frac{\delta^2(\lambda_1 - \lambda_2)}{\lambda_1 \log^2 \frac{n}{\delta(\lambda_1 - \lambda_2)}}\right)$, then there exists $t_1 = O\left(\frac{\log(n/\delta)}{\eta(\lambda_1 - \lambda_2)}\right)$ such that $\Pr[\tau \leq t_1] \geq 1 - \delta$. Specifically, we have*

$$t_1 = O\left(\frac{\lambda_1 \log \frac{n}{\delta} \log^2 \frac{n}{\delta(\lambda_1 - \lambda_2)}}{\delta^2(\lambda_1 - \lambda_2)^2}\right).$$

- (Phase 2: Toward arbitrarily small error) *Suppose $\frac{\langle \mathbf{w}_0, \mathbf{v}_1 \rangle^2}{\|\mathbf{w}_0\|_2^2} = \Omega(1)$. For any $t \in \mathbb{N}$, let $\eta_t = \eta = O\left(\frac{\epsilon(\lambda_1 - \lambda_2)}{\lambda_1 \log \frac{\log \log \frac{1}{\epsilon}}{\delta}}\right)$, then there exists $t_2 = O\left(\frac{\log(1/\epsilon)}{\eta(\lambda_1 - \lambda_2)}\right)$ such that $\Pr\left[\frac{\langle \mathbf{w}_{t_2}, \mathbf{v}_1 \rangle^2}{\|\mathbf{w}_{t_2}\|_2^2} \geq 1 - \epsilon\right] \geq 1 - \delta$. Specifically, we have*

$$t_2 = O\left(\frac{\lambda_1 \log \frac{1}{\epsilon} \log \frac{\log \log \frac{1}{\epsilon}}{\delta}}{\epsilon(\lambda_1 - \lambda_2)^2}\right).$$

Proof structure of Theorem 4.1 To prove Theorem 4.1, we first reduce the general setting where the covariance matrix A being PSD to the special case where $A = \text{diag}(\lambda)$ in Section 5.1. Next, an important property of biological Oja's rule showing the boundedness of $\|\mathbf{w}_t\|_2^2$ is provided in Section 5.2. The rest of the proof proceeds in two phases: In Phase 1 (see Section 6), we analyze the speed of biological Oja's rule getting constant error using linearization at 0. This is summarized in Theorem 6.1. In Phase 2 (see Section 7), we analyze how fast does biological Oja's rule go from constant error regime to the ϵ error regime. This is summarized in Theorem 7.1.

Theorem 4.2. For any $\epsilon, \delta \in (0, 1)$, suppose $\frac{\langle \mathbf{w}_0, \mathbf{v}_1 \rangle^2}{\|\mathbf{w}_0\|_2^2} \geq 1 - \epsilon/2$, we have the following results.

- (Fixed learning rate) For any $T, t \in \mathbb{N}$, let $\eta_t = \eta = O\left(\frac{\epsilon(\lambda_1 - \lambda_2)}{\lambda_1 \log \frac{T}{\delta}}\right)$, then

$$\Pr \left[\forall t \leq T, \frac{\langle \mathbf{w}_t, \mathbf{v}_1 \rangle^2}{\|\mathbf{w}_t\|_2^2} \geq 1 - \epsilon \right] \geq 1 - \delta.$$

- (Slowly diminishing learning rate) For any $t \in \mathbb{N}$, let $\eta_t = O\left(\frac{\epsilon(\lambda_1 - \lambda_2)}{\lambda_1 \log \frac{t}{\delta}}\right)$, then

$$\Pr \left[\forall t \in \mathbb{N}, \frac{\langle \mathbf{w}_t, \mathbf{v}_1 \rangle^2}{\|\mathbf{w}_t\|_2^2} \geq 1 - \epsilon \right] \geq 1 - \delta.$$

Proof structure of Theorem 4.2 The proof of Theorem 4.2 is a direct application of Lemma 7.4 from Phase 2. Since there is no new technique involves except properly setting the parameters, we postpone the whole proof to Appendix F.

5 Preprocessing

Before the main analysis of biological Oja's rule, we provide two useful observations on the dynamic in this section. Specifically, we show in Section 5.1 that considering the covariance matrix being *diagonal* is sufficient for the analysis and in Section 5.2 that $\|\mathbf{w}_t\|_2^2 = 1 \pm O(\eta)$ almost surely for all $t \in \mathbb{N}$.

5.1 A reduction to the diagonal case

In this subsection, we show that it suffices to analyze the case where the covariance matrix A is a diagonal matrix D . Recall that A is defined as the expectation of $\mathbf{x}\mathbf{x}^\top$ and thus it is positive semidefinite. Namely, there exists an orthonormal matrix U and a diagonal matrix D such that $A = UDU^\top$. Especially, the eigenvalues of A , i.e., $1 \geq \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$, are the entries of D from top left to bottom right on the diagonal. Thus, by a change of basis, we can focus on the case where $A = D$ without loss of generality.

To see this, consider $\tilde{\mathbf{w}}_t = U\mathbf{w}_t$ and $\tilde{\mathbf{x}}_t = U\mathbf{x}_t$. As $U^\top U = UU^\top = I$, we have $\tilde{\mathbf{x}}_t^\top \tilde{\mathbf{w}} = \mathbf{x}_t^\top \mathbf{w}$ and $\mathbb{E}[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\top] = D$. Let $\mathbf{v}_1$ be the top eigenvector of A (i.e., the first row of U), we also have

$$\|\mathbf{w}_t - \mathbf{v}_1\|_2 = \|U\mathbf{w}_t - U\mathbf{v}_1\|_2 = \|\tilde{\mathbf{w}}_t - \mathbf{e}_1\|_2$$

where $\mathbf{e}_1$ is the indicator vector for the first coordinate. Namely, it suffices to analyze how fast does $\tilde{\mathbf{w}}_t$ converge to $\mathbf{e}_1$. Thus, in the rest of this paper, we without loss of generality consider the diagonal case where the goal would then be showing that $\mathbf{w}_{t,1}^2 \geq 1 - \epsilon$.

5.2 Bounded conditions of Oja's rule

In this subsection, we show that the ℓ_2 norm of the weight vector is always close to 1 almost surely.

Lemma 5.1. For any $\eta \in (0, 0.1)$, if for all $t \in \mathbb{N}$, $\eta_t \leq \eta$, then for all $t \in \mathbb{N}_{\geq 0}$, $1 - 10\eta \leq \|\mathbf{w}_t\|_2^2 \leq 1 + 10\eta$ almost surely.

Proof. Here we only prove the upper bound while the lower bound can be proved using the same argument. The proof is based on induction. For the base case where $t = 0$, we have $\|\mathbf{w}_0\|_2^2 = 1$ from the problem setting. For the induction step, consider any $t \in \mathbb{N}$ such that $\mathbf{w}_{t-1}$ satisfies the bounds, we have

$$\begin{aligned}\|\mathbf{w}_t\|_2^2 &= \|\mathbf{w}_{t-1}\|_2^2 + 2\eta_t \cdot \mathbf{w}_{t-1}^\top [\mathbf{x}_t^\top \mathbf{w}_{t-1} \mathbf{x}_t - (\mathbf{x}_t^\top \mathbf{w}_{t-1})^2 \mathbf{w}_{t-1}] + \eta_t^2 \cdot \|\mathbf{x}_t^\top \mathbf{w}_{t-1} \mathbf{x}_t - (\mathbf{x}_t^\top \mathbf{w}_{t-1})^2 \mathbf{w}_{t-1}\|^2 \\ &= \|\mathbf{w}_{t-1}\|_2^2 - 2\eta_t (\mathbf{x}_t^\top \mathbf{w}_{t-1})^2 \cdot (\|\mathbf{w}_{t-1}\|_2^2 - 1) + 2\eta_t^2 (\mathbf{x}_t^\top \mathbf{w}_{t-1})^2 \cdot \max\{\|\mathbf{x}_t\|_2^2, (\mathbf{x}_t^\top \mathbf{w}_{t-1})^2 \|\mathbf{w}_{t-1}\|_2^2\}.\end{aligned}$$

Consider two cases: (i) $\|\mathbf{w}_{t-1}\|_2^2 \leq 1 + 8\eta$ and (ii) $1 + 8\eta < \|\mathbf{w}_{t-1}\|_2^2 \leq 1 + 10\eta$. Note that $\|\mathbf{w}_t\|_2^2 \leq 1 + 10\eta$ in both cases. This completes the induction and the proof. $\square$

6 Phase 1: Toward Constant Error

In this section, we study the stopping time τ for the event $\{\mathbf{w}_{t,1}^2 \geq \frac{4}{9}\}$. We show that for any failure probability $\delta \in (0, 1)$, we have $\Pr[\tau > t_1] < \delta$ for some $t_1 = \tilde{O}(\lambda_1 \delta^{-2} \text{gap}^{-2})$. The following is the formal theorem statement.

Theorem 6.1. *For any $n \in \mathbb{N}$, $\delta \in (0, 1)$, $\mathbf{w}_0$ a random unit vector in $\mathbb{R}^n$, and*

$$\eta = O\left(\frac{\delta^2(\lambda_1 - \lambda_2)}{\lambda_1 \log^2 \frac{n}{\delta(\lambda_1 - \lambda_2)}}\right)$$

there exists

$$t_1 = O\left(\frac{\log \frac{n}{\delta}}{\eta(\lambda_1 - \lambda_2)}\right)$$

such that if for all $t \in \mathbb{N}$, $\eta_t = \eta$, then

$$\Pr[\tau > t_1] < \delta.$$

Specifically, the stopping time τ is of order $O\left(\frac{\lambda_1 \log \frac{n}{\delta} \log^2 \frac{n}{\delta(\lambda_1 - \lambda_2)}}{\delta^2(\lambda_1 - \lambda_2)^2}\right)$ with probability at least $1 - \delta$.

The proof of [Theorem 6.1](#) contains three steps. First, we linearize the biological Oja's rule and apply the ODE trick to get a closed-form expression for the updates in [Section 6.1](#). Next, in [Section 6.2](#) we use the corollary of Freedman's inequality (*i.e.*, [Corollary 2.7](#)) to show that the the stopped process of the noise term in this closed-form expression is small and thus the updates will be dominated by the drifting term. Finally, we wrap up the proof in [Section 6.3](#) using stopping time arguments.

Dictionary for parameters in Phase 1 Since the analysis is recursive, some of the parameters, *e.g.*, the failure probability δ , are scaled by $\delta/\log(n/\delta)$ in some of the lemmas. Also, we use T instead of t_1 in most of the lemmas because some parameters depend on T so in the end we will fix $T = t_1$ and show that all the conditions hold. In [Table 2](#), we provide a dictionary of parameters for the reader to recall from time to time during the reading.

	Failure probability	Time length	Stopping time	Goal
The whole analysis	δ	t_1 and T	τ and ξ	$\mathbf{w}_{t,1}^2$ from $\frac{\delta^2}{n}$ to $\frac{4}{9}$
An interval analysis	$\delta' = \frac{\delta}{10 \log(n/\delta)}$	$t'_1 = \frac{t_1}{10 \log(n/\delta)}$	τ' and τ_i	$\mathbf{w}_{t,1}^2$ from a to $4a$

Table 2: Some potentially confusing parameters in Phase 1.

6.1 Linearization and ODE trick centered at 0

In this subsection, we describe the linearization of biological Oja's rule and ODE trick with center at 0. The point here is to get a closed-form expression in Equation 6.5 where the initial point $\mathbf{w}_{t_0,1}$ is decoupled with the noise terms. Specifically, the noise terms are martingales (or submartingales) with well-controlled bounded difference and conditional variance.

Lemma 6.2 (Linearization in Phase 1). *For any $t \in \mathbb{N}$ and $\eta \in (0, 0.1)$, we have*

$$\mathbf{w}_{t,1} = H_t \cdot \mathbf{w}_{t-1,1} + \eta D_t + \eta F_t + \eta^2 Q_t$$

almost surely where H_t, D_t, F_t, Q_t are some random processes depend on $\mathbf{w}_{t-1}$ and $\mathbf{x}_t$. Specifically,

$$D_t = \sum_{i=2}^n \mathbf{x}_{t,1} \mathbf{x}_{t,i} \mathbf{w}_{t-1,i} \quad \text{and} \quad H_t = \exp(\eta(\lambda_1 - \lambda_2)(1 - \mathbf{w}_{t-1,1}^2) + \eta E_t).$$

Furthermore, the following hold.

- (Bounded difference) *For any $t \in \mathbb{N}$, $E_t \geq 0$, $|F_t| = O(\mathbf{w}_{t-1,1})$, and $|Q_t| = O(\mathbf{w}_{t-1,1})$ almost surely.*
- (Conditional expectation) *For any $t \in \mathbb{N}$, $\mathbb{E}[F_t \mid \mathcal{F}_{t-1}] = 0$ and $|\mathbb{E}[Q_t \mid \mathcal{F}_{t-1}]| = O(\lambda_1 \mathbf{w}_{t-1,1})$.*
- (Conditional variance) *For any $t \in \mathbb{N}$, $\text{Var}[F_t \mid \mathcal{F}_{t-1}] = O(\lambda_1 \mathbf{w}_{t-1,1}^2)$ and $\text{Var}[Q_t \mid \mathcal{F}_{t-1}] = O(\lambda_1 \mathbf{w}_{t-1,1}^2)$.*

Proof of Lemma 6.2. The proof is based on Taylor's expansion and Cauchy-Schwarz inequality. See Appendix E for the full proof. $\square$

It turns out that the cross term, *i.e.*, D_t , is the most annoying term and we need the following extra lemma to control its bounded difference as well as its conditional variance. The way we handle D_t is introducing another stopping time ξ and a time parameter T such that (i) the bounded difference and the conditional variance of the stopped process $D_{t \wedge \xi}$ is under control and (ii) $\xi \geq T$ with high probability. Intuitively, ξ is the stopping time for some “bad events” and we want the probability of the “bad events” happening before T to be small.

Throughout the proof, the reader should think of T as t_1 . We decide not to directly state the lemma in terms of t_1 because there are some recursive dependencies between T and other parameters (*e.g.*, η). In the end of the proof of Theorem 6.1, we will fix T to be t_1 and show that every conditions hold.

Lemma 6.3 (Bounds for D_t). *Let $\mathbf{w}_0$ be a random unit vector in $\mathbb{R}^n$. For any $T \in \mathbb{N}$, $\delta \in (0, 1)$, if $\eta = O\left(\frac{\delta^2(\lambda_1 - \lambda_2)}{\lambda_1 \log^2 \frac{nT}{\delta}}\right)$, then there exists a stopping time ξ such that the following hold.*

1.

$$\Pr[\xi \geq T] \geq 1 - \delta.$$

2. (Bounded difference)

$$\Pr\left[|D_{t \wedge \xi}| = O\left(\frac{\mathbf{w}_{(t \wedge \xi)-1,1}}{\delta} \sqrt{\log \frac{nT}{\delta}}\right)\right] = 1, \quad \forall t \in [T].$$

3. (Conditional variance)

$$\text{Var}[D_{t \wedge \xi} \mid \mathcal{F}_{t-1}] = O\left(\frac{\lambda_1 \mathbf{w}_{(t \wedge \xi)-1,1}^2}{\delta^2} \log \frac{nT}{\delta}\right), \quad \forall t \in [T].$$

Proof of Lemma 6.3. The proof of Lemma 6.3 improves on the Markov-type inequality on a multi-dimensional chain of [AZL17] by considering a different multi-dimensional chain and using a vector form of ODE trick to gain a tight exponential bound in one-shot. Since the basic strategy is similar to the techniques in phase 2, we postpone the proof to Appendix D. $\square$

Finally, apply the ODE trick (*i.e.*, Lemma 2.11) on Lemma 6.2, we get the following corollary.

Corollary 6.4 (ODE trick in Phase 1). *For any $t_0 \in \mathbb{N}_{\geq 0}$, $t \in \mathbb{N}$, and $\eta \in (0, 0.1)$, we have*

$$\mathbf{w}_{t_0+t,1} = \left(\prod_{i=t_0+1}^{t_0+t} H_i \right) \cdot \left(\mathbf{w}_{t_0,1} + \sum_{i=t_0+1}^{t_0+t} \frac{\eta D_i + \eta F_i + \eta^2 Q_i}{\prod_{j=t_0+1}^i H_j} \right). \quad (6.5)$$

6.2 Concentration of the noise terms

After using the ODE trick to rewrite the dynamic of biological Oja's rule, we would like to show that the second parenthesis of Equation 6.5 is dominated by $\mathbf{w}_{0,1}$. In that case, it is then not difficult to argue that $\mathbf{w}_{t,1}$ grows as fast as one can expect. Let us start with the following lemma analyzing the concentration of the noise terms (*i.e.*, the D_i , F_i , and Q_i term in the second parenthesis). Note that since we do not have good control on the bounded difference and conditional variance on the noise terms, we are going to consider the *stopped processes* of them.

Lemma 6.6 (Concentration of noise in an interval). *Let $\mathbf{w}_0$ be a random unit vector in $\mathbb{R}^n$. For any $T \in \mathbb{N}$, $\delta \in (0, 0.5)$, choose $\eta, \delta' \in (0, 0.5)$ such that $\delta' = \Omega(\delta/nT)$ and $\eta = O\left(\frac{\delta^2(\lambda_1 - \lambda_2)}{\lambda_1 \log^2 \frac{nT}{\delta}}\right)$. Let ξ be the stopping time from Lemma 6.3 with the corresponding parameters (T, η, δ) .*

For any $a \in (0, 1/9]$, and $t_0 \in \mathbb{N}$ such that $\mathbf{w}_{t_0,1}^2 \in [a, 4a)$, let τ' be the stopping time for the event $\{\mathbf{w}_{t_0+\tau',1}^2 \geq 4a\}$. For all $t \in \mathbb{N}$ such that $t_0 + t \leq T$, we have

1.

$$\Pr \left[\left| \sum_{i=t_0+1}^{t_0+(t \wedge \tau')} \frac{\eta^2 Q_i}{\prod_{j=t_0+1}^i H_j} \right| \geq \Omega \left(\frac{\eta \lambda_1 \sqrt{a}}{\lambda_1 - \lambda_2} + \eta \sqrt{\frac{\eta a \lambda_1 \log \frac{1}{\delta'}}{\lambda_1 - \lambda_2}} \right) \right] < \delta',$$

2.

$$\Pr \left[\left| \sum_{i=t_0+1}^{t_0+(t \wedge \tau')} \frac{\eta F_i}{\prod_{j=t_0+1}^i H_j} \right| \geq \Omega \left(\sqrt{\frac{\eta \lambda_1 a \log \frac{1}{\delta'}}{\lambda_1 - \lambda_2}} \right) \right] < \delta', \text{ and}$$

3.

$$\Pr \left[\left| \sum_{i=t_0+1}^{(t_0+(t \wedge \tau')) \wedge \xi} \frac{\eta D_i}{\prod_{j=t_0+1}^i H_j} \right| \geq \Omega \left(\frac{1}{\delta} \sqrt{\frac{\eta \lambda_1 a \log \frac{1}{\delta'} \log \frac{nT}{\delta}}{\lambda_1 - \lambda_2}} \right) \right] < \delta'.$$

We state Lemma 6.6 in a very general form because we are going to apply it for several times in the later interval analysis. Concretely, the final analysis will start from setting $a_1 = \mathbf{w}_{0,1}^2$ and look at the stopping time for the event $\{\mathbf{w}_{\tau_1,1}^2 \geq 4a_1\}$. We show that the concentration lemma (*i.e.*, Lemma 6.6) and the ODE trick (*i.e.*, Corollary 6.4) imply τ_1 being small with high probability. Then, we are going to set $a_2 = 4a_1$ and repeat the same argument until some a_ℓ reaches $1/9$. From Lemma 2.1, such ℓ could be $O(\log(n/\delta))$ with high probability. We will pick $\delta' = \Theta(\delta/\log(n/\delta))$ in the end and apply union bound.

To see why Lemma 6.6 would be helpful, the following lemma plugs in the parameters we are going to use later and show that the deviations in the concentration inequalities are small.

Lemma 6.7 (Parameters for Phase 1). *For any $n \in \mathbb{N}$ and $\delta \in (0, 1)$, let η and t_1 be the parameters chosen in Theorem 6.1. For any $\delta' = \Omega(\delta/nt_1)$, there exists $C > 0$ such that for any $t \geq \frac{C}{\eta(\lambda_1 - \lambda_2)}$ the following holds.*

- (Drifting term)

$$\exp \left(\frac{5}{9} \eta (\lambda_1 - \lambda_2) t \right) \geq 16.$$

- (Deviation terms) For any $a > 0$

$$O\left(\sqrt{\frac{\eta\lambda_1 a \log \frac{1}{\delta'}}{\lambda_1 - \lambda_2}}\right), O\left(\frac{\eta\lambda_1 \sqrt{a}}{\lambda_1 - \lambda_2} + \eta\sqrt{\frac{\eta\lambda_1 a \log \frac{1}{\delta'}}{\lambda_1 - \lambda_2}}\right), O\left(\frac{1}{\delta}\sqrt{\frac{\eta\lambda_1 a \log \frac{1}{\delta'} \log \frac{nt_1}{\delta}}{\lambda_1 - \lambda_2}}\right) \leq \frac{\sqrt{a}}{10},$$

where all the O terms are from the concentration inequalities in [Lemma 6.6](#).

Proof of Lemma 6.7. Both statements immediately follow from checking the parameters. $\square$

From [Lemma 6.7](#), one can see that the deviations in the concentration inequalities for the noise terms are quite small ($\leq \sqrt{a}/10$) comparing with $|\mathbf{w}_{t_0,1}|$ ($\geq \sqrt{a}$). The rest of this subsection is devoted to the proof of the concentration lemma. Reader should feel comfortable to skip it during the first reading.

Proof of the concentration lemma

Proof of Lemma 6.6. Recall that the definition of H_j is $H_j = \exp(\eta(\lambda_1 - \lambda_2)(1 - \mathbf{w}_{j-1,1}^2) + \eta E_j)$ and τ' is the stopping time for the event $\{\mathbf{w}_{t_0+\tau',1}^2 \geq 4a\}$.

Let us start with the following observation on the multiplier term H_j . When $j \leq t_0 + \tau'$, we have $H_j \geq \exp(\eta(\lambda_1 - \lambda_2)/2)$. The reason is that (i) if $j \leq t_0 + \tau'$, then $\mathbf{w}_{j-1,1}^2 \leq 4a \leq 4/9$ and (ii) $E_j \geq 0$ almost surely for all $j \in \mathbb{N}$ by [Lemma E.5](#). By analyzing the geometric sum, we have

$$\begin{aligned} \sum_{i=t_0+1}^{t_0+(t \wedge \tau')} \left| \frac{1}{\prod_{j=1}^i H_j} \right|^2 &\leq O\left(\frac{1 - \exp(-\eta(\lambda_1 - \lambda_2)t/2)}{1 - \exp(-\eta(\lambda_1 - \lambda_2)/2)}\right) \\ (\because \text{Lemma 2.16}) &= O\left(\frac{1 - \exp(-\eta(\lambda_1 - \lambda_2)t/2)}{\eta(\lambda_1 - \lambda_2)}\right) \\ &= O\left(\frac{1}{\eta(\lambda_1 - \lambda_2)}\right) \end{aligned} \tag{6.8}$$

almost surely.

1. Now, for the Q_i term, define the following random process with $M_0 = 0$ and

$$M_t = \sum_{i=t_0+1}^{t_0+t} \frac{\eta^2 Q_i}{\prod_{j=t_0+1}^i H_j}, \quad \forall t \geq 1.$$

Note that $\{M_t\}$ is an adapted random process and so is its stopped process $M_{t \wedge \tau'}$. Since the process stops when $\mathbf{w}_{t_0+(t \wedge \tau'),1}^2 \geq 4a$, by [Lemma 6.2](#) and [Equation 6.8](#), we can upper bound the bounded difference of $\{M_{t \wedge \tau'}\}$ by $O(\eta^2 \sqrt{a})$, upper bound the summation of the conditional expectation of bounded difference by $O\left(\frac{\eta^2 \lambda_1 \sqrt{a}}{\eta(\lambda_1 - \lambda_2)}\right)$ and upper bound the sum of the conditional variance by $O\left(\frac{\eta^4 \lambda_1 a}{\eta(\lambda_1 - \lambda_2)}\right)$. Now by the corollary of Freedman's inequality (i.e., [Corollary 2.7](#)), we have

$$\Pr \left[\left| \sum_{i=1}^{t \wedge \tau'} \frac{\eta^2 Q_i}{\prod_{j=1}^i H_j} \right| \geq \Omega \left(\frac{\eta \lambda_1 \sqrt{a}}{\lambda_1 - \lambda_2} + \eta \sqrt{\frac{\eta \lambda_1 a \log \frac{1}{\delta'}}{\lambda_1 - \lambda_2}} \right) \right] < \exp \left(-\Omega \left(\frac{\frac{\eta^3 \lambda_1 a \log \frac{1}{\delta'}}{\lambda_1 - \lambda_2}}{\frac{\eta^4 \lambda_1 a}{\eta(\lambda_1 - \lambda_2)} + \eta^2 a \sqrt{\frac{\eta^3 \lambda_1 a \log \frac{1}{\delta'}}{\lambda_1 - \lambda_2}}} \right) \right) < \delta'.$$

Note that the $\frac{\eta \lambda_1 \sqrt{a}}{\lambda_1 - \lambda_2}$ term in the deviation is due to the sum of conditional expectation and $\eta \sqrt{\frac{\eta \lambda_1 a \log(1/\delta')}{\lambda_1 - \lambda_2}}$ is due to the bounded difference and conditional variance.

2. Similarly, for the F_i term, define $M_0 = 0$ and

$$M_t = \sum_{i=t_0+1}^{t_0+t} \frac{\eta F_i}{\prod_{j=t_0+1}^i H_j}, \quad \forall t \geq 1.$$

Note that $\{M_t\}$ is a martingale and so is its stopped process $\{M_{t \wedge \tau'}\}$. Since the process stops when $\mathbf{w}_{t_0+(t \wedge \tau'),1}^2 \geq 4a$, we can upper bound the difference of $\{M_{t \wedge \tau'}\}$ by $O(\eta\sqrt{a})$ almost surely and by [Equation 6.8](#) upper bound the sum of the conditional variance by $O\left(\frac{\eta^2 \lambda_1 a}{\eta(\lambda_1 - \lambda_2)}\right)$. By Freedman's inequality (see [Lemma 2.6](#)), we have

$$\Pr \left[\left| \sum_{i=1}^{t \wedge \tau'} \frac{\eta F_i}{\prod_{j=1}^i H_j} \right| \geq \Omega \left(\sqrt{\frac{\eta \lambda_1 a \log \frac{1}{\delta'}}{\lambda_1 - \lambda_2}} \right) \right] < \exp \left(-\Omega \left(\frac{\frac{\eta \lambda_1 a \log \frac{1}{\delta'}}{\lambda_1 - \lambda_2}}{\frac{\eta^2 \lambda_1 a}{\eta(\lambda_1 - \lambda_2)} + \eta \sqrt{a} \sqrt{\frac{\eta \lambda_1 a \log \frac{1}{\delta'}}{\lambda_1 - \lambda_2}}} \right) \right) < \delta'.$$

3. Similarly, for the D_i term, define $M_0 = 0$ and

$$M_t = \sum_{i=t_0+1}^{t_0+t} \frac{\eta D_i}{\prod_{j=t_0+1}^i H_j}, \quad \forall t \text{ s.t. } 1 \leq t \leq T - t_0.$$

Note that $\{M_t\}$ is a martingale and so is its stopped process $\{M_{t \wedge \tau' \wedge (\xi - t_0)}\}$. By [Lemma 6.3](#), [Equation 6.8](#), and the definition of τ' , we can upper bound the difference of $\{M_{t \wedge \tau' \wedge (\xi - t_0)}\}$ by $O\left(\eta \frac{\sqrt{a \log(nT/\delta)}}{\delta}\right)$ almost surely and upper bound the sum of the conditional variance by $O\left(\frac{\eta^2 \lambda_1 a \log(nT/\delta)}{\delta^2 \eta(\lambda_1 - \lambda_2)}\right)$. By Freedman's inequality (see [Lemma 2.6](#)), we have

$$\begin{aligned} \Pr \left[\left| \sum_{i=t_0+1}^{(t_0+(t \wedge \tau')) \wedge \xi} \frac{\eta D_i}{\prod_{j=t_0+1}^i H_j} \right| \geq \Omega \left(\frac{1}{\delta} \sqrt{\frac{\eta \lambda_1 a \log \frac{1}{\delta'} \log \frac{nT}{\delta}}{\lambda_1 - \lambda_2}} \right) \right] \\ < \exp \left(-\Omega \left(\frac{\frac{\eta \lambda_1 a \log \frac{1}{\delta'} \log \frac{nT}{\delta}}{\delta^2 (\lambda_1 - \lambda_2)}}{\frac{\eta^2 \lambda_1 a \log(nT/\delta)}{\delta^2 \eta(\lambda_1 - \lambda_2)} + \frac{\eta \sqrt{a \log(nT/\delta)}}{\delta} \sqrt{\frac{\eta \lambda_1 a \log \frac{1}{\delta'} \log \frac{nT}{\delta}}{\delta^2 (\lambda_1 - \lambda_2)}}} \right) \right) < \delta'. \end{aligned}$$

□

6.3 Wrap up

To prove [Theorem 6.1](#), we do an *interval analysis* by looking at the stopping time of the form $\{\mathbf{w}_{t_0+\tau}^2 \geq 4a\}$ with a geometrically growing from $\mathbf{w}_{0,1}^2$ to $4/9$. The following lemma summarizes how the analysis works in a single interval. We will prove [Theorem 6.1](#) after stating this lemma and provide the proof for the lemma in the rest of the section.

Lemma 6.9 (Interval analysis in Phase 1). *For any $n, T \in \mathbb{N}$ and $\delta \in (0, 0.5)$, choose any $\delta' \in (0, 0.5)$ such that $\delta' = \Omega(\delta/nT)$. For any*

$$\eta = O\left(\frac{\delta^2(\lambda_1 - \lambda_2)}{\lambda_1 \log^2 \frac{nT}{\delta}}\right)$$

there exists

$$t'_1 = \Theta\left(\frac{1}{\eta(\lambda_1 - \lambda_2)}\right)$$

such that the following holds. For any $a \in (0, 1/9]$ and $t_0 + t'_1 \leq T$ such that $\mathbf{w}_{t_0,1}^2 \in [a, 4a)$, let τ' be the stopping time for the event $\{\mathbf{w}_{t_0+\tau'}^2 \geq 4a\}$. Then, we have

$$\Pr[\tau' > t'_1] < \delta'.$$

Proof of the main theorem in Phase 1

Proof of Theorem 6.1. Now, the proof of Theorem 6.1 follows from recursively invoking Lemma 6.9 for $O(\log \frac{n}{\delta})$ times. First, we need to set the parameter T . Let $T = t_1 = 10 \log \frac{n}{\delta} t'_1$. Then for

$$\eta = O\left(\frac{\delta^2(\lambda_1 - \lambda_2)}{\lambda_1 \log^2 \frac{n}{\delta(\lambda_1 - \lambda_2)}}\right),$$

it satisfies the condition of Lemma 6.9.

Recall that τ is the stopping time for the event $\{\mathbf{w}_{\tau,1}^2 \geq 4/9\}$. Let us first invoke Lemma 6.9 with $a_1 = \mathbf{w}_{0,1}^2$ and failure probability $\delta' = \frac{\delta}{10 \log(n/\delta)}$. Let τ_1 be the stopping time for $\{\mathbf{w}_{\tau_1,1}^2 \geq 4a_1\}$, this gives

$$\Pr[\tau_1 > t'_1] < \frac{\delta}{10 \log \frac{n}{\delta}}.$$

Next, for each $i \in \mathbb{N}$, let $a_{i+1} = 4a_i$ if $4a_i < 1/9$ and let $a_{i+1} = 1/9$ otherwise. Also, define τ_{i+1} inductively to be the stopping time for the event $\{\mathbf{w}_{\tau_1+\dots+\tau_i+\tau_{i+1},1}^2 \geq 4a_{i+1}\}$. By invoking Lemma 6.9 with a_{i+1} , τ_{i+1} , and failure probability $\delta' = \frac{\delta}{10 \log(n/\delta)}$, we have

$$\Pr[\tau_{i+1} > t'_1] < \frac{\delta}{10 \log \frac{n}{\delta}}.$$

Now, by Lemma 2.1, we have $a_1 = \mathbf{w}_{0,1}^2 = \delta^2/100n$ with probability at least $1 - \delta/10$. Let ℓ be the index such that $a_\ell = \frac{1}{9}$, we have $\ell \leq 9 \log(n/\delta)$ with probability at least $1 - \delta/10$. Finally, by union bound, we have

$$\begin{aligned} \Pr[\tau > t_1] &\leq \Pr\left[\tau > 9 \log \frac{n}{\delta} t'_1\right] \leq \Pr\left[\ell > 9 \log \frac{n}{\delta} \vee \sum_{i=1}^{\ell} \tau_i > \ell \cdot t'_1\right] \\ &= \Pr\left[\ell > 9 \log \frac{n}{\delta}\right] + \Pr\left[\ell \leq 9 \log \frac{n}{\delta} \wedge \sum_{i=1}^{\ell} \tau_i > \ell \cdot t'_1\right] \\ &< \frac{\delta}{10} + 9 \log \frac{n}{\delta} \cdot \frac{\delta}{10 \log \frac{n}{\delta}} = \delta. \end{aligned}$$

This completes the proof of Theorem 6.1. □

Proof of the interval analysis

Proof of Lemma 6.9. Denote the failure event $\{\tau' > t'_1\}$ as BAD, the goal is to show that $\Pr[\text{BAD}] < \delta'$. The high-level strategy here is to use the concentration of (the stopped process of) the noise terms and ODE trick to guarantee $\mathbf{w}_{t,1}$ is going to follow the drifting term with high probability. To formalize this idea, we have to get rid of the stopping time in Lemma 6.6. Nevertheless, it turns out that in Phase 1 we do not need pull-out lemma to achieve this.

First, denote the good event where the noise terms in the ODE trick are small as follows.

$$\text{GOOD} := \left\{ \left| \sum_{i=t_0+1}^{t_0+(t'_1 \wedge \tau')} \frac{\eta F_i}{\prod_{j=t_0+1}^i H_j} \right|, \left| \sum_{i=t_0+1}^{t_0+(t'_1 \wedge \tau')} \frac{\eta^2 Q_i}{\prod_{j=t_0+1}^i H_j} \right|, \left| \sum_{i=t_0+1}^{(t_0+(t'_1 \wedge \tau')) \wedge \xi} \frac{\eta D_i}{\prod_{j=t_0+1}^i H_j} \right| \leq \frac{\sqrt{a}}{10} \right\}.$$

Note that with the parameter settings, from Lemma 6.7, the good events are exactly the events from the concentration inequalities in Lemma 6.6. Thus, we have $\Pr[\text{GOOD}] \geq 1 - \delta/2$. Also, by Lemma 6.3, we have $\Pr[\{\xi \geq T\}] \geq 1 - \delta/2$. The following claim shows that the events GOOD, BAD, and $\{\xi \geq T\}$ cannot happen simultaneously due to ODE trick.

Claim 6.10. $\Pr[\text{GOOD} \wedge \text{BAD} \wedge \{\xi \geq T\}] = 0$. Namely, $\Pr[\text{BAD}] \leq 1 - \Pr[\text{GOOD} \wedge \{\xi \geq T\}]$.

Proof of Claim 6.10. For the sake of contradiction, assume GOOD, BAD, and $\{\xi \geq T\}$ happen simultaneously. Note that $t'_1 \wedge \tau' = t'_1$ and $(t_0 + (t'_1 \wedge \tau')) \wedge \xi = t'_1$ under this assumption. Intuitively, the stopping times were *pulled out* in this situation.

Now, by the ODE trick (*i.e.*, Corollary 6.4), we have

$$\mathbf{w}_{t_0+t'_1,1} = \left(\prod_{i=t_0+1}^{t_0+t'_1} H_i \right) \cdot \left(\mathbf{w}_{t_0,1} + \sum_{i=t_0+1}^{t_0+t'_1} \frac{\eta D_i + \eta F_i + \eta^2 Q_i}{\prod_{j=t_0+1}^i H_j} \right).$$

Also, because $i \leq \tau'$ for all $i \in [t'_1]$, we have $\mathbf{w}_{t_0+i,1}^2 < 4a \leq 4/9$ and thus $H_i \geq \exp\left(\frac{5}{9}\eta(\lambda_1 - \lambda_2)\right)$. By the choice of parameters (*i.e.*, Lemma 6.7), we have

$$\prod_{i=t_0+1}^{t_0+t'_1} H_i \geq \exp\left(\frac{5}{9}\eta(\lambda_1 - \lambda_2)t'_1\right) \geq 16.$$

Further, the events GOOD implies

$$\left| \sum_{i=t_0+1}^{t_0+t'_1} \frac{\eta D_i + \eta F_i + \eta^2 Q_i}{\prod_{j=t_0+1}^i H_j} \right| \leq \frac{\sqrt{a}}{2}.$$

Combine the above three, we have

$$\mathbf{w}_{t_0+t'_1,1}^2 \geq \left(\prod_{i=t_0+1}^{t_0+t'_1} H_i \right)^2 \cdot \left(\mathbf{w}_{t_0,1} - \frac{\sqrt{a}}{2} \right)^2 \geq 16 \cdot \frac{a}{4} \geq 4a$$

which implies $\tau' \leq t'_1$ and contradicts to the happening of BAD. Thus, we conclude that the events GOOD, BAD, and $\{\xi \geq T\}$ cannot happen simultaneously and this completes the proof for the claim. $\square$

Finally, Lemma 6.6 immediately follows from Claim 6.10. Concretely,

$$\begin{aligned} \Pr[\tau' > t'_1] &= \Pr[\text{BAD}] \leq 1 - \Pr[\text{GOOD} \wedge \{\xi \geq T\}] \\ &\leq \Pr[\neg \text{GOOD}] + \Pr[\neg \{\xi \geq T\}] < \frac{\delta}{2} + \frac{\delta}{2} = \delta. \end{aligned}$$

This completes the proof of Lemma 6.6. $\square$

7 Phase 2: Toward Arbitrarily Small Error

Suppose $\mathbf{w}_{t_0,1} \geq 2/3$ for some $t_0 \in \mathbb{N}$. For any $\epsilon \in (0, 1)$, the goal of this section is to show that $1 - \mathbf{w}_{t_0+t,1}^2 \leq \epsilon$ for some $t = O\left(\frac{\lambda_1 \log(1/\epsilon)(\log \log \log(1/\epsilon) + \log(1/\delta))}{\epsilon(\lambda_1 - \lambda_2)^2}\right)$. Note that analogous result that starts from $\mathbf{w}_{t_0,1} \leq -2/3$ is an immediate corollary. Thus, combining with the previous section we would be able to complete the analysis of biological Oja's rule. Let us first state the main theorem in this section as follows.

Theorem 7.1. *For some $t_0 \in \mathbb{N}$, suppose $\mathbf{w}_{t_0,1} \geq 2/3$. For any $n \in \mathbb{N}$, $\epsilon, \delta \in (0, 1)$, there exists*

$$\eta = O\left(\frac{\epsilon(\lambda_1 - \lambda_2)}{\lambda_1 \log \frac{\log \log \frac{1}{\epsilon}}{\delta}}\right), \quad t_2 = O\left(\frac{\log \frac{1}{\epsilon}}{\eta(\lambda_1 - \lambda_2)}\right)$$

such that

$$\Pr[\mathbf{w}_{t_0+t_2,1} < 1 - \epsilon] < \delta.$$

Namely, the convergence rate is of order $t_0 + O\left(\frac{\lambda_1 \log \frac{1}{\epsilon}(\log \log \log \frac{1}{\epsilon} + \log \frac{1}{\delta})}{\epsilon(\lambda_1 - \lambda_2)^2}\right)$ with probability at least $1 - \delta$.

The proof structure of [Theorem 7.1](#) is the same as that of the main theorem in Phase 1. However, each step requires some non-trivial modifications. First, we derive a different linearization and ODE trick in [Section 7.1](#) using center at 1 instead of 0. Next, in [Section 7.2](#) we state the lemma showing that the noise terms in the ODE trick are small and further use the lemma to wrap up the proof for the main theorem in [Section 7.3](#). Nevertheless, it turns out that the noise terms in Phase 2 are more difficult to handle. We develop a *pull-out lemma* to get rid of the stopping time in the concentration inequality. The details of how we handle the noise is provided in [Section 7.4](#).

Dictionary for parameters in Phase 2 Similarly, the analysis in Phase 2 is recursive and thus some of the parameters, *e.g.*, the failure probability δ , are scaled by $\delta/\log \log(1/\epsilon)$ in some of the lemmas. We provide a dictionary of parameters for the reader to recall from time to time during the reading.

	Failure probability	Time length	Error	Goal
The whole analysis	δ	$t_2 = \sum_{k=1}^{\ell} t_{2,k}$	ϵ	$\mathbf{w}_{t,1}$ from $\frac{2}{3}$ to $1 - \epsilon$
An interval analysis	$\delta' = \frac{\delta}{\ell}$	$t_{2,k}$	$\epsilon' = \frac{\epsilon}{8}$	$\mathbf{w}_{t,1}$ from $1 - (\epsilon')^{1-2^{-(k-1)}}$ to $1 - (\epsilon')^{1-2^{-k}}$

Table 3: Some potentially confusing parameters in Phase 2. The parameter $\ell = \lceil \log \log(1/\epsilon) \rceil$ is the number of intervals we are going to use in the whole analysis.

7.1 Linearization and ODE trick centered at 1

In the analysis of Phase 2, we use a different linearization for the biological Oja's rule. The main difference from the linearization in Phase 1 is that here we do Taylor expansion with center at 1 instead of 0. The idea is inspired from the analysis of the continuous dynamics as explained in [Section 3](#).

To ease the notation, we define $\mathbf{w}'_{t,1} = \mathbf{w}_{t,1} - 1$ and the goal becomes showing that $\mathbf{w}'_{t_0+t_2,1} > -\epsilon$ with probability at least $1 - \delta$. The following lemma states the linearization for $\mathbf{w}'_{t,1}$.

Lemma 7.2 (Linearization in Phase 2). *For any $t \in \mathbb{N}$ and $\eta \in (0, 0.1)$, we have*

$$\mathbf{w}'_{t,1} = H \cdot \mathbf{w}'_{t-1,1} + \eta F_t + \eta P_t + \eta^2 Q_t$$

almost surely, where the multiplier is

$$H = \exp\left(-\frac{3}{4}\eta(\lambda_1 - \lambda_2)\right).$$

Moreover, F_t, P_t, Q_t satisfy the following properties.

- (Bounded difference) *For any $t \in \mathbb{N}$, $|F_t| = O(\sqrt{\mathbf{w}'_{t-1} + \eta})$, and $|Q_t| = O(1)$ almost surely. If $\mathbf{w}_{t-1,1} \geq 0.5$, then $P_t \geq -O(\eta\lambda_1)$ almost surely.*
- (Conditional expectation) *For any $t \in \mathbb{N}$, $\mathbb{E}[F_t \mid \mathcal{F}_{t-1}] = 0$ and $|\mathbb{E}[Q_t \mid \mathcal{F}_{t-1}]| = O(\lambda_1)$.*
- (Conditional variance) *For any $t \in \mathbb{N}$, $\text{Var}[F_t \mid \mathcal{F}_{t-1}] = O(\lambda_1(\mathbf{w}'_{t-1} + \eta))$ and $\text{Var}[Q_t \mid \mathcal{F}_{t-1}] = O(\lambda_1)$.*

Proof of Lemma 7.2. The proof is based on Taylor's expansion and Cauchy-Schwarz inequality. See [Appendix E](#) for the full proof. $\square$

Similarly, we apply the ODE trick (see [Lemma 2.11](#)) on [Lemma 7.2](#) and get the corollary as follows. Note that one major difference between the ODE trick in the two phases is that here the multiplier H is a constant and does not depend on $\mathbf{w}_t$.

Corollary 7.3 (ODE trick in Phase 2). *For any $t_0 \in \mathbb{N}_{\geq 0}$, $t \in \mathbb{N}$, and $\eta \in (0, 0.1)$, we have*

$$\mathbf{w}'_{t_0+t,1} = H^t \cdot \left(\mathbf{w}'_{t_0,1} + \sum_{i=t_0+1}^{t_0+t} \frac{\eta F_i + \eta P_i + \eta^2 Q_i}{H^{i-t_0}} \right).$$

7.2 Concentration of noise

The following key lemma shows the concentration of noise in Phase 2. Similarly, since we are doing an interval analysis so the lemma is stated in a general form.

Lemma 7.4 (Concentration of noise in an interval). *For any $\epsilon', \delta', \eta \in (0, 0.5)$ and $k \in \mathbb{N}_{\geq 0}$ such that $\eta = O\left(\frac{\epsilon'(\lambda_1 - \lambda_2)}{\lambda_1 \log \frac{1}{\delta'}}\right)$, let $t_{2,k} \in \mathbb{N}$ be*

$$t_{2,k} = \frac{4}{3} \cdot \frac{2^{-k} \log \frac{1}{\epsilon'}}{\eta(\lambda_1 - \lambda_2)} + \Theta\left(\frac{2^{-k}}{\eta(\lambda_1 - \lambda_2)}\right).$$

Suppose there exists some $t_0 \in \mathbb{N}$ and $\gamma \in (0, 0.5)$ where $\gamma = \Omega((\epsilon')^{1-2^{-k}})$ such that $\mathbf{w}_{t_0,1} \geq 1 - \frac{\gamma}{2}$. Then,

$$H^{t_{2,k}} = \exp\left(-\frac{3}{4}\eta(\lambda_1 - \lambda_2)t_{2,k}\right) \leq \frac{(\epsilon')^{2^{-k}}}{10}$$

and

$$\Pr\left[\min_{1 \leq t \leq t_{2,k}} \sum_{i=t_0+1}^{t_0+t} \frac{\eta F_i + \eta P_i + \eta^2 Q_i}{H^{i-t_0}} < -\frac{\gamma}{2}\right] < \delta'.$$

The proof of Lemma 7.4 involves stopping time analysis and maximal martingale inequality and is provided in Section 7.4. Let us first take a look at how to use Lemma 7.4 to prove the main theorem in Phase 2.

7.3 Wrap up

Now we are ready to prove the main theorem of this section.

Proof of Theorem 7.1. Recall that we have $\mathbf{w}_{\tau,1} \geq 2/3$ and the goal is to show that $\mathbf{w}_{\tau+t_2,1} \geq 1 - \epsilon$ with probability at least $1 - \delta$ for some $t_2 = O\left(\frac{\log(1/\epsilon)}{\eta(\lambda_1 - \lambda_2)}\right)$.

We use an interval analysis to achieve the goal. Specifically, there are going to be $\ell = \lceil \log \log(1/\epsilon) \rceil$ stages of the analysis. Let $\epsilon' = \epsilon/8$ and $\delta' = \delta/\ell$ be the error and failure probability parameters in each stage. For each $k \in [\ell]$, in the k^{th} stage, we are going to show that $\mathbf{w}_{t,1}$ improve from $1 - (\epsilon')^{1-2^{-(k-1)}}$ to $1 - (\epsilon')^{1-2^{-k}}$ in $t_{2,k}$ steps with probability at least $1 - \delta'$. The following is the formal statement of the guarantee we want to get for each stage in the interval analysis.

Lemma 7.5 (Interval analysis in Phase 2). *Let $n, \epsilon, \delta, \eta$ be the parameters from Theorem 7.1 and ϵ', δ' be the parameters as defined above. For each $k \in [\ell]$, let $t_{2,k} = \frac{4}{3} \cdot \frac{2^{-k} \log \frac{1}{\epsilon'}}{\eta(\lambda_1 - \lambda_2)} + \Theta\left(\frac{2^{-k}}{\eta(\lambda_1 - \lambda_2)}\right)$. Then we have*

$$\Pr\left[\mathbf{w}_{\tau + \sum_{j=1}^k t_{2,j},1} < 1 - (\epsilon')^{1-2^{-k}}\right] < k \cdot \delta'.$$

With the lemma for each interval, we are ready to prove Theorem 7.1. Recall that $\epsilon' = \epsilon/8$ and thus by Lemma 2.17, we have

$$(\epsilon')^{1-2^{-\ell}} \leq \left(\frac{\epsilon}{8}\right)^{1 - \frac{1}{\log(1/\epsilon)}} \leq \epsilon.$$

Pick $t_2 = \sum_{j=1}^{\ell} t_{2,j}$, we have

$$t_2 = \sum_{j=1}^{\ell} \frac{4}{3} \cdot \frac{2^{-j} \log \frac{1}{\epsilon'}}{\eta(\lambda_1 - \lambda_2)} + \Theta\left(\frac{2^{-j}}{\eta(\lambda_1 - \lambda_2)}\right) = O\left(\frac{\log \frac{1}{\epsilon}}{\eta(\lambda_1 - \lambda_2)}\right)$$

as desired. Finally, recall that $\delta' = \delta/\ell$ and thus Lemma 7.5 gives

$$\Pr[\mathbf{w}_{\tau+t_2,1} > 1 - \epsilon] \leq \Pr[\mathbf{w}_{\tau+t_2,1} > 1 - (\epsilon')^{1-2^{-\ell}}] < \ell \cdot \delta' = \delta.$$

This completes the proof of Theorem 7.1. The rest of this subsection is devoted to the proof of Lemma 7.5 on the interval analysis. $\square$

Proof of the interval analysis

Proof of Lemma 7.5. The proof is based on induction on $k \in [\ell]_{\geq 0}$. For the base case where $k = 0$, we have $\mathbf{w}_{\tau,1} \geq 2/3$ by the definition of τ .

For the induction step, for any $k \in [\ell]$, let GOOD_k denote the event $\left\{ \mathbf{w}_{\tau + \sum_{j=1}^{k-1} t_{2,j},1} \geq 1 - (\epsilon')^{1-2^{-(k-1)}} \right\}$. The induction hypothesis gives

$$\Pr[\neg \text{GOOD}_k] \leq (k-1) \cdot \delta'.$$

Now, invoke Lemma 7.4 with $t_0 = \tau + \sum_{j=1}^{k-1} t_{2,j}$, $\gamma = 2(\epsilon')^{1-2^{-(k-1)}}$, and ϵ', δ' as chosen before. We have

$$\Pr \left[\min_{1 \leq t' \leq t_{2,k}} \sum_{i=t_0+1}^{t_0+t'} \frac{\eta F_i + \eta P_i + \eta^2 Q_i}{H^{i-t_0}} < -\frac{\gamma}{10} \mid \text{GOOD}_k \right] < \delta'$$

and

$$\Pr \left[H^{t_{2,k}} \leq \frac{(\epsilon')^{2^{-k}}}{10} \mid \text{GOOD}_k \right] = 1.$$

Thus, by the ODE trick (*i.e.*, Corollary 7.3), we have

$$\mathbf{w}'_{t_0+t_{2,k},1} = H^{t_{2,k}} \cdot \left(\mathbf{w}'_{t_0,1} + \sum_{i=t_0+1}^{t_0+t_{2,k}} \frac{\eta F_i + \eta P_i + \eta^2 Q_i}{H^{i-t_0}} \right).$$

Condition on the event GOOD_k , we have

$$\geq \frac{(\epsilon')^{2^{-k}}}{10} \cdot \left(-\frac{\gamma}{2} - \frac{\gamma}{10} \right) \geq -(\epsilon')^{2^{-k}} \cdot (\epsilon')^{1-2^{-(k-1)}} \geq -(\epsilon')^{1-2^{-k}}$$

happens with probability at least $1 - \delta'$. Namely,

$$\begin{aligned} \Pr \left[\mathbf{w}_{\tau + \sum_{j=1}^i t_{2,k},1} < 1 - (\epsilon')^{1-2^{-k}} \right] &= \Pr \left[\mathbf{w}'_{\tau + \sum_{j=1}^i t_{2,k},1} < -(\epsilon')^{1-2^{-k}} \right] \\ &\leq \Pr \left[\mathbf{w}'_{\tau + \sum_{j=1}^i t_{2,k},1} < -(\epsilon')^{1-2^{-k}} \mid \text{GOOD}_k \right] + \Pr[\neg \text{GOOD}_k] \\ &< \delta' + (k-1) \cdot \delta' = k \cdot \delta'. \end{aligned}$$

This completes the induction and the proof for the interval analysis. □

7.4 Proof of Lemma 7.4

Now, the only missing part in this section is the proof of Lemma 7.4. Recall that the goal is to show that the noise term in the ODE trick (Corollary 7.3) is small with high probability.

High-level idea on the proof of Lemma 7.4 The most natural way to prove such statement is using martingale concentration inequality as the noise term is a martingale by construction. However, the difficulty here is that $\mathbf{w}_{t,1}$ might *go back* to the small region (*e.g.*, $\mathbf{w}_{t,1} < 1/2$) and thus the bounded difference might be large and ruin the condition of Freedman's inequality. Note that this “*going back phenomenon*” is the major difference between Phase 1 and Phase 2 and thus the simple proof of the former cannot be applied to the later. Nevertheless, the continuous dynamic (see Section 3) suggests that this situation should happen with small probability. To enforce the analysis, we consider a *stopped process* where the dynamic stops once $\mathbf{w}_{t,1} < 1/2$. Concretely, let t_0 be the time when we enter Phase 2, *i.e.*, $\mathbf{w}_{t_0,1} \geq 2/3$, and let τ be the stopping time for the event $\{\mathbf{w}_{t_0+\tau,1} < 1/2\}$. Denote the noise term in Corollary 7.3 as $\{M_{t_0+t'}\}_{t' \in \mathbb{N}}$, which is a martingale by construction. Consider its stopped process $\{M_{t_0+(t' \wedge \tau)}\}_{t' \in \mathbb{N}}$, which is also a martingale due to

the property of stopping time. Moreover, the stopped process satisfies good bounded difference condition by its construction and thus Freedman's inequality works well. See [Lemma 7.6](#) for a formal and more general statement on the above intuition.

After obtaining a good control on the noise term in the stopped process, we would like to remove the stopping time and get back to the original non-stopped process in order to prove [Lemma 7.4](#). This can be done by [Lemma 7.9](#) which *pulls out* the stopping time from the concentration inequality for the stopped process. In general, pulling out the stopping time is impossible, however, there are some structures in the martingale we are looking at. Intuitively, suppose $\tau \geq t$ for some t , then this would imply that all the noise terms before time $t_0 + t$ are small with high probability (using maximal martingale inequality). Next, the noise being small at time $t_0 + t$ would further imply that $\tau \geq t + 1$ (using ODE trick). This would form a chain of implications as pictured in [Figure 3](#).

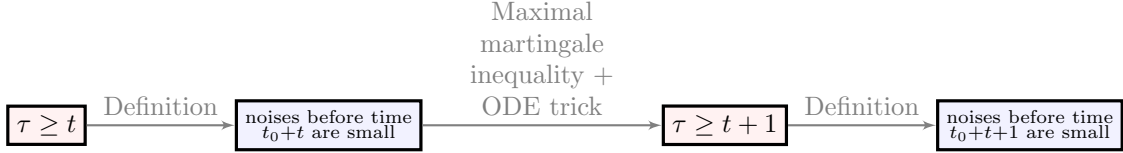


Figure 3: Intuition on why it is possible to pull out stopping time in Phase 2.

With the above *chain* structure in the noise terms, we are then able to pull out the stopping time in [Lemma 7.6](#) by introducing another stopping time to help us properly partition the probability space. The rest of this subsection is devoted to formalize the above intuition and complete the proof for [Lemma 7.4](#).

Noise concentration for the stopped process The first step towards proving [Lemma 7.4](#) is showing the concentration of the noise terms in the stopped process.

Lemma 7.6 (Bounds for the noise terms in the stopped process in Phase 2). *For any $t \in \mathbb{N}$, $\eta, \epsilon, \delta \in (0, 1)$, $t_0 \in \mathbb{N}$ and $\gamma \in (0, 0.5)$ such that $\eta = O\left(\frac{\lambda_1 - \lambda_2}{\lambda_1 \log(1/\delta)}\right)$, define stopping time τ to be the first time of the event*

$$\{\mathbf{w}_{t_0+\tau,1} < 1 - \gamma\}.$$

Then, we have

1.

$$\Pr \left[\min_{1 \leq t' \leq t} \sum_{i=t_0+1}^{t_0+(t' \wedge \tau)} \frac{\eta P_i}{H^{i-t_0}} \geq -O \left(\exp \left(\frac{3}{4} \eta (\lambda_1 - \lambda_2) t \right) \frac{\eta \lambda_1}{\lambda_1 - \lambda_2} \right) \right] = 1,$$

2.

$$\Pr \left[\max_{1 \leq t' \leq t} \left| \sum_{i=t_0+1}^{t_0+(t' \wedge \tau)} \frac{\eta F_i}{H^{i-t_0}} \right| \geq \Omega \left(\exp \left(\frac{3}{4} \eta (\lambda_1 - \lambda_2) t \right) \sqrt{\frac{\eta(\gamma + \eta) \lambda_1}{\lambda_1 - \lambda_2} \log \frac{1}{\delta}} \right) \right] < \delta, \text{ and}$$

3.

$$\Pr \left[\max_{1 \leq t' \leq t} \left| \sum_{i=t_0+1}^{t_0+(t' \wedge \tau)} \frac{\eta^2 Q_i}{H^{i-t_0}} \right| \geq \Omega \left(\exp \left(\frac{3}{4} \eta (\lambda_1 - \lambda_2) t \right) \left(\sqrt{\frac{\eta^3 \lambda_1}{\lambda_1 - \lambda_2} \log \frac{1}{\delta}} + \frac{\eta \lambda_1}{\lambda_1 - \lambda_2} \right) \right) \right] < \delta.$$

To see why [Lemma 7.6](#) would be helpful, the following lemma plugs in the parameters we are going to use later and show that the deviations in the concentration inequalities are small.

Lemma 7.7. For any $k \in \mathbb{N}$, let $\epsilon', \delta', \gamma, \eta, t_{2,k}$ be the parameters from Lemma 7.4. Then we have

1.

$$O\left(\exp\left(\frac{3}{4}\eta(\lambda_1 - \lambda_2)t_{2,k}\right) \frac{\eta\lambda_1}{\lambda_1 - \lambda_2}\right) \leq \frac{\gamma}{10},$$

2.

$$O\left(\exp\left(\frac{3}{4}\eta(\lambda_1 - \lambda_2)t_{2,k}\right) \sqrt{\frac{\eta(\gamma + \eta)\lambda_1}{\lambda_1 - \lambda_2} \log \frac{1}{\delta}}\right) \leq \frac{\gamma}{10},$$

3.

$$O\left(\exp\left(\frac{3}{4}\eta(\lambda_1 - \lambda_2)t_{2,k}\right) \left(\sqrt{\frac{\eta^3\lambda_1}{\lambda_1 - \lambda_2} \log \frac{1}{\delta}} + \frac{\eta\lambda_1}{\lambda_1 - \lambda_2}\right)\right) \leq \frac{\gamma}{10}, \text{ and}$$

4.

$$H^{t_{2,k}} = \exp\left(-\frac{3}{4}\eta(\lambda_1 - \lambda_2)t_{2,k}\right) \leq (\epsilon')^{2^{-k}}.$$

All the O terms are from the concentration inequalities of Lemma 7.4.

Proof of Lemma 7.7. These items directly follow from the choice of parameters $\epsilon', \delta', \gamma, \eta, t_{2,k}$. $\square$

Proof of Lemma 7.6. The proof is based on using the properties of these martingales (see Lemma 7.2) and maximal martingale concentration inequality (e.g., Freedman's inequality as stated in Lemma 2.6). First, notice that by $\eta(\lambda_1 - \lambda_2) = O(1)$, for any $1 \leq t' \leq t$, we have

$$\sum_{i=t_0+1}^{t_0+t'} \frac{1}{H^{i-t_0}} = O\left(\frac{\exp(\frac{3}{4}\eta(\lambda_1 - \lambda_2)t) - 1}{\exp(\frac{3}{4}\eta(\lambda_1 - \lambda_2)) - 1}\right) \stackrel{\text{Lemma 2.15}}{=} O\left(\frac{\exp(\frac{3}{4}\eta(\lambda_1 - \lambda_2)t)}{\eta(\lambda_1 - \lambda_2)}\right). \quad (7.8)$$

1. From Lemma 7.2, we know that $P_{t_0+t'} \geq -O(\eta\lambda_1)$ for all $1 \leq t' \leq \tau$. Combine with Equation 7.8, for any $1 \leq t' \leq t$, we have

$$\sum_{i=t_0+1}^{t_0+(t' \wedge \tau)} \frac{\eta P_i}{H^{i-t_0}} \geq -O\left(\frac{\eta\lambda_1 \exp(\frac{3}{4}\eta(\lambda_1 - \lambda_2)t)}{\lambda_1 - \lambda_2}\right)$$

almost surely.

2. Let $M_0 = 0$ and for each $1 \leq t' \leq t$, let $M_{t'} = \sum_{i=t_0+1}^{t_0+t'} \frac{\eta F_i}{H^{i-t_0}}$. Note that both $\{M_{t'}\}$ and $\{M_{t' \wedge \tau}\}$ form martingales. Also, recall that $H = \exp(-3\eta(\lambda_1 - \lambda_2)/4)$. Moreover, by Lemma 7.2, the bounded difference of $\{M_{t' \wedge \tau}\}$ is $O(\eta\sqrt{\gamma + \eta} \exp(\frac{3}{4}\eta(\lambda_1 - \lambda_2)t))$ and the sum of the conditional variance of $\{M_{t' \wedge \tau}\}$ can be upper bounded by $O\left(\frac{\eta^2(\gamma + \eta)\lambda_1 \exp(\frac{3}{4}\eta(\lambda_1 - \lambda_2)t)}{\eta(\lambda_1 - \lambda_2)}\right)$. Now, by the Freedman's inequality (see Lemma 2.6), we have

$$\begin{aligned} & \Pr \left[\max_{1 \leq t' \leq t} \left| \sum_{i=t_0+1}^{t_0+(t' \wedge \tau)} \frac{\eta F_i}{H^{i-t_0}} \right| \geq \Omega \left(\exp\left(\frac{3}{4}\eta(\lambda_1 - \lambda_2)t\right) \sqrt{\frac{\eta(\gamma + \eta)\lambda_1}{\lambda_1 - \lambda_2} \log \frac{1}{\delta}} \right) \right] \\ & < \exp \left(\frac{\frac{\eta(\gamma + \eta)\lambda_1 \exp(\frac{3}{4}\eta(\lambda_1 - \lambda_2)t)}{\lambda_1 - \lambda_2} \log \frac{1}{\delta}}{\frac{\eta(\gamma + \eta)\lambda_1 \exp(\frac{3}{4}\eta(\lambda_1 - \lambda_2)t)}{\lambda_1 - \lambda_2} + \eta\sqrt{\gamma + \eta} \exp(\frac{3}{4}\eta(\lambda_1 - \lambda_2)t) \sqrt{\frac{\eta(\gamma + \eta)\lambda_1 \exp(\frac{3}{4}\eta(\lambda_1 - \lambda_2)t)}{\lambda_1 - \lambda_2} \log \frac{1}{\delta}}} \right) < \delta. \end{aligned}$$

The last inequality holds because $\eta = O(\frac{\lambda_1 - \lambda_2}{\lambda_1 \log(1/\delta)})$ and thus the first term in the denominator dominates the second term.

3. Let $M_0 = 0$ and for each $1 \leq t' \leq t$, let $M_{t'} = \sum_{i=t_0+1}^{t_0+t'} \frac{\eta^2 Q_i}{H^{i-t_0}}$. By [Lemma 7.2](#), the bounded difference of $\{M_{t' \wedge \tau}\}$ is $O(\eta^2 \exp(\frac{3}{4}\eta(\lambda_1 - \lambda_2)t))$, the sum of the absolute value of the conditional expectation is bounded by $O\left(\frac{\eta^2 \lambda_1 \exp(\frac{3}{4}\eta(\lambda_1 - \lambda_2)t)}{\eta(\lambda_1 - \lambda_2)}\right)$ and the sum of conditional variance of can be upper bounded by $O\left(\frac{\eta^4 \lambda_1 \exp(\frac{3}{2}\eta(\lambda_1 - \lambda_2)t)}{\eta(\lambda_1 - \lambda_2)}\right)$. Now, by the corollary of Freedman's inequality (i.e., [Corollary 2.7](#)), we have

$$\Pr \left[\max_{1 \leq t' \leq t} \left| \sum_{i=t_0+1}^{t_0+(t' \wedge \tau)} \frac{\eta^2 Q_i}{H^{i-t_0}} \right| \geq \Omega \left(\exp \left(\frac{3}{4}\eta(\lambda_1 - \lambda_2)t \right) \left(\frac{\eta \lambda_1}{\lambda_1 - \lambda_2} + \sqrt{\frac{\eta^3 \lambda_1}{\lambda_1 - \lambda_2} \log \frac{1}{\delta}} \right) \right) \right] \\ < \exp \left(\frac{\frac{\eta^3 \lambda_1 \exp(\frac{3}{2}\eta(\lambda_1 - \lambda_2)t)}{\lambda_1 - \lambda_2} \log \frac{1}{\delta}}{\frac{\eta^4 \lambda_1 \exp(\frac{3}{2}\eta(\lambda_1 - \lambda_2)t)}{\eta(\lambda_1 - \lambda_2)} + \eta^2 \exp \left(\frac{3}{4}\eta(\lambda_1 - \lambda_2)t \right) \sqrt{\frac{\eta^3 \lambda_1 \exp(\frac{3}{2}\eta(\lambda_1 - \lambda_2)t)}{\lambda_1 - \lambda_2} \log \frac{1}{\delta}}} \right) < \delta.$$

Similarly, the last inequality holds because $\eta = O\left(\frac{\lambda_1 - \lambda_2}{\lambda_1 \log(1/\delta)}\right)$ and thus the first term in the denominator dominates the second term. □

Pulling out the stopping time In the second step towards proving [Lemma 7.4](#), we are going to pull out the stopping time τ in [Lemma 7.6](#). The following lemma shows that under certain *chain* condition, it is possible to pull out the stopping time.

Lemma 7.9. *Let $\{M_t\}_{t \in \mathbb{N}_{\geq 0}}$ be an adapted stochastic process and τ be a stopping time. Let $\{M_t^*\}_{t \in \mathbb{N}_{\geq 0}}$ be the maximal process of $\{M_t\}_{t \in \mathbb{N}_{\geq 0}}$ where $M_t^* = \max_{1 \leq t' \leq t} M_{t'}$. For any $t \in \mathbb{N}$, $a \in \mathbb{R}$, and $\delta \in (0, 1)$, suppose*

1. $\Pr[M_{t \wedge \tau}^* \geq a] < \delta$ and
2. $\Pr[\tau \geq t' + 1 \mid M_{t'}^* < a] = 1$ for any $1 \leq t' < t$.

Then, we have

$$\Pr[M_t^* \geq a] < \delta.$$

Proof of Lemma 7.9. The key idea is to introduce another stopping time which helps us partition the probability space. Let τ' be the stopping time for the event $\{M_{t \wedge \tau}^* \geq a\}$. The following claim shows that if τ stopped before time t , then τ' should stop earlier than τ .

Claim 7.10. *Let τ and τ' be stopping times as defined above. Suppose the conditions in [Lemma 7.9](#) hold. Then we have*

$$\Pr[\tau < t, \tau' > \tau] = 0.$$

Proof of Claim 7.10. The claim can be proved by contradiction as follows. Suppose both $\tau < t$ and $\tau' > \tau$. By the definition of τ' , we know that $M_\tau^* < a$ since $\tau < \tau'$. However, by the second condition of the lemma, we then have

$$\Pr[\tau \geq \tau + 1 \mid M_\tau^* < a] = 1$$

which is a contradiction. □

Next, we would like to show that $\Pr[M_t^* \geq a] \leq \Pr[M_{t \wedge \tau}^* \geq a]$. The idea is partitioning the probability space as follows.

$$\Pr[M_t^* \geq a] = \Pr[M_t^* \geq a, \tau \geq t] + \Pr[M_t^* \geq a, \tau < t, \tau' \leq \tau] + \Pr[M_t^* \geq a, \tau < t, \tau' > \tau]. \quad (7.11)$$

For the first term of [Equation 7.11](#), when $\tau \geq t$, we have $t = t \wedge \tau$ and thus $M_t^* = M_{t \wedge \tau}^*$. As for the second term, when $\tau' \leq \tau < t$, we have both $M_t^*, M_{t \wedge \tau}^* \geq a$. Thus, [Equation 7.11](#) becomes

$$= \Pr[M_{t \wedge \tau}^* \geq a, \tau \geq t] + \Pr[M_{t \wedge \tau}^* \geq a, \tau < t, \tau' \leq \tau] + \Pr[M_t^* \geq a, \tau < t, \tau' > \tau].$$

Combining the first two terms, we have

$$\leq \Pr[M_{t \wedge \tau}^* \geq a] + \Pr[M_t^*, \tau < t, \tau' > \tau].$$

Finally, from [Claim 7.10](#), we know that $\Pr[\tau < t, \tau' > \tau] = 0$. Thus, we conclude that $\Pr[M_t^* \geq a] \leq \Pr[M_{t \wedge \tau}^* \geq a] < \delta$ as desired. $\square$

Wrap up the proof of Lemma 7.4 Finally, we are ready to prove [Lemma 7.4](#) by using [Lemma 7.6](#) which bounds the the stopped process of the noise terms and [Lemma 7.9](#) which pulls out the stopping time from the concentration inequalities. Let us restate the lemma as follows for the completeness.

Lemma 7.4 (Concentration of noise in an interval). *For any $\epsilon', \delta', \eta \in (0, 0.5)$ and $k \in \mathbb{N}_{\geq 0}$ such that $\eta = O\left(\frac{\epsilon'(\lambda_1 - \lambda_2)}{\lambda_1 \log \frac{1}{\delta'}}\right)$, let $t_{2,k} \in \mathbb{N}$ be*

$$t_{2,k} = \frac{4}{3} \cdot \frac{2^{-k} \log \frac{1}{\epsilon'}}{\eta(\lambda_1 - \lambda_2)} + \Theta\left(\frac{2^{-k}}{\eta(\lambda_1 - \lambda_2)}\right).$$

Suppose there exists some $t_0 \in \mathbb{N}$ and $\gamma \in (0, 0.5)$ where $\gamma = \Omega((\epsilon')^{1-2^{-k}})$ such that $\mathbf{w}_{t_0,1} \geq 1 - \frac{\gamma}{2}$. Then,

$$H^{t_{2,k}} = \exp\left(-\frac{3}{4}\eta(\lambda_1 - \lambda_2)t_{2,k}\right) \leq \frac{(\epsilon')^{2^{-k}}}{10}$$

and

$$\Pr\left[\min_{1 \leq t \leq t_{2,k}} \sum_{i=t_0+1}^{t_0+t} \frac{\eta F_i + \eta P_i + \eta^2 Q_i}{H^{i-t_0}} < -\frac{\gamma}{2}\right] < \delta'.$$

Proof of Lemma 7.4. The first statement of the lemma directly follows from the forth item of [Lemma 7.7](#). For the second statement of the lemma, the proof is based on [Lemma 7.9](#) which pulls out the stopping time.

Let τ be the stopping time for the event $\{\mathbf{w}_{t_0+\tau,1} < 1 - \gamma\}$ as defined in [Lemma 7.6](#). Also, define

$$M_t = - \sum_{i=t_0+1}^{t_0+t} \frac{\eta F_i + \eta P_i + \eta^2 Q_i}{H^{i-t_0}}$$

for any $t \in \mathbb{N}$. Now by [Lemma 7.6](#) and [Lemma 7.7](#), we have

$$\Pr\left[\max_{1 \leq t \leq t_{2,k}} M_{t \wedge \tau} > \frac{\gamma}{2}\right] < \delta. \quad (7.12)$$

Now in order to apply [Lemma 7.9](#) to pull out the stopping time, we need to check the following.

$$\Pr\left[\tau \geq j+1 \mid \max_{1 \leq t \leq j} M_t \leq \frac{\gamma}{2}\right] = 1 \quad (7.13)$$

for any $1 \leq j < t_{2,k}$. By the ODE trick (*i.e.*, [Corollary 7.3](#)) we have

$$\mathbf{w}'_{t_0+j,1} = H^j \cdot \left(\mathbf{w}'_{t_0,1} + \sum_{i=t_0+1}^{t_0+j} \frac{\eta F_i + \eta P_i + \eta^2 Q_i}{H^{i-t_0}}\right).$$

By the condition in [Equation 7.13](#) and $\mathbf{w}'_{t_0,1} \geq -\frac{\gamma}{2}$, the ODE trick gives us

$$\geq H^j \cdot \left(-\frac{\gamma}{2} - \frac{\gamma}{2}\right) \geq -\gamma.$$

Note that the above implies that $\mathbf{w}_{t_0+j,1} \geq 1 - \gamma$ which means that $\tau \geq j+1$ as desired. This shows that [Equation 7.13](#) holds.

Finally, invoke the pull-out lemma (*i.e.*, [Lemma 7.9](#)) with [Equation 7.12](#) and [Equation 7.13](#), we have the desiring bound as follows.

$$\Pr\left[\min_{1 \leq t \leq t_{2,k}} \sum_{i=t_0+1}^{t_0+t} \frac{\eta F_i + \eta P_i + \eta^2 Q_i}{H^{i-t_0}} < -\frac{\gamma}{2}\right] = \Pr\left[\max_{1 \leq t \leq t_{2,k}} M_t > \frac{\gamma}{2}\right] \leq \Pr\left[\max_{1 \leq t \leq t_{2,k}} M_{t \wedge \tau} > \frac{\gamma}{2}\right] < \delta.$$

This completes the proof of [Lemma 7.4](#). $\square$

8 Conclusions and Future Directions

In this work, our contributions are two-fold. First, we give the first convergence rate analysis for the biological Oja’s rule in solving streaming PCA. In particular, the rate we show is nearly optimal. This provides theoretical evidences for a biologically-plausible streaming PCA mechanism to converge in a biologically-realistic time scale in the retina-optical nerve pathway. Second, we introduce a novel one-shot framework to analyze stochastic dynamics. The framework is simple and captures the intrinsic behaviors of the dynamics. Finally, as a byproduct, the convergence rate we get for the biological Oja’s rule outperforms the state-of-the-art upper bound for streaming PCA (using ML Oja’s rule).

In this section, we discuss the biological and algorithmic significance of our results and point out potential future directions.

8.1 Biological aspect

Spiking Oja’s Rule In this paper, we simplify the biological dynamic using a rate-based model. It would be interesting to design a spiking version of the learning rule to solve streaming PCA. On the other hand, it has been shown that Spike Timing Dependent Plasticity (STDP) has self-normalizing behaviors [AN00], so the higher order terms in the biological Oja’s rule might not be needed for the normalization in the spiking version.

Convergence rate analysis for other biological-plausible learning rules As mentioned in Section 1.4, there are plenty of Hebbian-type learning rules that had been proposed to solve some computational problems [Sej77, BCM82, San89, XOS92, OBL00, Apa12, PHC15]. Nevertheless, most of them do not have efficiency guarantee and we think it would be of interest to use our frameworks to systematically analyze the convergence rates of these update rules. This is not only a natural theoretical question but also could potentially provide insights on how these biologically-plausible algorithms are different from standard algorithms.

Designing biologically-plausible learning rule for online k -PCA In this work, we focus on the biological Oja’s rule in finding the top eigenvector of the covariance matrix. It is a natural question to ask: whether there is a biologically-plausible algorithm for finding top k eigenvectors (a.k.a. the k -PCA problem)? In the setting of ML Oja’s rule, this can be achieved by *QR decomposition* [AZL17] while it is unclear how to implement such decomposition in a biologically-plausible way. Another standard way to solve k -PCA is performing the *Gram-Schmidt process* after having a good enough approximation to the top eigenvectors. Nevertheless, it is also unclear how to implement Gram-Schmidt using a simple plasticity update rule. Thus, in our opinion, it would be very exciting to have a variant of biological Oja’s rule that solves the streaming k -PCA problem.

8.2 Algorithmic aspect

Improving the guarantees for biological Oja’s rule In this paper, we mainly focus on the situation when $\lambda_1 > \lambda_2$ while some of the previous works also considered the gap-free setting. We believe our framework can be easily extended to the gap-free setting and leave it as a future work. Also, there are some logarithmic terms (e.q. additive $\log \log \log(1/\epsilon)$ in the local convergence) in the convergence rate and do not seem to be inherent. It would be interesting to find out the optimal logarithmic dependency.

On the other hand, we suspect the $\log(1/\epsilon)$ term in the convergence rate of biological Oja’s rule might be necessary. Thus, showing a lower bound with $\log(1/\epsilon)$ would be of great interest. Note that there exists (non-streaming) algorithm which solves PCA using only $O(\lambda_1 \epsilon^{-1} \text{gap}^{-2})$ samples so the lower bound should be tailored to the dynamic.

Tighter analysis for ML Oja’s rule Using the objective function from [AZL17], one can also easily generalize our framework to ML Oja’s rule and tighten the bounds for both the local and global convergence rates.

Other Stochastic Dynamics There are many stochastic optimization problems in machine learning where the optimal analysis still remains elusive, *e.g.*, stochastic gradient dynamics of matrix completion, low-rank approximation, nonnegative matrix factorization, etc. It is of great interest to apply our *one-shot* framework to analyze other important stochastic dynamics.

Acknowledgements

We thank Nancy Lynch for valuable comments on the presentation of this paper, thank Kai-Min Chung for helpful discussions, and thank Rohit Agrawal, Boaz Barak, Chun-Hsiang Chan, Yanlin Chen, and Santhoshini Velusamy for comments on the draft of this paper. We thank Nancy Lynch again for organizing a brain algorithm reading group at MIT in Spring 2019 and all the participants for the inspiring conversation.

A Oja’s derivation for the biological Oja’s rule

Recall that Oja wanted to use the following normalized update rule to solve the streaming PCA problem.

$$\mathbf{w}_t = \frac{(I + \eta_t \mathbf{x}_t \mathbf{x}_t^\top) \mathbf{w}_{t-1}}{\|(I + \eta_t \mathbf{x}_t \mathbf{x}_t^\top) \mathbf{w}_{t-1}\|_2}. \quad (\text{A.1})$$

Oja applied *Taylor’s expansion* on the normalization term and truncated the higher-order term of η_t . Concretely, we have

$$\begin{aligned} \|(I + \eta_t \mathbf{x}_t \mathbf{x}_t^\top) \mathbf{w}_{t-1}\|_2^{-1} &= \left(\sum_{i=1}^n (\mathbf{w}_{t-1,i} + \eta_t y_t \mathbf{x}_{t,i})^2 \right)^{-1/2} \\ &= \left(\sum_{i=1}^n \mathbf{w}_{t-1,i}^2 + 2\eta_t y_t \mathbf{x}_{t,i} \mathbf{w}_{t-1,i} + O(\eta_t^2) \right)^{-1/2}. \end{aligned}$$

As $y_t = \mathbf{x}_t^\top \mathbf{w}_{t-1}$ and $\|\mathbf{w}_{t-1}\|_2$ is expected to be 1, the equation approximately becomes

$$= (1 + 2\eta_t y_t^2 + O(\eta_t^2))^{-1/2} = 1 - \eta_t y_t^2 + O(\eta_t^2). \quad (\text{A.2})$$

Replace the denominator of Equation A.1 with Equation A.2 and truncate the $O(\eta_t^2)$ term, one gets exactly Equation 1.4.

B Details of the Linearizations in Continuous Oja’s Rule

Recall that the dynamic of the continuous Oja’s rule is the following.

$$\frac{d\mathbf{w}_t}{dt} = \text{diag}(\lambda) \mathbf{w}_t - \mathbf{w}_t^\top \text{diag}(\lambda) \mathbf{w}_t \mathbf{w}_t.$$

Before proving the two convergence theorems of continuous Oja’s rule using different linearizations, let us first prove the following lemma on some basic properties.

Lemma B.1 (Properties of continuous Oja’s rule). *Let $\mathbf{w}_0 \in \mathbb{R}^n$ such that $\|\mathbf{w}_0\|_2 = 1$ and $\mathbf{w}_{0,1} > 0$. For any $t \geq 0$, we have*

1. $\|\mathbf{w}_t\|_2 = 1$,
2. $\frac{d\mathbf{w}_{t,1}}{dt} \geq (\lambda_1 - \lambda_2) \mathbf{w}_{t,1} (1 - \mathbf{w}_{t,1}^2)$, and
3. $\mathbf{w}_{t,1}$ is non-decreasing

almost surely.

Proof of Lemma B.1. In the following, everything holds almost surely so we would not mention this condition every time. First, consider

$$\begin{aligned}\frac{d\|\mathbf{w}_t\|_2^2}{dt} &= 2\mathbf{w}_t^\top \frac{d\mathbf{w}_t}{dt} = 2\mathbf{w}_t^\top (\text{diag}(\lambda)\mathbf{w}_t - \mathbf{w}_t^\top \text{diag}(\lambda)\mathbf{w}_t\mathbf{w}_t) \\ &= 2\mathbf{w}_t^\top \text{diag}(\lambda)\mathbf{w}_t \cdot (1 - \|\mathbf{w}_t\|_2^2) .\end{aligned}$$

As $1 - \|\mathbf{w}_0\|_2^2 = 0$, by induction, we have $\|\mathbf{w}_t\|_2 = 1$ for all $t \geq 0$.

For the second item of the lemma, we have

$$\begin{aligned}\frac{d\mathbf{w}_{t,1}}{dt} &= \left(\lambda_1 - \left(\sum_{i \in [n]} \lambda_i \mathbf{w}_{t,i}^2 \right) \right) \mathbf{w}_{t,1} \geq (\lambda_1 - \lambda_2) \mathbf{w}_{t,1} (1 - \mathbf{w}_{t,1}^2) \\ &= \lambda_1 (\mathbf{w}_{t,1} - \mathbf{w}_{t,1}^3) - \sum_{i=2}^n \lambda_i \mathbf{w}_{t,i}^2 \mathbf{w}_{t,1} .\end{aligned}$$

From the first item, we have $\sum_{i=2}^n \mathbf{w}_{t,i}^2 = 1 - \mathbf{w}_{t,1}^2$. Thus, we have

$$\geq \lambda_1 (\mathbf{w}_{t,1} - \mathbf{w}_{t,1}^3) - \lambda_2 (1 - \mathbf{w}_{t,1}^2) \mathbf{w}_{t,1} = (\lambda_1 - \lambda_2) \mathbf{w}_{t,1} (1 - \mathbf{w}_{t,1}^2) .$$

The last item of the lemma is then an immediate corollary of the first two items. $\square$

Now, we restate and prove Theorem 3.4 as follows.

Theorem 3.4 (Linearization at 0). *Suppose $\mathbf{w}_{0,1} > 0$. For any $\epsilon \in (0, 1)$, when $t \geq \Omega\left(\frac{\log(1/\mathbf{w}_{0,1}^2)}{\epsilon(\lambda_1 - \lambda_2)}\right)$, we have $\mathbf{w}_{t,1}^2 > 1 - \epsilon$.*

Proof of Theorem 3.4. Observe that for any $t \geq 0$ such that $\mathbf{w}_{t,1}^2 \leq 1 - \epsilon$, by the second item of Lemma B.1, we have

$$\frac{d\mathbf{w}_{t,1}}{dt} \geq (\lambda_1 - \lambda_2) \mathbf{w}_{t,1} (1 - \mathbf{w}_{t,1}^2) \geq \epsilon(\lambda_1 - \lambda_2) \mathbf{w}_{t,1} .$$

Let $\tau = \frac{10 \log(1/\mathbf{w}_{0,1}^2)}{\epsilon(\lambda_1 - \lambda_2)}$ and assume $\mathbf{w}_{\tau,1}^2 \leq 1 - \epsilon$ for the sake of contradiction. From the above linearization and $\mathbf{w}_{t,1}$ being non-decreasing (the third item of Lemma B.1), we have

$$\mathbf{w}_{\tau,1} \geq e^{\epsilon(\lambda_1 - \lambda_2)\tau} \cdot \mathbf{w}_{0,1} > 1 ,$$

which is a contradiction to the first item of Lemma B.1. Thus, we conclude that for any $t = \Omega\left(\frac{\log(1/\mathbf{w}_{0,1}^2)}{\epsilon(\lambda_1 - \lambda_2)}\right)$, $\mathbf{w}_{t,1}^2 > 1 - \epsilon$. $\square$

Now, we restate and prove Theorem 3.5 as follows.

Theorem 3.5 (Linearization at 1). *Suppose $\mathbf{w}_{0,1} > 0$. For any $\epsilon \in (0, 1)$, when $t \geq \Omega\left(\frac{\log(1/\epsilon)}{\mathbf{w}_{0,1}(\lambda_1 - \lambda_2)}\right)$, we have $\mathbf{w}_{t,1}^2 > 1 - \epsilon$.*

Proof of Theorem 3.5. Observe that for any $t \geq 0$, by the second item of Lemma B.1, we have

$$\begin{aligned}\frac{d(\mathbf{w}_{t,1} - 1)}{dt} &\geq (\lambda_1 - \lambda_2) \mathbf{w}_{t,1} (1 - \mathbf{w}_{t,1}^2) \\ &= -(\lambda_1 - \lambda_2) (\mathbf{w}_{t,1} - 1) (\mathbf{w}_{t,1} + \mathbf{w}_{t,1}^2) .\end{aligned}$$

As $\mathbf{w}_{t,1}$ is non-decreasing (the third item of Lemma B.1) and at most 1, we have

$$\geq -(\lambda_1 - \lambda_2) \mathbf{w}_{0,1} (\mathbf{w}_{t,1} - 1) .$$

By solving the linear ODE, we have

$$\mathbf{w}_{t,1} - 1 \geq (\mathbf{w}_{0,1} - 1) \cdot e^{-(\lambda_1 - \lambda_2) \mathbf{w}_{0,1} t} .$$

Thus, for any $t \geq \Omega\left(\frac{\log(1/\epsilon)}{\mathbf{w}_{0,1}(\lambda_1 - \lambda_2)}\right)$, we have $\mathbf{w}_{t,1}^2 > 1 - \epsilon$. $\square$

C Why the Analysis of ML Oja's Rule Cannot be Applied to Biological Oja's Rule?

In this section, we would like to discuss what makes biological Oja's rule much harder to analyze comparing to the previous approaches for ML Oja's rule. We study this problem through the lens of their corresponding continuous dynamics. Observe that, to study ML Oja's rule, it suffices to study the following dynamic

$$\frac{d\mathbf{w}_t}{dt} = \text{diag}(\lambda)\mathbf{w}_t.$$

The dynamic of the objective function $\sum_{i=2}^n \mathbf{w}_{t,i}^2 / \mathbf{w}_{t,1}^2$ would be

$$\begin{aligned} \frac{d \frac{\sum_{i=2}^n \mathbf{w}_{t,i}^2}{\mathbf{w}_{t,1}^2}}{dt} &= \frac{-2 \sum_{i=2}^n \mathbf{w}_{t,i}^2}{\mathbf{w}_{t,1}^3} \lambda_1 \mathbf{w}_{t,1} + \sum_{i=2}^n \frac{2 \mathbf{w}_{t,i}}{\mathbf{w}_{t,1}^2} \lambda_i \mathbf{w}_{t,i} \\ &\leq -2(\lambda_1 - \lambda_2) \frac{\sum_{i=2}^n \mathbf{w}_{t,i}^2}{\mathbf{w}_{t,1}^2}. \end{aligned}$$

Namely, the continuous dynamic is just a linear ODE with *slope* being independent to the value of $\mathbf{w}_t$. In comparison, the dynamic of the biological Oja's rule is the following.

$$\frac{d\mathbf{w}_{t,1}}{dt} \geq (\lambda_1 - \lambda_2) \mathbf{w}_{t,1} (1 - \mathbf{w}_{t,1}^2)$$

where you have to use at least two objective functions with different linearizations to get tight analysis. Furthermore, for any linearization, there exist some values of $\mathbf{w}_t$ that make the improvement extremely small or even vanishing. It is also not obvious to choose which two objective functions to analyze unless you are guided by the continuous dynamics.

We remark that the discussion here only suggests the difficulty of applying previous techniques of ML Oja's rule to biological Oja's rule. It might still be the case that the two dynamics are coupled but we argue here that even this is the case, previous techniques cannot show this.

D Proof of Lemma 6.3

Let us start with the definition of the central object of this section.

Definition D.1. For each $j \in [n]$, $t \in [T]$, and $\mathbf{w} \in \mathbb{R}^n$, define

$$f_{t,j}(\mathbf{w}) = \frac{\sum_{i=2}^j \mathbf{x}_{t,i} \mathbf{w}_i}{\mathbf{w}_1}.$$

The goal of this section is to prove the following lemma which implies Lemma 6.3.

Lemma D.2. Let $\mathbf{w}_0$ be a random unit vector in $\mathbb{R}^n$. For any $n, T \in \mathbb{N}$ and $\delta \in (0, 1)$ and $\eta = O\left(\frac{\delta^2(\lambda_1 - \lambda_2)}{\lambda_1 \log^2 \frac{nT}{\delta}}\right)$. There exists a stopping time ξ such that the following hold.

1.

$$\Pr[\xi \geq T] \geq 1 - \delta,$$

2.

$$\Pr \left[f_{t \wedge \xi, n}(\mathbf{w}_{t \wedge \xi - 1}) = O \left(\frac{1}{\delta} \sqrt{\log \frac{nT}{\delta}} \right) \right] = 1$$

for $t \in [T]$.

Note that Lemma 6.3 would be an immediate corollary of Lemma D.2.

Proof of Lemma 6.3. Pick exactly the same ξ as in Lemma D.2 and the first item holds directly. As for the second item, we have

$$\begin{aligned} |D_{t \wedge \xi}| &= \left| \mathbf{x}_{t \wedge \xi, 1} \sum_{i=2}^n \mathbf{x}_{t \wedge \xi, i} \mathbf{w}_{t \wedge \xi - 1, i} \right| \\ &\leq \left| \sum_{i=2}^n \mathbf{x}_{t \wedge \xi, i} \mathbf{w}_{t \wedge \xi - 1, i} \right| = |f_{t \wedge \xi, n}(\mathbf{w}_{t \wedge \xi - 1}) \cdot \mathbf{w}_{t \wedge \xi - 1, 1}|. \end{aligned}$$

By the second item in Lemma D.2, we have

$$= O\left(\frac{\mathbf{w}_{t \wedge \xi - 1, 1}}{\delta} \sqrt{\log \frac{nT}{\delta}}\right)$$

almost surely. As for the third item, we have

$$\begin{aligned} \text{Var}[D_{t \wedge \xi} \mid \mathcal{F}_{t-1}] &\leq \mathbb{E} \left[\mathbf{x}_{t \wedge \xi, 1}^2 \left(\sum_{i=2}^n \mathbf{x}_{t \wedge \xi, i} \mathbf{w}_{t \wedge \xi - 1, i} \right)^2 \mid \mathcal{F}_{t-1} \right] \\ &= \mathbb{E} \left[\mathbf{x}_{t \wedge \xi, 1}^2 f_{t \wedge \xi, n}(\mathbf{w}_{t \wedge \xi - 1})^2 \cdot \mathbf{w}_{t \wedge \xi - 1, 1}^2 \mid \mathcal{F}_{t-1} \right]. \end{aligned}$$

By the second item in Lemma D.2, we have

$$\begin{aligned} &\leq \mathbb{E} \left[\mathbf{x}_{t \wedge \xi, 1}^2 \mathbf{w}_{t \wedge \xi - 1, 1}^2 \mid \mathcal{F}_{t-1} \right] \cdot O\left(\frac{1}{\delta^2} \log \frac{nT}{\delta}\right) \\ &= O\left(\frac{\lambda_1 \mathbf{w}_{t \wedge \xi - 1, 1}^2}{\delta^2} \log \frac{nT}{\delta}\right). \end{aligned}$$

□

Organization of this section The goal of the rest of this section is to prove Lemma D.2. In Section D.1, we first use an initialization lemma from [AZL17] and design the stopping time ξ . Specifically, we show in Lemma D.4 that proving Equation D.5 would be sufficient for the goal. Next, we introduce the linearization and ODE trick for handling $f_{t,j}(\mathbf{w}_s)$ in Section D.2. In particular, we use a *vector form* of the ODE trick to get tight characterization. Then, we are able to show the concentration of the stopped process in Section D.3 and wrap up the proof in Section D.4.

D.1 Initialization

Before describing the stopping time ξ we are going to use, let us first introduce an useful initialization lemma (due to [AZL17]) which shows that some nice initial conditions hold with good probability (over the randomness of $\mathbf{w}_0$). We then *embed* the nice initial conditions into the stopping time ξ to enable future analysis.

Lemma D.3 (Initialization lemma in [AZL17, Lemma 5.1]). *For any $n, T \in \mathbb{N}$, and $\mathcal{D}$ a distribution over unit vectors in $\mathbb{R}^n$. Let $\mathbf{w}_0 \in \mathbb{R}^n$ be a random unit vector, then for any $j \in [n]$ and $p, \delta' \in (0, 1)$, we have*

$$\frac{\sum_{i=2}^n \mathbf{w}_{0,i}^2}{\mathbf{w}_{0,1}^2} \leq O\left(\frac{n}{p^2} \log \frac{n}{p}\right)$$

and

$$\Pr_{\mathbf{x}_1, \dots, \mathbf{x}_T \sim \mathcal{D}} \left[\exists j \in [n], t \in [T], |f_{t,j}(\mathbf{w}_0)| > \Omega\left(\frac{1}{p} \sqrt{\log \frac{nT}{\delta'}}\right) \right] < \delta'$$

with probability at least $1 - p - \delta'$ where the randomness is over $\mathbf{w}_0$.

Fixing parameters There are plenty of parameters here and might be confusing. So let us fix them before moving on. Recall that $n, T \in \mathbb{N}$, $\delta \in (0, 1)$, and $\eta = O\left(\frac{\delta^2(\lambda_1 - \lambda_2)}{\lambda_1 \log^2 \frac{nT}{\delta}}\right)$ are given and we are going to specify the deviation Λ and another auxiliary failure probability δ' . First, take $\delta' = \frac{\delta}{32n^2T}$ and $p = \delta/4$ and let the deviation inside the [Lemma D.3](#) be $\frac{\Lambda}{10} = O\left(\frac{1}{p} \sqrt{\log \frac{nT}{\delta'}}\right) = O\left(\frac{1}{\delta} \sqrt{\log \frac{nT}{\delta}}\right)$.

The choice of stopping time Now, we are able to formally define the stopping time ξ as follows. First, consider the initialization lemma (*i.e.*, [Lemma D.3](#)) with $p = \delta/4$ and $\delta' \ll \delta$ chosen later. If $\mathbf{w}_0$ does not satisfy the conditions in [Lemma D.3](#), then we define $\xi = 0$. Note that $\{\xi = 0\}$ happens with probability at most $p + \delta' \leq \delta/2$. Next, if $\mathbf{w}_0$ satisfies the conditions, then define ξ to be the first time t such that $|f_{t,n}(\mathbf{w}_{t-1})| > \Lambda$. Observe that the event $\{\xi = 0\}$ only depends on the randomness from $\mathbf{w}_0$ (not on that from $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T$). Thus, ξ is well-defined with respect to the natural filtration.

Note that ξ satisfies the second item of [Lemma D.2](#) immediately from its definition. To achieve the first condition and complete the proof, the following lemma shows that it suffices to show the concentration for a bunch of events described in [Equation D.5](#).

Lemma D.4. *For any $n, T \in \mathbb{N}$ and $\delta \in (0, 1)$. Suppose for any $t \in [T]$ and $j \in [n]$, we have*

$$\Pr[|f_{t,j}(\mathbf{w}_{(t-1) \wedge \xi})| > \Lambda \mid \xi > 0] < \frac{\delta}{4nT}. \quad (\text{D.5})$$

Then

$$\Pr[\xi < T] < \delta.$$

Proof of Lemma D.4. The proof is based on induction on T . Our induction hypothesis is that for any $t \in [T]$

$$\Pr[\xi < t] < \frac{\delta}{2} + \frac{\delta t}{2T}.$$

For the base case where $t = 1$, we have

$$\Pr[\xi < 1] = \Pr[\xi = 0] = \Pr[\mathbf{w}_0 \text{ does not satisfy the conditions in Lemma D.3}] < \frac{\delta}{2}.$$

For the induction step, assume the induction hypothesis holds for $t - 1$ for some $1 < t \leq T$. Consider

$$\Pr[\xi < t] = \Pr[\xi < t - 1] + \Pr[\xi = t - 1].$$

By the induction hypothesis, the first term can be bounded by $\frac{\delta}{2} + \frac{\delta(t-1)}{2T}$. The second term can also be upper bounded from definition as follows.

$$\begin{aligned} &< \frac{\delta}{2} + \frac{\delta(t-1)}{2T} + \Pr[\exists j \in [n], |f_{t,j}(\mathbf{w}_{t-1})| > \Lambda, \xi \geq t-1] \\ &= \frac{\delta}{2} + \frac{\delta(t-1)}{2T} + \Pr[\exists j \in [n], |f_{t,j}(\mathbf{w}_{(t-1) \wedge \xi})| > \Lambda, \xi \geq t-1]. \end{aligned}$$

Finally, from the assumption (*i.e.*, [Equation D.5](#)) in the lemma statement, the second term can be union bounded by $n \cdot \frac{1}{\Pr[\xi > 0]} \cdot \frac{\delta}{4nT}$. As $\Pr[\xi > 0] \geq 0.5$, the inequality becomes

$$\leq \frac{\delta}{2} + \frac{\delta(t-1)}{2T} + n \cdot \frac{1}{\Pr[\xi > 0]} \cdot \frac{\delta}{4nT} \leq \frac{\delta}{2} + \frac{\delta t}{2T}.$$

This completes the induction. Thus, $\Pr[\xi < T] \leq \delta$ as desired. $\square$

To get [Equation D.5](#), we need to show the concentration of $f_{t,j}(\mathbf{w}_{t-1})$ and as before the linearization and ODE trick would be our main tools. In the end, we prove [Equation D.5](#) in [Section D.4](#).

D.2 Linearization and ODE trick

Let us start with the linearization and ODE trick for function $f_{t,j}$ in this subsection.

Lemma D.6 (Linearization). *For any $t \in [T]$ and $j \in [n]$. For any $s \in [t]$, write $\mathbf{w}_s = \mathbf{w}_{s-1} + \eta \mathbf{z}_s$ where $\mathbf{z}_s = y_s(\mathbf{x}_s - y_s \mathbf{w}_{s-1})$. There exists $\bar{\mathbf{w}}_{s-1} = \mathbf{w}_{s-1} + c\eta \mathbf{z}_s$ for some $c \in [0, 1]$ such that*

$$f_{t,j}(\mathbf{w}_s) = (1 - \eta(\lambda_1 - \lambda_j))f_{t,j}(\mathbf{w}_{s-1}) + \eta A_{t,s,j} + \eta B_{t,s,j} + \eta^2 C_{t,s,j}$$

where

$$\begin{aligned} A_{t,s,j} &= \sum_{i=2}^{j-1} (\lambda_i - \lambda_{i+1}) f_{t,i}(\mathbf{w}_{s-1}), \\ B_{t,s,j} &= \sum_{i=1}^n \frac{\partial f_{t,j}}{\partial \mathbf{w}_i}(\mathbf{w}_{s-1}) \cdot (\mathbf{z}_{s,i} - \mathbb{E}[\mathbf{z}_{s,i} \mid \mathcal{F}_{s-1}]), \text{ and} \\ C_{t,s,j} &= \sum_{i,i'=1}^n \frac{\partial^2 f_{t,j}}{\partial \mathbf{w}_i \partial \mathbf{w}_{i'}}(\bar{\mathbf{w}}_{s-1}) \cdot \mathbf{z}_{s,i} \mathbf{z}_{s,i'}. \end{aligned}$$

Vector form of linearization For any $t \in [T]$, let $\mathbf{f}_t(\mathbf{w}), \mathbf{B}_{t,s}, \mathbf{C}_{t,s} \in \mathbb{R}^{n-1}$ be $(n-1)$ -dimensional vectors where the i^{th} coordinates of them are $f_{t,i+1}(\mathbf{w}), B_{t,s,i+1}, C_{t,s,i+1}$ respectively. The following is an immediate corollary of Lemma D.6 by rewriting everything into a vector form.

Corollary D.7 (Linearization in vector form). *For any $t \in [T]$ and $s \in [t]$, we have*

$$\mathbf{f}_t(\mathbf{w}_s) = H \mathbf{f}_t(\mathbf{w}_{s-1}) + \eta \mathbf{B}_{t,s} + \eta^2 \mathbf{C}_{t,s}$$

where

$$H = \begin{pmatrix} 1 - \eta(\lambda_1 - \lambda_2) & 0 & 0 & \cdots & 0 \\ \eta(\lambda_2 - \lambda_3) & 1 - \eta(\lambda_1 - \lambda_3) & 0 & \cdots & 0 \\ \eta(\lambda_2 - \lambda_3) & \eta(\lambda_3 - \lambda_4) & 1 - \eta(\lambda_1 - \lambda_4) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \eta(\lambda_2 - \lambda_3) & \eta(\lambda_3 - \lambda_4) & \eta(\lambda_4 - \lambda_5) & \cdots & 1 - \eta(\lambda_1 - \lambda_n) \end{pmatrix}.$$

It turns out that the multiplier matrix H in Corollary D.7 is well-behaved: it is invertible and we can explicitly write down its eigenvectors as follows. For each $i \in [n-1]$, the i^{th} eigenvalue-eigenvector pair $(\mu_i, \mathbf{v}_i)$ is

$$\mu_i = 1 - \eta(\lambda_1 - \lambda_{i+1}) \quad \text{and} \quad \mathbf{v}_i = [\underbrace{0, \dots, 0}_{i-1}, \underbrace{1, \dots, 1}_{n-i}]^\top.$$

With the eigenvectors, we can easily diagonalize H as follows. Let $V \in \mathbb{R}^{(n-1) \times (n-1)}$ be the matrix with the i^{th} column being $\mathbf{v}_i$, we have $H = V D V^{-1}$ where

$$D = \text{diag}(1 - \eta(\lambda_1 - \lambda_2), 1 - \eta(\lambda_1 - \lambda_3), \dots, 1 - \eta(\lambda_1 - \lambda_n)).$$

Thus, the inverse of H is $H^{-1} = V D^{-1} V^{-1}$.

By the ODE trick for vector (see Lemma 2.13), we immediately have the following corollary for a closed-form solution to $\mathbf{f}_t(\mathbf{w}_s)$.

Corollary D.8 (ODE trick). *For any $t \in [T]$, $s \in [t-1]$, and $\eta > 0$, we have*

$$\mathbf{f}_t(\mathbf{w}_s) = H^s \mathbf{f}_t(\mathbf{w}_0) + \sum_{s'=1}^s H^{s-s'} (\eta \mathbf{B}_{t,s'} + \eta^2 \mathbf{C}_{t,s'}).$$

D.3 Concentration of the noise terms

In this subsection, we prove Equation D.5. However, same as the situation before, we cannot get concentration for the noise terms of the ODE trick (see Corollary D.8) directly. As a consequence, we have to introduce a new stopping time τ_t to make sure the stopped processes are in a good shape for the martingale concentration inequality.

D.3.1 Stopping time

For a fixed $t \in [T]$, we define a stopping time τ_t for the noise terms from $s = 1, 2, \dots, t-1$ as follows. First, we work on a slightly different filtration $\{\mathcal{F}_s^{(t)}\}_{s \in [t-1]}$ than the natural filtration $\{\mathcal{F}_s\}_{s \in [t-1]}$. The key idea is that the stopping time can depend on $\mathbf{x}_t$ since we only look at the noise term up to $t-1$. Concretely, for each $s \in [t-1]$, let $\mathcal{F}_s^{(t)}$ be the σ -algebra generated by $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_s\} \cup \{\mathbf{x}_t\}$. Note that $\{\mathcal{F}_s^{(t)}\}_{s \in [t-1]}$ is well-defined and both $\{B_{t,s,j}\}_{s \in [t-1]}$ and $\{C_{t,s,j}\}_{s \in [t-1]}$ are adapted random processes with respect to $\{\mathcal{F}_s^{(t)}\}_{s \in [t-1]}$, i.e., $B_{t,s,j}$ and $C_{t,s,j}$ lie in $\mathcal{F}_s^{(t)}$ for all $s \in [t-1]$. Also, note that $\mathbb{E}[\mathbf{z}_s \mid \mathcal{F}_{s-1}] = \mathbb{E}[\mathbf{z}_s \mid \mathcal{F}_{s-1}^{(t)}]$. That is, the conditional expectation and conditional variance of $\mathbf{z}$ are the same with respect to $\{\mathcal{F}_s\}$ and $\{\mathcal{F}_s^{(t)}\}$.

Now we define τ_t to be the stopping time for the event $\{\|\mathbf{f}_t(\mathbf{w}_{\tau_t})\|_\infty > \Lambda\}$. Note that the stopped process $\left\{ \sum_{s'=1}^s H^{-(s'-s_0)} \mathbf{B}_{t,(s'-1) \wedge \xi \wedge \tau_t} \right\}_{s \in [t-1]}$ is a martingale and $\left\{ \sum_{s'=1}^s H^{-(s'-s_0)} \mathbf{C}_{t,(s'-1) \wedge \xi \wedge \tau_t} \right\}_{s \in [t-1]}$ is an adapted stochastic process with respect to $\{\mathcal{F}_s^{(t)}\}_{s \in [t-1]}$. Furthermore, both of them have small bounded difference, conditional expectation, and conditional variance. Concretely, we have the following lemma.

Lemma D.9 (Structure of the stopped processes). *Let T, η, Λ be the parameters and ξ, τ_t be the stopping times as chosen before. For any $t \in [T]$, $s \in [t-1]$, and $j \in [n]$, let $M_{t,s,j}^B$ and $M_{t,s,j}^C$ be the j^{th} entry of $\sum_{s'=1}^s H^{s-s'} \mathbf{B}_{t,s' \wedge \xi \wedge \tau_t}$ and $\sum_{s'=1}^s H^{s-s'} \mathbf{C}_{t,s' \wedge \xi \wedge \tau_t}$ respectively. The following hold.*

- (Bounded difference) For any $j \in [n]$, we have

$$|M_{t,s,j}^B - M_{t,s-1,j}^B| = O(\Lambda^2) \quad \text{and} \quad |M_{t,s,j}^C - M_{t,s-1,j}^C| = O(\Lambda^3) \text{ almost surely.}$$

- (Conditional expectation) For any $s \in [t-1]$ and $j \in [n]$, we have

$$\mathbb{E} \left[M_{t,s,j}^B - M_{t,s-1,j}^B \mid \mathcal{F}_{s-1}^{(t)} \right] = 0 \quad \text{and} \quad \sum_{s'=1}^s \left| \mathbb{E} \left[M_{t,s',j}^C - M_{t,s'-1,j}^C \mid \mathcal{F}_{s'-1}^{(t)} \right] \right| = O \left(\frac{\lambda_1 \Lambda^3}{\eta(\lambda_1 - \lambda_2)} \right).$$

- (Conditional variance) For any $s \in [t-1]$ and $j \in [n]$, we have

$$\sum_{s'=1}^s \text{Var} \left[M_{t,s',j}^B \mid \mathcal{F}_{s'-1}^{(t)} \right] = O \left(\frac{\lambda_1 \Lambda^4}{\eta(\lambda_1 - \lambda_2)} \right) \quad \text{and} \quad \sum_{s'=1}^s \text{Var} \left[M_{t,s',j}^C \mid \mathcal{F}_{s'-1}^{(t)} \right] = O \left(\frac{\lambda_1 \Lambda^6}{\eta(\lambda_1 - \lambda_2)} \right).$$

Proof of Lemma D.9. Before looking at $M_{t,s,j}^B$ and $M_{t,s,j}^C$, which are weighted by some factor of H , let us first understand the structure of the unweighted version in the following lemma.

Lemma D.10. *Let T, η, Λ be the parameters and ξ, τ_t be the stopping times as chosen before. For any $t \in [T]$, $s \in [t-1]$, and $j \in [n]$, the following hold.*

- (Bounded difference) We have

$$|B_{t,s \wedge \xi \wedge \tau_t,j}| = O(\Lambda^2) \quad \text{and} \quad |C_{t,s \wedge \xi \wedge \tau_t,j}| = O(\Lambda^3) \text{ almost surely.}$$

- (Conditional expectation) We have

$$\mathbb{E}[B_{t,s \wedge \xi \wedge \tau_t,j} \mid \mathcal{F}_{s-1}^{(t)}] = 0 \quad \text{and} \quad \left| \mathbb{E}[C_{t,s \wedge \xi \wedge \tau_t,j} \mid \mathcal{F}_{s-1}^{(t)}] \right| = O(\lambda_1 \Lambda^3).$$

- (Conditional variance) We have

$$\text{Var}[B_{t,s \wedge \xi \wedge \tau_t,j} \mid \mathcal{F}_{s-1}^{(t)}] = O(\lambda_1 \Lambda^4) \quad \text{and} \quad \text{Var}[C_{t,s \wedge \xi \wedge \tau_t,j} \mid \mathcal{F}_{s-1}^{(t)}] = O(\lambda_1 \Lambda^6).$$

Proof of Lemma D.10. The proof is based on properly expanding the noise terms and utilizes the stopping time condition. We postpone the details to Section E.3. $\square$

To move the unweighted structure (*i.e.*, Lemma D.10) to the desiring weighted structure (*i.e.*, for $M_{t,s,j}^B$ and $M_{t,s,j}^C$), we need the following claim on translating the bounded difference and variance after being multiplied by some power of H .

Claim D.11. *Let $H \in \mathbb{R}^{(n-1) \times (n-1)}$ be the matrix from Corollary D.7. For any $s \in \mathbb{N}$, $c, \mu, \sigma^2 > 0$, and random vector $\mathbf{u} \in \mathbb{R}^{n-1}$ such that (i) $\|\mathbf{u}\|_\infty \leq c$ almost surely, (ii) $\|\mathbb{E}[\mathbf{u}]\|_\infty \leq \mu$, and (iii) $\text{Var}[\mathbf{u}_i] \leq \sigma^2$ for each $i \in [n-1]$. The following hold.*

- (Bounded difference) For any $s \in \mathbb{N}$, $\|H^s \mathbf{u}\|_\infty \leq (1 - \eta(\lambda_1 - \lambda_2))^s \cdot c$ almost surely.
- (Expectation) $\|\mathbb{E}[H^s \mathbf{u}]\|_\infty \leq (1 - \eta(\lambda_1 - \lambda_2))^s \cdot \mu$.
- (Variance) $\text{Var}[(H^s \mathbf{u})_i] \leq (1 - \eta(\lambda_1 - \lambda_2))^{2s} \cdot \sigma^2$ for each $i \in [n-1]$.

Proof of Claim D.11. Recall that $H = VDV^{-1}$ is invertible where

$$V = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix} \quad \text{and} \quad V^{-1} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}.$$

Also, observe that for any diagonal matrix $D' = \text{diag}(d_1, d_2, \dots, d_{n-1})$, we have

$$VD'V^{-1} = \begin{pmatrix} d_1 & 0 & 0 & \cdots & 0 \\ d_1 - d_2 & d_2 & 0 & \cdots & 0 \\ d_1 - d_2 & d_2 - d_3 & d_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_1 - d_2 & d_2 - d_3 & d_3 - d_4 & \cdots & d_{n-1} \end{pmatrix}.$$

Note that if $d_1 \geq d_2 \geq \dots \geq d_{n-1}$, the ℓ_1 norm of each row of $VD'V^{-1}$ is exactly d_1 and the ℓ_2 norm is at most d_1 . To see the bound for the ℓ_2 norm, for any $i \in [n-1]$, the square of the ℓ_2 norm of the i^{th} row of $VD'V^{-1}$ is

$$\sum_{j=1}^{i-1} (d_j - d_{j+1})^2 + d_i^2 \leq \left(\sum_{j=1}^{i-1} (d_j - d_{j+1}) + d_i \right)^2 = d_1^2.$$

The first inequality holds because $d_j - d_{j+1} \geq 0$ for all $j \in [n-2]$.

Now, let us go back to $H^s = VD^sV^{-1}$ and recall that the i^{th} entries of D are $(1 - \eta(\lambda_1 - \lambda_{i+1}))$ for each $i \in [n-1]$. Namely, the diagonal entries of D^s from top left to bottom right are non-increasing and the corresponding $d_1 = (1 - \eta(\lambda_1 - \lambda_2))^s$.

Finally, let $\mathbf{u} \in \mathbb{R}^{n-1}$ be a random vector with the properties as stated in the claim, for any $i \in [n-1]$, we have

$$(H^s \mathbf{u})_i = (VD^sV^{-1} \mathbf{u})_i = \mathbf{v}_{s,i}^\top \mathbf{u}$$

where $\mathbf{v}_{s,i}$ is the i^{th} row of VD^sV^{-1} . By the above discussion, $\|\mathbf{v}_{s,i}\|_1, \|\mathbf{v}_{s,i}\|_2 \leq d_1 = (1 - \eta(\lambda_1 - \lambda_2))^s$. Thus, we have both

$$|(H^s \mathbf{u})_i| = |\mathbf{v}_{s,i}^\top \mathbf{u}| \leq \|\mathbf{v}_{s,i}\|_1 \cdot \|\mathbf{u}\|_\infty \quad \text{and} \quad |(H^s \mathbf{u})_i| = |\mathbf{v}_{s,i}^\top \mathbf{u}| \leq \|\mathbf{v}_{s,i}\|_2 \cdot \|\mathbf{u}\|_2$$

almost surely due to Hölder's inequality. This gives $|(H^s \mathbf{u})_i| \leq (1 - \eta(\lambda_1 - \lambda_2))^s \cdot c$ almost surely and $\text{Var}[(H^s \mathbf{u})_i] \leq (1 - \eta(\lambda_1 - \lambda_2))^{2s} \cdot \sigma^2$. The expectation of $(H^s \mathbf{u})_i$ is handled by linearity as follows.

$$\mathbb{E}[(H^s \mathbf{u})_i] = \mathbb{E}[\mathbf{v}_{s,i}^\top \mathbf{u}] = \mathbf{v}_{s,i}^\top \mathbb{E}[\mathbf{u}] \leq \|\mathbf{v}_{s,i}\|_1 \cdot \|\mathbb{E}[\mathbf{u}]\|_\infty.$$

This completes the proof of the claim. $\square$

Now, the structure for the vector form (*i.e.*, [Lemma D.9](#)) immediately follows from the structure of the individual noise term (*i.e.*, [Lemma D.10](#)), [Claim D.11](#), Cauchy-Schwarz inequality, and the following formula.

$$\sum_{s=0}^{\infty} (1 - \eta(\lambda_1 - \lambda_2))^s = \frac{1}{\eta(\lambda_1 - \lambda_2)}.$$

□

D.3.2 Concentration for the stopped processes

As a consequence of [Lemma D.10](#), we are able to prove the following concentration for the stopped processes of the noise terms.

Lemma D.12 (Concentration for the stopped process of the noise vectors). *Let T, η, Λ be the parameters and ξ, τ_t be the stopping times as chosen before. For any $t \in [T]$, $s \in [t-1]$, and $i \in [n-1]$, the following hold.*

1.

$$\Pr \left[\sup_{s' \in [s]} \sum_{s''=1}^{s'} \eta \left(H^{s'-s''} \mathbf{B}_{t, s'' \wedge \xi \wedge \tau_t} \right)_i \geq \Omega \left(\Lambda^2 \sqrt{\frac{\eta \lambda_1 \log \frac{1}{\delta'}}{\lambda_1 - \lambda_2}} \right) \right] < \delta'.$$

2.

$$\Pr \left[\sup_{s' \in [s]} \sum_{s''=1}^{s'} \eta^2 \left(H^{s'-s''} \mathbf{C}_{t, s'' \wedge \xi \wedge \tau_t} \right)_i \geq \Omega \left(\frac{\eta^3 \lambda_1 \Lambda^3}{\lambda_1 - \lambda_2} + \eta \Lambda^3 \sqrt{\frac{\eta \lambda_1 \log \frac{1}{\delta'}}{(\lambda_1 - \lambda_2)}} \right) \right] < \delta'.$$

Proof of Lemma D.12. The proof is based on applying the corollary of Freedman's inequality (see [Corollary 2.7](#)) on each coordinate using [Lemma D.9](#). Since the calculation is similar to that of [Lemma 6.6](#) and [Lemma 7.6](#), we omit the details here for simplicity. □

The following lemma calculates the deviation in [Lemma D.12](#) after fixing the parameters and weighting the noise terms with η .

Lemma D.13. *With the parameter settings in [Section D.1](#), by direct verification we have*

$$O \left(\Lambda^2 \sqrt{\frac{\eta \lambda_1 \log \frac{1}{\delta'}}{\lambda_1 - \lambda_2}} \right), O \left(\frac{\eta^3 \lambda_1 \Lambda^3}{\lambda_1 - \lambda_2} + \eta \Lambda^3 \sqrt{\frac{\eta \lambda_1 \log \frac{1}{\delta'}}{(\lambda_1 - \lambda_2)}} \right) < \frac{\Lambda}{10}$$

where the O terms are from the concentration inequalities of [Lemma D.12](#).

D.3.3 Pull-out lemma

Now that we have good concentration on the stopped process of the noise vector in the ODE trick, the last step is pulling out the stopping time. Note that there are two differences between the setting here and the setting in Phase 2: here we are working on a vector of random processes instead of a single one and there is an initial condition (*i.e.*, [Lemma D.3](#)) we have to condition on throughout the analysis. Thus, we have to develop the following generalization of [Lemma 7.9](#).

Lemma D.14. *Let $m \in \mathbb{N}$ and $\{M_{s,j}\}_{s \in \mathbb{N}_{\geq 0}}$ be m -dimensional random processes indexed by $j \in [m]$. Denote its maximal process as $\{M_{s,j}^*\}_{s \in \mathbb{N}_{\geq 0}}$ where $M_{s,j}^* = \sup_{s' \in [s]} M_{s',j}$ for any $s \in \mathbb{N}$ and $j \in [m]$.*

For any $s \in \mathbb{N}$, $a \in \mathbb{R}$, $\delta \in (0, 1)$, stopping time τ , and an event $\mathcal{E}$, suppose

1. $\Pr[\exists j \in [n], M_{s \wedge \tau, j}^* \geq a \mid \mathcal{E}] < \delta$ and
2. $\Pr[\tau \geq s' + 1 \mid \forall j \in [n], M_{s', j}^* < a, \mathcal{E}] = 1$ for any $1 \leq s' < s$.

Then, we have

$$\Pr[\exists j \in [n], M_{s,j}^* \geq a] < \delta.$$

Proof. Let stopping time τ' be the first time t such that $\exists j \in [n], M_{t \wedge \tau, j} \geq a$. Similarly, we first have the following observation.

Claim D.15. *For any $s \in \mathbb{N}$, we have*

$$\Pr[\tau < s, \tau' > \tau \mid \mathcal{E}] = 0.$$

Proof of Claim D.15. Let us assume both $\tau < s$ and $\tau' > \tau$ for the sake of contradiction. Since $\tau' > \tau$, we know that for any $j \in [n]$ and $s'' \leq \tau$, $M_{s'',j} < a$. From the second condition of Lemma D.14, we have $\tau \geq \tau + 1$ almost surely, which is a contradiction. Thus, we conclude that the two events cannot simultaneously happen. $\square$

Now, let us partition the probability space of the error event as follows.

$$\begin{aligned} \Pr[\exists j \in [n], M_{s,j}^* \geq a \mid \mathcal{E}] &= \Pr[\exists j \in [n], M_{s,j}^* \geq a, \tau \geq s \mid \mathcal{E}] + \Pr[\exists j \in [n], M_{s,j}^* \geq a, \tau < s, \tau' \leq \tau \mid \mathcal{E}] \\ &\quad + \Pr[\exists j \in [n], M_{s,j}^* \geq a, \tau < s, \tau' > \tau \mid \mathcal{E}]. \end{aligned}$$

Note that when $\tau \geq s$, we have $s = s \wedge \tau$ and thus $M_{s,j}^* = M_{s \wedge \tau, j}^*$ for all $j \in [n]$. Also, when $\tau' \leq \tau < s$, we know that there exists $j \in [n]$ such that both $M_{s,j}^*, M_{s \wedge \tau, j}^* \geq a$. Thus, the equation becomes

$$\begin{aligned} &= \Pr[\exists j \in [n], M_{s \wedge \tau, j}^* \geq a, \tau \geq s \mid \mathcal{E}] + \Pr[\exists j \in [n], M_{s \wedge \tau, j}^* \geq a, \tau < s, \tau' \leq \tau \mid \mathcal{E}] \\ &\quad + \Pr[\exists j \in [n], M_{s,j}^* \geq a, \tau < s, \tau' > \tau \mid \mathcal{E}]. \end{aligned}$$

Moreover, by Claim D.15, we know that the last term is 0. Finally, we have

$$\leq \Pr[\exists j \in [n], M_{s \wedge \tau, j}^* \geq a \mid \mathcal{E}] < \delta.$$

This completes the proof of the pull-out lemma. $\square$

D.4 Wrap up

Now, let us formally prove Equation D.5 and thus by Lemma D.4 and Lemma D.2 this would finish the proof of Lemma 6.3 and fulfill the goal of this section.

Proof of Equation D.5. Recall that the goal is to show that

$$\Pr[|f_{t,j}(\mathbf{w}_{(t-1) \wedge \xi})| > \Lambda] < \frac{\delta}{4nT}$$

for all $t \in [T]$ and $j \in [n]$. The high-level idea is first showing that the stopped process of the noise vectors would concentrate (using Lemma D.12) and then using the pull-out lemma (*i.e.*, Lemma D.14) to get the concentration for the original noise vectors.

For the ease of notation, for any $t \in [T]$, let us define

$$\text{BAD}_t := \left\{ \exists s' \in [t-1], i \in [n-1], \left| \sum_{s''=1}^{s'} \eta \left(H^{s'-s''} \mathbf{B}_{t, s'' \wedge \xi} \right)_i + \eta^2 \left(H^{s'-s''} \mathbf{C}_{t, s'' \wedge \xi} \right)_i \right| > \frac{\Lambda}{5} \right\}$$

and

$$\text{BAD}_{t, \tau_t} := \left\{ \exists s' \in [t-1], i \in [n-1], \left| \sum_{s''=1}^{s'} \eta \left(H^{s'-s''} \mathbf{B}_{t, s'' \wedge \xi \wedge \tau_t} \right)_i + \eta^2 \left(H^{s'-s''} \mathbf{C}_{t, s'' \wedge \xi \wedge \tau_t} \right)_i \right| > \frac{\Lambda}{5} \right\}.$$

For any $t \in [T]$, by [Lemma D.12](#), [Lemma D.13](#), and union bound, we have

$$\Pr[\text{BAD}_{t,\tau_t}] < 4n\delta'. \quad (\text{D.16})$$

To apply the pull-out lemma (*i.e.*, [Lemma D.14](#)), we have to check the second condition as follows. For any $1 \leq s \leq t-1$, assume for any $i \in [n-1]$ and $s' \in [s]$, we have

$$\sum_{s''=1}^{s'} \eta \left(H^{s'-s''} \mathbf{B}_{t,s'' \wedge \xi} \right)_i + \eta^2 \left(H^{s'-s''} \mathbf{C}_{t,s'' \wedge \xi} \right)_i < \frac{\Lambda}{5}.$$

Now, we would like to bound each coordinate of $\mathbf{f}(\mathbf{w}_{s \wedge \xi})$. Specifically, we only care about the case where $\xi > 0$. By the ODE trick (*i.e.*, [Corollary D.8](#)), for any $i \in [n-1]$, we have

$$\mathbf{f}_t(\mathbf{w}_{s \wedge \xi}) \stackrel{\xi > 0}{=} H^{s \wedge \xi} \mathbf{f}_t(\mathbf{w}_0) + \sum_{s'=1}^s H^{(s \wedge \xi) - s'} (\eta \mathbf{B}_{t,s' \wedge \xi} + \eta^2 \mathbf{C}_{t,s' \wedge \xi}).$$

As $\xi > 0$ implies $\|\mathbf{f}_t(\mathbf{w}_0)\|_\infty < \Lambda/2$, combining with the above assumption, we have $\|\mathbf{f}_t(\mathbf{w}_{s \wedge \xi})\|_\infty < \Lambda$ and thus $\tau_t \geq s+1$. Namely,

$$\Pr \left[\tau_t \geq s+1 \mid \neg \text{BAD}_s, \xi > 0 \right] = 1, \quad (\text{D.17})$$

i.e., the second condition of the pull-out lemma holds. Finally, invoke the pull-out lemma (*i.e.*, [Lemma D.14](#)) with [Equation D.17](#) and event $\mathcal{E} = \{\xi > 0\}$, we have

$$\Pr[\text{BAD}_t \mid \xi > 0] \leq \Pr[\text{BAD}_{t,\tau_t} \mid \xi > 0] \leq \frac{\Pr[\text{BAD}_{t,\tau_t}, \xi > 0]}{\Pr[\xi > 0]} < 8n\delta'$$

where the last inequality is due to $\Pr[\xi > 0] \geq 0.5$ and [Equation D.16](#). Finally, using the ODE trick, we have

$$\Pr[\exists j \in [n], |f_{t,j}(\mathbf{w}_{(t-1) \wedge \xi})| > \Lambda \mid \xi > 0] \stackrel{(\because \text{ODE trick})}{\leq} \Pr[\text{BAD}_t \mid \xi > 0] \stackrel{(\because \text{Equation D.16})}{<} 8n\delta' \leq \frac{\delta}{4nT}.$$

Recall that the last inequality holds because we piked $\delta' = \frac{\delta}{32n^2T}$ in [Section D.1](#), thus we get [Equation D.5](#). This completes the proof. $\square$

E Details of Linearizations in Biological Oja's Rule

In this section, we provide the full proofs for the linearization lemmas for biological Oja's rule. In [Section E.1](#) we provide the proof of the linearization in Phase 1, in [Section E.2](#) we provide the proof of the linearization in Phase 2, and in [Section E.3](#) we provide the proof of the linearization used in [Appendix D](#) for the proof of [Lemma 6.3](#).

We remark that all of these linearization lemmas are based on Taylor's expansion and the proofs are elementary. The key of the proofs is to identify the *right way* to rearrange the terms so that everything is analyzable. The rule of thumb is separating the *drifting terms* and the *noise terms* where in most of the case the later are either (i) martingales or (ii) something with small bounded difference and small conditional expectation. Namely, martingale concentration inequality (*e.g.*, [Corollary 2.7](#)) can show that the noise terms are small with high probability and thus the dynamic is dominated by the drifting terms.

E.1 Linearization in Phase 1

In this subsection, we prove [Lemma 6.2](#) using [Lemma E.1](#) and [Lemma E.3](#).

Lemma E.1 (Linearization in Phase 1). *For any $t \in \mathbb{N}$ and $\eta \in (0, 0.1)$, we have*

$$\mathbf{w}_{t,1} = H_t \cdot \mathbf{w}_{t-1,1} + \eta D_t + \eta F_t + \eta^2 Q_t$$

almost surely, where

$$\begin{aligned} A_t &= -\sum_{i \neq j} \mathbf{x}_{t,i} \mathbf{x}_{t,j} \mathbf{w}_{t-1,i} \mathbf{w}_{t-1,j}, & B_t &= \sum_{i=3}^n (\mathbf{x}_{t,2}^2 - \mathbf{x}_{t,i}^2 - (\lambda_2 - \lambda_i)) \mathbf{w}_{t-1,i}^2, \\ C_t &= (1 - \mathbf{w}_{t-1,1}^2) (\mathbf{x}_{t,1}^2 - \mathbf{x}_{t,2}^2 - (\lambda_1 - \lambda_2)), & D_t &= \sum_{i=2}^n \mathbf{x}_{t,1} \mathbf{x}_{t,i} \mathbf{w}_{t-1,i}, \\ E_t &= \sum_{i=3}^n (\lambda_2 - \lambda_i) \mathbf{w}_{t-1,i}^2, & F_t &= \mathbf{w}_{t-1,1} (A_t + B_t + C_t), \\ Q_t &= O(\mathbf{x}_{t,2}^2 \mathbf{w}_{t-1,1} + \lambda_1 \mathbf{w}_{t-1,1}) & H_t &= \exp(\eta(\lambda_1 - \lambda_2)(1 - \mathbf{w}_{t-1,1}^2) + \eta E_t). \end{aligned}$$

Proof of Lemma E.1. Recall that from Equation 1.4, we have

$$\begin{aligned} \mathbf{w}_{t,1} &= \mathbf{w}_{t-1,1} + \eta \cdot \left[\sum_{i=1}^n \mathbf{x}_{t,i} \mathbf{w}_{t-1,i} \mathbf{x}_{t,1} - \left(\sum_{i=1}^n \mathbf{x}_{t,i} \mathbf{w}_{t-1,i} \right)^2 \mathbf{w}_{t-1,1} \right] \\ &= \mathbf{w}_{t-1,1} + \eta \cdot \left[\mathbf{x}_{t,1}^2 \mathbf{w}_{t-1,1} + D_t - \left(\mathbf{x}_{t,1}^2 \mathbf{w}_{t-1,1} + \mathbf{x}_{t,2}^2 \mathbf{w}_{t-1,2}^2 + \sum_{i=3}^n \mathbf{x}_{t,i}^2 \mathbf{w}_{t-1,i}^2 - A_t \right) \mathbf{w}_{t-1,1} \right]. \end{aligned}$$

From Lemma 5.1, we can replace $\mathbf{w}_{t-1,2}^2$ with $1 - \mathbf{w}_{t-1,1}^2 - \sum_{i=3}^n \mathbf{w}_{t-1,i}^2 + O(\eta)$. Let $B'_t = \sum_{i=3}^n (\mathbf{x}_{t,2}^2 - \mathbf{x}_{t,i}^2) \mathbf{w}_{t-1,i}^2$, we have

$$= \mathbf{w}_{t-1,1} + \eta \cdot [(\mathbf{x}_{t,1}^2 - \mathbf{x}_{t,2}^2) \mathbf{w}_{t-1,1} (1 - \mathbf{w}_{t-1,1}^2) + \mathbf{w}_{t-1,1} B'_t + \mathbf{w}_{t-1,1} A_t + D_t] + O(\eta^2 \mathbf{x}_{t,2}^2 \mathbf{w}_{t-1,1}).$$

By adding and subtracting $\eta(\lambda_1 - \lambda_2) \mathbf{w}_{t-1,1} (1 - \mathbf{w}_{t-1,1}^2)$ and $\eta \mathbf{w}_{t-1,1} E_t$, we have

$$= \mathbf{w}_{t-1,1} + \eta \cdot [(\lambda_1 - \lambda_2) \mathbf{w}_{t-1,1} (1 - \mathbf{w}_{t-1,1}^2) + \mathbf{w}_{t-1,1} (A_t + B_t + C_t + E_t) + D_t] + O(\eta^2 \mathbf{x}_{t,2}^2 \mathbf{w}_{t-1,1}). \quad (\text{E.2})$$

Let $F_t = \mathbf{w}_{t-1,1} (A_t + B_t + C_t)$ and $Q_t = O(\mathbf{x}_{t,2}^2 \mathbf{w}_{t-1,1} + \lambda_1 \mathbf{w}_{t-1,1})$. By Lemma 2.16, we have

$$= \mathbf{w}_{t-1,1} \cdot \exp(\eta(\lambda_1 - \lambda_2)(1 - \mathbf{w}_{t-1,1}^2) + \eta E_t) + \eta D_t + \eta F_t + \eta^2 Q_t.$$

Finally, let $H_t = \exp(\eta(\lambda_1 - \lambda_2)(1 - \mathbf{w}_{t-1,1}^2) + \eta E_t)$, we have

$$= H_t \cdot \mathbf{w}_{t-1,1} + \eta D_t + \eta F_t + \eta^2 Q_t.$$

□

Lemma E.3 (Bounded difference and conditional variance in Phase 1). *For each $t \in \mathbb{N}$, we have the following.*

- (Bounded difference) For any $t \in \mathbb{N}$, $E_t \geq 0$, $|F_t| = O(\mathbf{w}_{t-1,1})$, and $|Q_t| = O(\mathbf{w}_{t-1,1})$ almost surely.
- (Conditional expectation) For any $t \in \mathbb{N}$, $\mathbb{E}[F_t \mid \mathcal{F}_{t-1}] = 0$ and $|\mathbb{E}[Q_t \mid \mathcal{F}_{t-1}]| = O(\lambda_1 \mathbf{w}_{t-1,1})$.
- (Conditional variance) For any $t \in \mathbb{N}$, $\text{Var}[F_t \mid \mathcal{F}_{t-1}] = O(\lambda_1 \mathbf{w}_{t-1,1}^2)$ and $\text{Var}[Q_t \mid \mathcal{F}_{t-1}] = O(\lambda_1 \mathbf{w}_{t-1,1}^2)$.

Proof of Lemma E.3.

- (Bounded difference) Let us start with bounding $|F_t| = |\mathbf{w}_{t-1,1} (A_t + B_t + C_t)|$. A straightforward application of Cauchy-Schwarz inequality would give $|B_t|, |C_t| = O(1)$ almost surely. The bound for $|A_t|$ is slightly trickier as follows.

$$\begin{aligned} |A_t| &\leq \sum_{i=1}^n |\mathbf{x}_{t,i} \mathbf{w}_{t-1,i}| \cdot \sum_{j \neq i} |\mathbf{x}_{t,j} \mathbf{w}_{t-1,j}| \\ (\because \text{Cauchy-Schwarz}) &= O(1) \end{aligned}$$

almost surely.

As for the other two quantities, follow from their definitions we have $Q_t = O(\mathbf{w}_{t-1,1})$ and $E_t \geq 0$ almost surely.

- (Conditional expectation) Since $\mathbb{E}[\mathbf{x}_{t,i}\mathbf{w}_{t,j} \mid \mathcal{F}_{t-1}] = \lambda_i \cdot \mathbf{1}_{i=j}$ for any $i, j \in [n]$, we immediately have $\mathbb{E}[F_t \mid \mathcal{F}_{t-1}] = 0$ and $|\mathbb{E}[Q_t \mid \mathcal{F}_{t-1}]| = O(\lambda_1 \mathbf{w}_{t-1,1}^2)$.
- (Conditional variance) First rewrite

$$A_t = -2 \sum_{i=2}^n \mathbf{x}_{t,1} \mathbf{x}_{t,i} \mathbf{w}_{t-1,1} \mathbf{w}_{t-1,i} - \left(\sum_{i=2}^n \mathbf{x}_{t,i} \mathbf{w}_{t-1,i} \right)^2 + \sum_{i=2}^n \mathbf{x}_{t,i}^2 \mathbf{w}_{t-1,i}^2.$$

Consider the first term of A_t ,

$$\begin{aligned} \mathbb{E} \left[\sum_{i,j=2}^n \mathbf{x}_{t,1}^2 \mathbf{x}_{t,i} \mathbf{x}_{t,j} \mathbf{w}_{t-1,1}^2 \mathbf{w}_{t-1,i} \mathbf{w}_{t-1,j} \mid \mathcal{F}_{t-1} \right] &= \mathbb{E} \left[\mathbf{x}_{t,1}^2 \mathbf{w}_{t-1,1}^2 \cdot \left(\sum_{i=2}^n \mathbf{x}_{t,i} \mathbf{w}_{t-1,i} \right)^2 \mid \mathcal{F}_{t-1} \right] \\ (\because \text{Cauchy-Schwarz}) &\leq \mathbb{E} \left[\mathbf{x}_{t,1}^2 \mathbf{w}_{t-1,1}^2 \cdot \left(\sum_{i=2}^n \mathbf{x}_{t,i}^2 \sum_{i=2}^n \mathbf{w}_{t-1,i}^2 \right) \mid \mathcal{F}_{t-1} \right] \\ &\leq \mathbb{E} \left[O(\mathbf{x}_{t,1}^2 \mathbf{w}_{t-1,1}^2) \mid \mathcal{F}_{t-1} \right] \\ &\leq O(\lambda_1 \mathbf{w}_{t-1,1}^2). \end{aligned}$$

The conditional variance of last two terms of A_t and the B_t, C_t terms can be bounded similarly by noticing that $\mathbb{E}[x_{t,i}^4] \leq \lambda_1$. So $\text{Var}[F_t \mid \mathcal{F}_{t-1}] \leq O(\lambda_1 \mathbf{w}_{t-1,1}^2)$. On the other hand, the conditional variance of Q_t is $O(\lambda_1 \mathbf{w}_{t-1,1}^2)$ by checking the definition. $\square$

E.2 Linearization in Phase 2

In this subsection, we prove [Lemma 7.2](#) using [Lemma E.4](#) and [Lemma E.5](#).

Lemma E.4 (Linearization in Phase 2). *For any $t \in \mathbb{N}$, we have*

$$\mathbf{w}'_{t,1} = H \cdot \mathbf{w}'_{t-1,1} + \eta F_t + \eta P_t + \eta^2 Q_t$$

almost surely, where

$$\begin{aligned} A_t &= - \sum_{i \neq j} \mathbf{x}_{t,i} \mathbf{x}_{t,j} \mathbf{w}_{t-1,i} \mathbf{w}_{t-1,j}, & B_t &= \sum_{i=3}^n (\mathbf{x}_{t,2}^2 - \mathbf{x}_{t,i}^2 - (\lambda_2 - \lambda_i)) \mathbf{w}_{t-1,i}^2, \\ C_t &= (1 - \mathbf{w}_{t-1,1}^2) (\mathbf{x}_{t,1}^2 - \mathbf{x}_{t,2}^2 - (\lambda_1 - \lambda_2)), & D_t &= \sum_{i=2}^n \mathbf{x}_{t,1} \mathbf{x}_{t,i} \mathbf{w}_{t-1,i}, \\ E_t &= \sum_{i=3}^n (\lambda_2 - \lambda_i) \mathbf{w}_{t-1,i}^2, & F_t &= \mathbf{w}_{t-1,1} (A_t + B_t + C_t) + D_t, \\ G_t &= (\lambda_1 - \lambda_2) (1 - \mathbf{w}_{t-1,1}) (\mathbf{w}_{t-1,1}^2 + \mathbf{w}_{t-1,1} - \frac{3}{4}), & Q_t &= O(\mathbf{x}_{t,2}^2 + \lambda_1) \\ P_t &= \mathbf{w}_{t-1,1} E_t + G_t, \text{ and} & H &= \exp(-3\eta(\lambda_1 - \lambda_2)/4). \end{aligned}$$

Proof of Lemma E.4. For the simplicity of the proof, we do not derive the linearization from Taylor's expansion here and only prove the correctness of the statement. See the continuous analysis in [Section 3](#) for more intuition.

From [Equation E.2](#) in the linearization of Phase 1 (i.e., [Lemma 6.2](#)), we have

$$\mathbf{w}_{t,1} = \mathbf{w}_{t-1,1} + \eta \cdot [(\lambda_1 - \lambda_2) \mathbf{w}_{t-1,1} (1 - \mathbf{w}_{t-1,1}^2) + \mathbf{w}_{t-1,1} (A_t + B_t + C_t + E_t) + D_t] + \eta^2 Q_t.$$

Observe that $\mathbf{w}_{t-1,1} (1 - \mathbf{w}_{t-1,1}^2) = -(\mathbf{w}_{t-1,1} - 1)(\mathbf{w}_{t-1,1}^2 + \mathbf{w}_{t-1,1})$, we have

$$= \mathbf{w}_{t-1,1} + \eta \cdot [-(\lambda_1 - \lambda_2)(\mathbf{w}_{t-1,1} - 1)(\mathbf{w}_{t-1,1}^2 + \mathbf{w}_{t-1,1}) + \mathbf{w}_{t-1,1} (A_t + B_t + C_t + E_t) + D_t] + \eta^2 Q_t.$$

Now, by adding and subtracting $\frac{3}{4}\eta(-(\lambda_1 - \lambda_2)(\mathbf{w}_{t-1,1}))$, the equation becomes

$$= \mathbf{w}_{t-1,1} + \eta \cdot \left[-\frac{3}{4}(\lambda_1 - \lambda_2)(\mathbf{w}_{t-1,1} - 1) + \mathbf{w}_{t-1,1} (A_t + B_t + C_t + E_t) + D_t + G_t \right] + \eta^2 Q_t.$$

Finally, the lemma is concluded by replacing $\mathbf{w}_{t,1}$ and $\mathbf{w}_{t-1,1}$ with $\mathbf{w}'_{t,1} + 1$ and $\mathbf{w}'_{t-1,1}$ respectively and apply [Lemma 2.16](#) as we did in the proof of [Lemma 6.2](#). $\square$

The following lemma shows the nice martingale structure of the noise terms.

Lemma E.5 (Bounded difference and conditional variance in Phase 2). *For each $t \in \mathbb{N}$, we have the following.*

- (Bounded difference) For any $t \in \mathbb{N}$, $|F_t| = O(\sqrt{\mathbf{w}'_{t-1} + \eta})$, $Q_t = O(1)$ almost surely. If $\mathbf{w}_{t-1,1} \geq 0.5$, then $P_t \geq -O(\eta\lambda_1)$ almost surely.
- (Conditional expectation) For any $t \in \mathbb{N}$, $\mathbb{E}[F_t \mid \mathcal{F}_{t-1}] = 0$ and $|\mathbb{E}[Q_t \mid \mathcal{F}_{t-1}]| = O(\lambda_1)$.
- (Conditional variance) $\text{Var}[F_t \mid \mathcal{F}_{t-1}] \leq O(\lambda_1(\mathbf{w}'_{t-1} + \eta))$ and $\text{Var}[Q_t \mid \mathcal{F}_{t-1}] \leq O(\lambda_1)$.

Proof of Lemma E.5. The proof is based on a careful manipulation of Cauchy-Schwarz inequality.

- (Bounded difference) Let us start with bounding $|F_t| = |\mathbf{w}_{t-1,1}(A_t + B_t + C_t) + D_t|$. A straightforward application of Cauchy-Schwarz inequality and Lemma 5.1 would give $|B_t|, |C_t|, |D_t| = O((\|\mathbf{w}_{t-1}\|_2^2 - \mathbf{w}_{t-1,1}^2)^{1/2}) = O(\sqrt{\mathbf{w}'_{t-1} + \eta})$ almost surely. The bound for $|A_t|$ is slightly trickier as follows. First rewrite

$$A_t = -2 \sum_{i=2}^n \mathbf{x}_{t,1} \mathbf{x}_{t,i} \mathbf{w}_{t-1,1} \mathbf{w}_{t-1,i} - \left(\sum_{i=2}^n \mathbf{x}_{t,i} \mathbf{w}_{t-1,i} \right)^2 + \sum_{i=2}^n \mathbf{x}_{t,i}^2 \mathbf{w}_{t-1,i}^2. \quad (\text{E.6})$$

By Cauchy-Schwarz and Lemma 5.1, the last two terms can be bounded as follows.

$$O(\|\mathbf{w}_{t-1}\|_2^2 - \mathbf{w}_{t-1,1}^2) = O(\mathbf{w}_{t-1,1} + \eta) = O(\sqrt{\mathbf{w}'_{t-1} + \eta}).$$

As for the first term, we have

$$\begin{aligned} \sum_{i=2}^n \mathbf{x}_{t,1} \mathbf{x}_{t,i} \mathbf{w}_{t-1,1} \mathbf{w}_{t-1,i} &\leq \left(\sum_{i=2}^n \mathbf{x}_{t,1}^2 \mathbf{x}_{t,i}^2 \sum_{j=2}^n \mathbf{w}_{t-1,1}^2 \mathbf{w}_{t-1,i}^2 \right)^{1/2} \\ &\leq O((\|\mathbf{w}_{t-1}\|_2^2 - \mathbf{w}_{t-1,1}^2)^{1/2}) \\ &= O(\sqrt{\mathbf{w}'_{t-1} + \eta}) \end{aligned}$$

almost surely.

As for the P_t term, first observe that $E_t \geq 0$ trivially from the definition and

$$\begin{aligned} G_t &= (\lambda_1 - \lambda_2)(1 - \mathbf{w}_{t-1,1})(\mathbf{w}_{t-1,1}^2 + \mathbf{w}_{t-1,1} - \frac{3}{4}) \\ &\geq -(\lambda_1 - \lambda_2)O(\eta)(\mathbf{w}_{t-1,1}^2 + \mathbf{w}_{t-1,1} - \frac{3}{4}) = -O(\eta\lambda_1) \end{aligned}$$

where the inequality is due to $\|\mathbf{w}_t\|^2 - 1 = O(\eta)$ by Lemma 5.1 which implies that $1 - \mathbf{w}_{t-1,1} \geq -O(\eta)$ almost surely. As a result, we have

$$P_t = G_t + \mathbf{w}_{t-1,1} E_t \geq -O(\eta\lambda_1)$$

almost surely.

- (Conditional variance) It suffices to show that $\mathbb{E}[F_t^2 \mid \mathcal{F}_{t-1}] = O(\lambda_1(\mathbf{w}'_{t-1} + \eta)^2)$ and $\mathbb{E}[Q_t^2 \mid \mathcal{F}_{t-1}] = O(\lambda_1)$.

First, rewrite the A_t term as we did in Equation E.6. Consider the first term of A_t ,

$$\begin{aligned} \mathbb{E} \left[\sum_{i,j=2}^n \mathbf{x}_{t,1}^2 \mathbf{x}_{t,i} \mathbf{x}_{t,j} \mathbf{w}_{t-1,1}^2 \mathbf{w}_{t-1,i} \mathbf{w}_{t-1,j} \mid \mathcal{F}_{t-1} \right] &= \mathbb{E} \left[\mathbf{x}_{t,1}^2 \mathbf{w}_{t-1,1}^2 \cdot \left(\sum_{i=2}^n \mathbf{x}_{t,i} \mathbf{w}_{t-1,i} \right)^2 \mid \mathcal{F}_{t-1} \right] \\ (\because \text{Cauchy-Schwarz}) &\leq \mathbb{E} \left[\mathbf{x}_{t,1}^2 \mathbf{w}_{t-1,1}^2 \cdot \left(\sum_{i=2}^n \mathbf{x}_{t,i}^2 \sum_{i=2}^n \mathbf{w}_{t-1,i}^2 \right) \mid \mathcal{F}_{t-1} \right] \\ &\leq \mathbb{E} \left[O(\mathbf{x}_{t,1}^2 \mathbf{w}_{t-1,1}^2 (\mathbf{w}'_{t-1} + \eta)) \mid \mathcal{F}_{t-1} \right] \\ &\leq \lambda_1 (\mathbf{w}'_{t-1,1} + \eta). \end{aligned}$$

Similarly, the conditional variance of the last two terms and B_t, C_t, D_t are all $O(\mathbf{w}'_{t-1,1} + \eta)$ and thus $\text{Var}[F_t \mid \mathcal{F}_{t-1}] = O(\lambda_1(\mathbf{w}'_{t-1,1} + \eta))$.

As for the Q_t term, as $\mathbb{E}[x_{t,2}^4 \mid \mathcal{F}_{t-1}] \leq \lambda_1$, we have $\text{Var}[Q_t \mid \mathcal{F}_{t-1}] = O(\lambda_1)$.

□

E.3 Linearization in Appendix D

In this subsection, we provide the full proof for [Lemma D.6](#) on the linearization and [Lemma D.10](#) on the bounded difference, conditional expectation, and the conditional variance of the noise terms.

Proof of Lemma D.6

Proof of Lemma D.6. The proof is based on Taylor's expansion. Concretely,

$$f_{t,j}(\mathbf{w}_s) = f_{t,j}(\mathbf{w}_{s-1}) + \eta \sum_{i=1}^n \frac{\partial f_{t,j}}{\partial \mathbf{w}_i}(\mathbf{w}_{s-1}) \cdot \mathbf{z}_{s,i} + \eta^2 \sum_{i,i'=1}^n \frac{\partial^2 f_{t,j}}{\partial \mathbf{w}_i \partial \mathbf{w}_{i'}}(\bar{\mathbf{w}}_{s-1}) \cdot \mathbf{z}_{s,i} \mathbf{z}_{s,i'}.$$

Note that $\frac{\partial f_{t,j}(\mathbf{w})}{\partial \mathbf{w}_1} = -f_{t,j}(\mathbf{w})/\mathbf{w}_1$ and $\frac{\partial f_{t,j}(\mathbf{w})}{\partial \mathbf{w}_i} = \mathbf{1}_{i \leq j} \cdot \mathbf{x}_{t,i}/\mathbf{w}_1$ for $i = 2, \dots, n$. Denote the η^2 term as $\eta^2 C_{t,s,j}$, we have

$$= f_{t,j}(\mathbf{w}_{s-1}) - \eta \frac{f_{t,j}(\mathbf{w}_{s-1})}{\mathbf{w}_{s-1,1}} \cdot \mathbf{z}_{s,1} + \eta \frac{\sum_{i=2}^j \mathbf{x}_{s,i} \mathbf{z}_{s,i}}{\mathbf{w}_{s-1,1}} + \eta^2 C_{t,s,j}.$$

Next, recall that $\mathbb{E}[\mathbf{z}_{s,i} \mid \mathcal{F}_{s-1}] = (\lambda_i - \mathbf{w}_{s-1}^\top \text{diag}(\lambda) \mathbf{w}_{s-1}) \cdot \mathbf{w}_{s-1,i}$. Denote $B_{t,s,j} = \sum_{i=1}^n \partial_{\mathbf{w}_i} f_{t,j}(\mathbf{w}_{s-1}) \cdot (\mathbf{z}_{s,i} - \mathbb{E}[\mathbf{z}_{s,i} \mid \mathcal{F}_{t-1}])$. By adding and subtracting the expectations, the equation becomes

$$\begin{aligned} &= f_{t,j}(\mathbf{w}_{s-1}) - \eta \lambda_1 f_{t,j}(\mathbf{w}_{s-1}) + \eta \frac{\sum_{i=2}^j \lambda_i \mathbf{x}_{s,i} \mathbf{w}_{s-1,i}}{\mathbf{w}_{s-1,1}} \\ &+ \eta (\mathbf{w}_{s-1}^\top \text{diag}(\lambda) \mathbf{w}_{s-1}) \cdot \left(f_{t,j}(\mathbf{w}_{s-1}) - \frac{\sum_{i=2}^j \mathbf{x}_{s,i} \mathbf{w}_{s-1,i}}{\mathbf{w}_{s-1,1}} \right) + \eta B_{t,s,j} + \eta^2 C_{t,s,j} \end{aligned}$$

Observe that the two terms in the parenthesis becomes 0 after cancelling out with each other. Finally, by adding and subtracting $\eta \lambda_i f_{t,i}(\mathbf{w}_{s-1})$ for each $i = 2, 3, \dots, j$, we have

$$= (1 - \eta(\lambda_1 - \lambda_j)) \cdot f_{t,j}(\mathbf{w}_{s-1}) + \eta \sum_{i=2}^{j-1} (\lambda_i - \lambda_{i+1}) f_{t,i}(\mathbf{w}_{s-1}) + \eta B_{t,s,j} + \eta^2 C_{t,s,j}.$$

Finally, denote $A_{t,s,j} = \sum_{i=2}^{j-1} (\lambda_i - \lambda_{i+1}) f_{t,i}(\mathbf{w}_{s-1})$, we have

$$= (1 - \eta(\lambda_1 - \lambda_j)) \cdot f_{t,j}(\mathbf{w}_{s-1}) + \eta A_{t,s,j} + \eta B_{t,s,j} + \eta^2 C_{t,s,j}.$$

□

Proof of Lemma D.10 It is not difficult to see that $f_{t,j}(\mathbf{w}_s)$ can be nicely bounded by $O(\Lambda)$ as follows, though the bound is not good enough.

Lemma E.7. For any $t \in [T]$, $s \in [t]$, $j \in [n]$, and $c \in [0, 1]$, we have,

$$|f_{t,j}(\mathbf{w}_{(s \wedge \tau_t)-1} + c\eta \mathbf{z}_s)| = O(\Lambda)$$

almost surely.

The reason why we consider $c\eta\mathbf{z}_s$ is that it shows up later in $C_{t,s,j}$ due to the mean value theorem used in the linearization. Note that ideally we want $|f_{t,j}(\mathbf{w}_s)| \leq \Lambda$. To get the bound, we need to apply the ODE trick which requires the concentration of $B_{t,s,j}$, and $C_{t,s,j}$. Recall that this is exactly what we have done in [Section D.3](#).

First, we expand the differential terms in $B_{t,s,j}$, $C_{t,s,j}$ and rewrite them as follows.

Lemma E.8 (Rewrite $B_{t,s,j}$ and $C_{t,s,j}$). *For any $t \in [T]$, $s \in [t-1]$, and $j \in [n]$, we have*

$$\begin{aligned} B_{t,s,j} &= f_{t,j}(\mathbf{w}_{s-1}) \cdot (-\mathbf{x}_{s,1}^2 - f_{s,n}(\mathbf{w}_{s-1})\mathbf{x}_{s,1} + \lambda_1) \\ &\quad + \sum_{i=2}^j f_{s,n}(\mathbf{w}_{s-1})\mathbf{x}_{t,i}\mathbf{x}_{s,i} + \sum_{i=2}^j (\lambda_i - \lambda_{i+1})f_{t,i}(\mathbf{w}_{s-1}) \end{aligned}$$

and

$$C_{t,s,j} = \frac{2f_{t,j}(\mathbf{w}_{s-1} + c\eta\mathbf{z}_s)\mathbf{z}_{s,1}^2 - \sum_{i=2}^j \mathbf{x}_{t,i}\mathbf{z}_{s,i}\mathbf{z}_{s,1}}{\mathbf{w}_{s-1,1}^2 [1 + c\eta(\mathbf{x}_{s,1} + f_{s,n}(\mathbf{w}_{s-1}))(\mathbf{x}_{s,1} - y_s\mathbf{w}_{s-1,1})]^2}. \quad (\text{E.9})$$

Now, the bounded difference, conditional expectation, and the conditional variance of $B_{t,s,j}$ and $C_{t,s,j}$ can be easily proved with the help of [Lemma E.8](#).

Proof of [Lemma D.10](#). The bounded difference immediately follows from [Lemma E.8](#) while the conditional expectation and the conditional variance require some small extra cares.

- (Bounded difference) For the $B_{t,s \wedge \xi \wedge \tau_t,j}$ term, from [Lemma E.7](#), we have

$$|f_{t,j}(\mathbf{w}_{(s \wedge \tau_t)-1})|, |f_{s,n}(\mathbf{w}_{(s \wedge \tau_t)-1})| = O(\Lambda)$$

almost surely. Thus, by [Lemma E.8](#), we have

$$|B_{t,s \wedge \xi \wedge \tau_t,j}| \leq (\Lambda^2) + \Lambda \cdot O\left(\sum_{i=2}^j \mathbf{x}_{t,i}\mathbf{x}_{s,i} + \sum_{i=2}^j (\lambda_i - \lambda_{i+1})\right).$$

By Cauchy-Schwarz inequality and telescoping sum, the parenthesis can be bounded by $O(1)$ and thus we have $|B_{t,s \wedge \xi \wedge \tau_t,j}| = O(\Lambda^2)$ almost surely.

As for $C_{t,s,j}$, from [Lemma E.7](#) and Cauchy-Schwarz inequality, we can bound the numerator in [Equation E.9](#) by $O(y_s^2\Lambda)$ almost surely where the y_s^2 terms are borrowed from the $\mathbf{z}$ terms. To see the denominator cannot be too small, first by the choice of η, Λ , and [Lemma E.7](#), we have

$$c\eta(\mathbf{x}_{s,1} + f_{s,n}(\mathbf{w}_{(s \wedge \tau_t)-1})) = O(\lambda_1) \quad (\text{E.10})$$

almost surely. Thus, the bracket in the denominator of [Equation E.9](#) would be $\Omega(1)$. Next, as $|f_{t,n}(\mathbf{w}_{(s \wedge \tau_t)-1})| = O(\Lambda)$ almost surely by [Lemma E.7](#), we have

$$|C_{t,s \wedge \xi \wedge \tau_t}| = O\left(\frac{y_s^2\Lambda}{\mathbf{w}_{(s \wedge \tau_t)-1,1}^2}\right) = O(f_t(\mathbf{w}_{(s \wedge \tau_t)-1})^2\Lambda) = O(\Lambda^3).$$

- (Conditional expectation) Recall the original definition of $B_{t,s,j}$ in [Lemma D.6](#), it is naturally a martingale. Thus, its stopped process $\{B_{t,s \wedge \xi \wedge \tau_t,j}\}$ is also a martingale and the conditional expectation of the stopped process is zero.

As for the $C_{t,s\wedge\xi\wedge\tau_t,j}$ term, let us start with rewriting the numerator of Equation E.9 by replacing $\mathbf{z}_{s,i}$ with $y_s(\mathbf{x}_{s,i} - y_s\mathbf{w}_{s-1,i})$ as follows.

$$\begin{aligned} & 2y_s^2 f_{t,j}(\mathbf{w}_{s-1} + c\eta\mathbf{z}_s)(\mathbf{x}_{s,1} - y_s\mathbf{w}_{s-1,1})^2 - \sum_{i=2}^j y_s^2 \mathbf{x}_{t,i}(\mathbf{x}_{s,i} - y_s\mathbf{w}_{s-1,i})(\mathbf{x}_{s,1} - y_s\mathbf{w}_{s-1,1}) \\ &= 2y_s^2 f_{t,j}(\mathbf{w}_{s-1} + c\eta\mathbf{z}_s)(\mathbf{x}_{s,1} - y_s\mathbf{w}_{s-1,1})^2 - \sum_{i=2}^j y_s^2 \mathbf{x}_{t,i} \mathbf{x}_{s,i} \mathbf{x}_{s,1} \\ &+ \sum_{i=2}^j y_s^3 \mathbf{x}_{t,i} \mathbf{x}_{s,i} \mathbf{w}_{s-1,1} + \sum_{i=2}^j y_s^4 \mathbf{x}_{t,i} \mathbf{w}_{s-1,i} (\mathbf{x}_{s,1} - y_s\mathbf{w}_{s-1,1}). \end{aligned}$$

Let $D = [1 + c\eta(\mathbf{x}_{s\wedge\xi\wedge\tau_t,1} + f_{s\wedge\xi\wedge\tau_t,n}(\mathbf{w}_{(s\wedge\xi\wedge\tau_t)-1}))(\mathbf{x}_{s\wedge\xi\wedge\tau_t,1} - y_s\mathbf{w}_{(s\wedge\xi\wedge\tau_t)-1,1})]^2$ be the bracket term in the denominator of Equation E.9, from the previous discussion we know that $|D| = \Theta(1)$ almost surely. Now, Consider the first term combining with the denominator, we have

$$\begin{aligned} & \left| \mathbb{E} \left[\frac{2y_s^2 f_{t,j}(\mathbf{w}_{(s\wedge\xi\wedge\tau_t)-1} + c\eta\mathbf{z}_{s\wedge\xi\wedge\tau_t})(\mathbf{x}_{s\wedge\xi\wedge\tau_t,1} - y_s\mathbf{w}_{(s\wedge\xi\wedge\tau_t)-1,1})^2}{\mathbf{w}_{(s\wedge\xi\wedge\tau_t)-1,1}^2 D} \mid \mathcal{F}_{s-1}^{(t)} \right] \right| \\ &= O \left(\Lambda^3 \cdot \mathbb{E}[(\mathbf{x}_{s\wedge\xi\wedge\tau_t,1} - y_s\mathbf{w}_{(s\wedge\xi\wedge\tau_t)-1,1})^2 \mid \mathcal{F}_{s-1}^{(t)}] \right) = O(\lambda_1 \Lambda^3). \end{aligned}$$

Note that the first equality holds because the quantity in the expectation is always non-negative. Next, consider the second term as follows.

$$\begin{aligned} & \left| \mathbb{E} \left[\frac{\sum_{i=2}^j y_s^2 f_{t,i}(\mathbf{x}_{s,i} - y_s\mathbf{w}_{s-1,i})(\mathbf{x}_{s,1} - y_s\mathbf{w}_{s-1,1})}{\mathbf{w}_{(s\wedge\xi\wedge\tau_t)-1,1}^2 D} \mid \mathcal{F}_{s-1}^{(t)} \right] \right| \leq \left| \mathbb{E} \left[\frac{\sum_{i=2}^j y_s^2 \mathbf{x}_{t,i}(\mathbf{x}_{s\wedge\xi\wedge\tau_t,i}^2 + \mathbf{x}_{s\wedge\xi\wedge\tau_t,1}^2)}{\mathbf{w}_{(s\wedge\xi\wedge\tau_t)-1,1}^2 D} \mid \mathcal{F}_{s-1}^{(t)} \right] \right| \\ & \leq O(\Lambda^2 \cdot \mathbb{E}[\mathbf{x}_{s\wedge\xi\wedge\tau_t,i}^2 + \mathbf{x}_{s\wedge\xi\wedge\tau_t,1}^2]) = O(\lambda_1 \Lambda^2). \end{aligned}$$

The third term can be upper bounded as follows.

$$\left| \mathbb{E} \left[\frac{\sum_{i=2}^j y_s^3 \mathbf{x}_{t,i} \mathbf{x}_{s\wedge\xi\wedge\tau_t,i} \mathbf{w}_{(s\wedge\xi\wedge\tau_t)-1,1}}{\mathbf{w}_{(s\wedge\xi\wedge\tau_t)-1,1}^2 D} \mid \mathcal{F}_{s-1}^{(t)} \right] \right| = O(\Lambda \cdot \mathbb{E}[y_{s\wedge\xi\wedge\tau_t}^2 \mid \mathcal{F}_{s-1}^{(t)}]) = O(\lambda_1 \Lambda).$$

The last term can be upper bounded as follows.

$$\begin{aligned} & \left| \mathbb{E} \left[\frac{\sum_{i=2}^j y_s^4 \mathbf{x}_{t,i} \mathbf{w}_{(s\wedge\xi\wedge\tau_t)-1,i} (\mathbf{x}_{s\wedge\xi\wedge\tau_t,1} - y_s\mathbf{w}_{(s\wedge\xi\wedge\tau_t)-1,1})}{\mathbf{w}_{(s\wedge\xi\wedge\tau_t)-1,1}^2 D} \mid \mathcal{F}_{s-1}^{(t)} \right] \right| \\ &= O(\Lambda^2 \cdot \mathbb{E}[y_{s\wedge\xi\wedge\tau_t}^2 \mid \mathcal{F}_{s-1}^{(t)}]) = O(\lambda_1 \Lambda^2). \end{aligned}$$

We conclude that $|\mathbb{E}[C_{t,s\wedge\xi\wedge\tau_t,j} \mid \mathcal{F}_{s-1}^{(t)}]| = O(\lambda_1 \Lambda^3)$.

- (Conditional variance) The conditional variance of $B_{t,s\wedge\xi\wedge\tau_t,j}$ and $C_{t,s\wedge\xi\wedge\tau_t,j}$ can be upper bounded using the same argument as we did in the calculation of conditional expectation. Thus, we omit the details here for simplicity.

□

Proof of Lemma E.8 Let us complete this subsection with the proof for the lemma that rewrites $B_{t,s,j}$ and $C_{t,s,j}$.

Proof of Lemma E.8. Recall that

$$B_{t,s,j} = \sum_{i=1}^n \frac{\partial f_{t,j}}{\partial \mathbf{w}_i}(\mathbf{w}_{s-1}) \cdot (\mathbf{z}_{s,i} - \mathbb{E}[\mathbf{z}_{s,i} \mid \mathcal{F}_{s-1}]) \quad \text{and} \quad C_{t,s,j} = \sum_{i,i'=1}^n \frac{\partial^2 f_{t,j}}{\partial \mathbf{w}_i \partial \mathbf{w}_{i'}}(\mathbf{w}_{s-1}) \cdot \mathbf{z}_{s,i} \mathbf{z}_{s,i'}$$

For simplicity, in the following we denote $\mathbb{E}_{t-1}[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$. Observe that

$$\frac{\partial f_{t,j}(\mathbf{w})}{\partial \mathbf{w}_i} = \begin{cases} \frac{-f_{t,j}(\mathbf{w})}{\mathbf{w}_1^2} & , \text{ if } i = 1 \\ \frac{\mathbf{x}_{t,i}}{\mathbf{w}_1} & , \text{ if } i = 2, 3, \dots, j \end{cases}$$

and

$$\frac{\partial^2 f_{t,j}(\mathbf{w})}{\partial \mathbf{w}_i \partial \mathbf{w}_{i'}} = \begin{cases} \frac{2f_{t,j}(\mathbf{w})}{\mathbf{w}_1^2} & , \text{ if } i = i' = 1 \\ \frac{-\mathbf{x}_{t,i'}}{\mathbf{w}_1^2} & , \text{ if } i = 1 \text{ and } i' = 2, 3, \dots, j \\ \frac{-\mathbf{x}_{t,i}}{\mathbf{w}_1^2} & , \text{ if } i = 2, 3, \dots, j \text{ and } i' = 1 \\ 0 & \text{ else.} \end{cases}$$

Let us start with rewriting $B_{t,s,j}$ as follows.

$$B_{t,s,j} = -\frac{-f_{t,j}(\mathbf{w}_{s-1}) \cdot (\mathbf{z}_{s,1} - \mathbb{E}_{s-1}[\mathbf{z}_{s,1}])}{\mathbf{w}_{s-1,1}} + \frac{\sum_{i=2}^j \mathbf{x}_{t,i} \cdot (\mathbf{z}_{s,i} - \mathbb{E}_{s-1}[\mathbf{z}_{s,i}])}{\mathbf{w}_{s-1,1}}.$$

Recall that $\mathbf{z}_{s,i} = y_s(\mathbf{x}_{s,i} - y_s \mathbf{w}_{s-1,i})$ and $\mathbb{E}_{s-1}[\mathbf{z}_{s,i}] = (\lambda_i - \mathbf{w}_{s-1}^\top \text{diag}(\lambda) \mathbf{w}_{s-1}) \cdot \mathbf{w}_{s-1,i}$. The equation then becomes

$$\begin{aligned} &= \frac{-f_{t,j}(\mathbf{w}_{s-1}) y_s \mathbf{x}_{s,1}}{\mathbf{w}_{s-1,1}} + (y_s^2 + \lambda_1 - \mathbf{w}_{s-1}^\top \text{diag}(\lambda) \mathbf{w}_{s-1}) \cdot f_{t,j}(\mathbf{w}_{s-1}) \\ &+ \frac{\sum_{i=2}^j y_s \mathbf{x}_{t,i} \mathbf{x}_{s,i} - (y_s^2 + \lambda_i - \mathbf{w}_{s-1}^\top \text{diag}(\lambda) \mathbf{w}_{s-1}) \cdot \mathbf{x}_{t,i} \mathbf{w}_{s-1,i}}{\mathbf{w}_{s-1,1}}. \end{aligned}$$

Recall that $y_s = \sum_{i=1}^n \mathbf{x}_{s,i} \mathbf{w}_{s-1,i} = \mathbf{x}_{s,1} \mathbf{w}_{s-1,1} + f_{s,n}(\mathbf{w}_{s-1}) \mathbf{w}_{s-1,1}$. The equation becomes

$$= f_{t,j}(\mathbf{w}_{s-1}) \cdot (-\mathbf{x}_{s,1}^2 - f_{s,n}(\mathbf{w}_{s-1}) \mathbf{x}_{s,1} + \lambda_1) + \sum_{i=2}^j f_{s,n}(\mathbf{w}_{s-1}) \mathbf{x}_{t,i} \mathbf{x}_{s,i} + \sum_{i=2}^j (\lambda_i - \lambda_{i+1}) f_{t,i}(\mathbf{w}_{s-1}).$$

Note that now $B_{t,s,j}$ only depends on $f_{t',j'}(\mathbf{w}_{s-1})$ for any $t' \in [t]$ and $j' \in [n]$.

Next, let us rewrite $C_{t,s,j}$ by expanding the second order derivative as follows.

$$C_{t,s,j} = \frac{2f_{t,j}(\bar{\mathbf{w}}_{s-1}) \mathbf{z}_{s,1}^2}{\bar{\mathbf{w}}_{s-1,1}^2} - \frac{\sum_{i=2}^j \mathbf{x}_{t,i} \mathbf{z}_{s,i} \mathbf{z}_{s,1}}{\bar{\mathbf{w}}_{s-1,1}^2}.$$

By mean value theorem, we can replace $\bar{\mathbf{w}}_{s-1}$ with $\mathbf{w}_{s-1} + c\eta \mathbf{z}_s$ for some $c \in [0, 1]$ and by definition, we can replace the $\mathbf{z}_s$ in the denominator with $y_s(\mathbf{x}_s - y_s \mathbf{w}_{s-1})$. The equation becomes

$$= \frac{2f_{t,j}(\mathbf{w}_{s-1} + c\eta \mathbf{z}_s) \mathbf{z}_{s,1}^2 - \sum_{i=2}^j \mathbf{x}_{t,i} \mathbf{z}_{s,i} \mathbf{z}_{s,1}}{(\mathbf{w}_{s-1,1} + c\eta y_s(\mathbf{x}_{s,1} - y_s \mathbf{w}_{s-1,1}))^2}.$$

Next, replace y_s with $\mathbf{x}_{s,1} \mathbf{w}_{s-1,1} + f_{s,n}(\mathbf{w}_{s-1}) \mathbf{w}_{s-1,1}$ as we did before and get

$$= \frac{2f_{t,j}(\mathbf{w}_{s-1} + c\eta \mathbf{z}_s) \mathbf{z}_{s,1}^2 - \sum_{i=2}^j \mathbf{x}_{t,i} \mathbf{z}_{s,i} \mathbf{z}_{s,1}}{\mathbf{w}_{s-1,1}^2 [1 + c\eta(\mathbf{x}_{s,1} + f_{s,n}(\mathbf{w}_{s-1}))(\mathbf{x}_{s,1} - y_s \mathbf{w}_{s-1,1})]^2}.$$

Note that now $C_{t,s,j}$ is under control in the sense that the numerator is bounded by $O(\Lambda)$ and the denominator is $\Omega(\mathbf{w}_{s-1,1}^2)$ under the stopping time condition. $\square$

F Proof of Theorem 4.2

Let us first state a corollary of Lemma 7.4 as follows.

Corollary F.1. For any $\epsilon, \delta', \eta \in (0, 0.5)$ such that $\eta = O\left(\frac{\epsilon(\lambda_1 - \lambda_2)}{\lambda_1 \log \frac{1}{\delta'}}\right)$, let $t_3 \in \mathbb{N}$ be

$$t_3 = \frac{4}{3} \cdot \frac{1}{\eta(\lambda_1 - \lambda_2)} + \Theta\left(\frac{1}{\eta(\lambda_1 - \lambda_2) \log \frac{1}{\epsilon}}\right).$$

Suppose there exists some $t_0 \in \mathbb{N}$ such that $\mathbf{w}_{t_0,1} \geq 1 - \frac{\epsilon}{2}$. Then,

$$H^{t_3} = \exp\left(-\frac{3}{4}\eta(\lambda_1 - \lambda_2)t_3\right) \leq \frac{1}{10}$$

and

$$\Pr\left[\min_{1 \leq t \leq t_3} \sum_{i=t_0+1}^{t_0+t} \frac{\eta F_i + \eta P_i + \eta^2 Q_i}{H^{i-t_0}} < -\frac{\epsilon}{2}\right] < \delta'.$$

Proof of Corollary F.1. This is an immediate corollary of Lemma 7.4 by setting $\epsilon' = \epsilon$, $k = \log \log(1/\epsilon)$, and $\gamma = \epsilon$. $\square$

Proof of Theorem 4.2. The first statement of the theorem is an immediate corollary of the second so we only focus on the for-all-time guarantee. The proof is based on an inductive argument. Concretely, let t_3 be the time parameter from Corollary F.1 with $\delta' = \delta/10s^2$.

Claim F.2. For any $s \in \mathbb{N}$, denote $t_0 = (s-1)t_3$ we have

$$\Pr\left[\min_{1 \leq t \leq t_3} \mathbf{w}_{t_0+t,1} < 1 - \epsilon \bigvee \mathbf{w}_{t_0+t_3,1} < 1 - \frac{\epsilon}{2} \mid \mathbf{w}_{t_0,1} \geq 1 - \frac{\epsilon}{2}\right] < \frac{\delta}{10s^2}.$$

Proof of Claim F.2. For any $s \in \mathbb{N}$, let $t_0 = (s-1)t_3$ and condition on the event $\{\mathbf{w}_{t_0,1} \geq 1 - \epsilon/2\}$. Next, invoke Corollary F.1 with $\delta' = \frac{\delta}{10s^2}$. Check that $10 \log \frac{t_0+t}{\delta} \geq 10 \log \frac{s}{\delta} \geq \log \frac{10s^2}{\delta} = \log \frac{1}{\delta'}$. Thus,

$$\eta_{t_0+t} = O\left(\frac{\epsilon(\lambda_1 - \lambda_2)}{\lambda_1 \log \frac{t_0+t}{\delta}}\right) = O\left(\frac{\epsilon(\lambda_1 - \lambda_2)}{\lambda_1 \log \frac{1}{\delta'}}\right)$$

for all $1 \leq t \leq t_3$. Then, by Lemma 7.4 we have

$$\Pr\left[\min_{1 \leq t \leq t_3} \sum_{i=t_0+1}^{t_0+t} \frac{\eta F_i + \eta P_i + \eta^2 Q_i}{H^{i-t_0}} < -\frac{\epsilon}{2} \mid \mathbf{w}_{t_0,1} \geq 1 - \frac{\epsilon}{2}\right] < \frac{\delta}{10s^2}$$

and $H^{t_3} \leq \frac{1}{10}$. For any $1 \leq t \leq t_3$, recall that $\mathbf{w}'_{t_0+t,1} = \mathbf{w}_{t_0+t,1} - 1$. Now, apply the ODE trick in Phase 2 (i.e., Corollary 7.3), we have

$$\mathbf{w}'_{t_0+t,1} \geq H^t \cdot \left(\mathbf{w}'_{t_0,1} + \sum_{i=t_0+1}^{t_0+t} \frac{\eta F_i + \eta P_i + \eta^2 Q_i}{H^{i-t_0}}\right).$$

Conditioning on the event $\{\mathbf{w}_{t_0,1} \geq 1 - \epsilon/2\}$, with probability at least $1 - \frac{\delta}{10s^2}$, the inequality becomes

$$\geq H^t \cdot \left(-\frac{\epsilon}{2} - \frac{\epsilon}{2}\right) \geq \begin{cases} -\epsilon & , 1 \leq t < t_3 \\ -\frac{\epsilon}{2} & , t = t_3. \end{cases}$$

This completes the proof of the claim. $\square$

By the chain rule for conditional probability, Claim F.2 gives

$$\begin{aligned} & \Pr[\exists t \geq 0, \mathbf{w}_{t,1} < 1 - \epsilon] \\ & \leq \sum_{s=0}^{\infty} \Pr\left[\min_{(s-1)t_3+1 \leq t \leq s \cdot t_3} \mathbf{w}_{t,1} < 1 - \epsilon \bigvee \mathbf{w}_{s \cdot t_3,1} < 1 - \frac{\epsilon}{2} \mid \mathbf{w}_{(s-1)t_3} \geq 1 - \frac{\epsilon}{2}\right] \\ & \leq \sum_{s=0}^{\infty} \frac{\delta}{10s^2} \leq \delta. \end{aligned}$$

This completes the proof of Theorem 4.2. $\square$

References

- [A95] Wandell B A. *Foundations of Vision*. Sunderland, MA: Sinauer, 1995.
- [ACS13] Raman Arora, Andy Cotter, and Nati Srebro. Stochastic optimization of pca with capped msg. In *Advances in Neural Information Processing Systems*, pages 1815–1823, 2013.
- [AN00] L F Abbott and Sacha B Nelson. Synaptic plasticity: taming the beast. *Nature Neuroscience*, 3:1178–1183, 2000.
- [Apa12] Vladimir Aparin. Simple modification of oja rule limits l_1 -norm of weight vector and leads to sparse connectivity. *Neural computation*, 24(3):724–743, 2012.
- [AZL17] Zeyuan Allen-Zhu and Yuanzhi Li. First efficient convergence for streaming k-pca: a global, gap-free, and near-optimal rate. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 487–492. IEEE, 2017.
- [Azu67] Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.
- [BCM82] Elie L Bienenstock, Leon N Cooper, and Paul W Munro. Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience*, 2(1):32–48, 1982.
- [BDWY16] Maria-Florina Balcan, Simon Shaolei Du, Yining Wang, and Adams Wei Yu. An improved gap-dependency analysis of the noisy power method. In *Conference on Learning Theory*, pages 284–309, 2016.
- [CCL19] Chi-Ning Chou, Kai-Min Chung, and Chi-Jen Lu. On the algorithmic power of spiking neural networks. In *10th Innovations in Theoretical Computer Science Conference, ITCS 2019, January 10-12, 2019, San Diego, California, USA*, pages 26:1–26:20, 2019.
- [CG90] Pierre Comon and Gene H Golub. Tracking a few extreme singular values and vectors in signal processing. *Proceedings of the IEEE*, 78(8):1327–1343, 1990.
- [CKS96] Andrzej Cichocki, Włodzimierz Kasprzak, and Władysław Skarbek. Adaptive learning algorithm for principal component analysis with partial data. *CYBERNETICS AND SYSTEMS RESEARCH*, pages 1014–1019, 1996.
- [CL94] Hong Chen and Ruey-Wen Lin. An on-line unsupervised learning machine for adaptive feature extraction. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 41(2):87–98, 1994.
- [DK96] Konstantinos I Diamantaras and Sun Yuan Kung. *Principal component neural networks: theory and applications*. John Wiley & Sons, Inc., 1996.
- [DSOR15] Christopher De Sa, Kunle Olukotun, and Christopher Ré. Global convergence of stochastic gradient descent for some non-convex matrix problems. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37, ICML’15*, pages 2332–2341. JMLR.org, 2015.
- [Duf13] Marie Duflo. *Random iterative models*, volume 34. Springer Science & Business Media, 2013.
- [Fre75] David A Freedman. On tail probabilities for martingales. *The Annals of Probability*, pages 100–118, 1975.
- [GS12] Surya Ganguli and Haim Sompolinsky. Compressed Sensing, Sparsity and neural data. *Annual review of neuroscience*, 35(1):463–483, 2012.
- [Heb49] Donald O. Hebb. *The organization of behavior: A neuropsychological theory*. Wiley, New York, June 1949.

- [HKP91] John Hertz, Anders Krogh, and Richard G Palmer. Introduction to the theory of neural computation. *Santa Fe Institute Studies in the Sciences of Complexity; Lecture Notes, Redwood City, Ca.: Addison-Wesley, 1991*, 1991.
- [HMRAR13] Michel Habib, Colin McDiarmid, Jorge Ramirez-Alfonsin, and Bruce Reed. *Probabilistic methods for algorithmic discrete mathematics*, volume 16. Springer Science & Business Media, 2013.
- [Hot33] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.
- [HP94] George F Harpur and Richard W Prager. *Experiments with simple Hebbian-based learning rules in pattern classification tasks*. Citeseer, 1994.
- [HP14] Moritz Hardt and Eric Price. The noisy power method: A meta algorithm with applications. In *Advances in Neural Information Processing Systems*, pages 2861–2869, 2014.
- [HP19] Yael Hitron and Merav Parter. Counting to ten with two fingers: Compressed counting with spiking neurons. *arXiv preprint arXiv:1902.10369*, 2019.
- [HV98] M Haft Hemmen and J L Van. Theory and implementation of infomax filters for the retina. *Network: Computation in Neural Systems*, 9(1):39–71, 1998.
- [JJK⁺16] Prateek Jain, Chi Jin, Sham M Kakade, Praneeth Netrapalli, and Aaron Sidford. Streaming pca: Matching matrix bernstein and near-optimal finite sample guarantees for oja’s algorithm. In *Conference on learning theory*, pages 1147–1164, 2016.
- [Kar96] Nicolaos B Karayiannis. Accelerating the training of feedforward neural networks using generalized hebbian rules for initializing the internal representations. *IEEE transactions on neural networks*, 7(2):419–426, 1996.
- [KDT94] Sun-Yuan Kung, Konstantinos I Diamantaras, and Jin-Shiuh Taur. Adaptive principal component extraction (apex) and applications. *IEEE Transactions on Signal Processing*, 42(5):1202–1217, 1994.
- [LG16] Jean-François Le Gall. *Brownian motion, martingales, and stochastic calculus*, volume 274. Springer, 2016.
- [LM18] Nancy Lynch and Cameron Musco. A basic compositional model for spiking neural networks. *arXiv preprint arXiv:1808.03884*, 2018.
- [LMP17a] Nancy Lynch, Cameron Musco, and Merav Parter. Spiking neural networks: An algorithmic perspective. In *Workshop on Biological Distributed Algorithms (BDA), July 28th, 2017, Washington DC, USA*, 2017.
- [LMP17b] Nancy A. Lynch, Cameron Musco, and Merav Parter. Computational tradeoffs in biological neural networks: Self-stabilizing winner-take-all networks. In *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA*, pages 15:1–15:44, 2017.
- [LMP17c] Nancy A. Lynch, Cameron Musco, and Merav Parter. Neuro-ram unit with applications to similarity testing and compression in spiking neural networks. In *31st International Symposium on Distributed Computing, DISC 2017, October 16-20, 2017, Vienna, Austria*, pages 33:1–33:16, 2017.
- [LMPV18] Robert A. Legenstein, Wolfgang Maass, Christos H. Papadimitriou, and Santosh Srinivas Vempala. Long term memory and the densest k-subgraph problem. In *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA*, pages 57:1–57:15, 2018.

- [LMT19] Nancy Lynch and Frederik Mallmann-Trenn. Learning hierarchically structured concepts. *arXiv preprint arXiv:1909.04559*, 2019.
- [LTYH09] Jian Cheng Lv, Kok Kiong Tan, Zhang Yi, and Sunan Huang. A family of fuzzy learning algorithms for robust principal component analysis neural networks. *IEEE Transactions on Fuzzy Systems*, 18(1):217–226, 2009.
- [LWLZ18] Chris Junchi Li, Mengdi Wang, Han Liu, and Tong Zhang. Near-optimal stochastic approximation for online principal component estimation. *Mathematical Programming*, 167(1):75–97, 2018.
- [OBL00] Shan Ouyang, Zheng Bao, and Gui-Sheng Liao. Robust recursive least squares learning algorithm for principal component analysis. *IEEE Transactions on Neural Networks*, 11(1):215–221, 2000.
- [Oja82] Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15(3):267–273, 1982.
- [Oja92] Erkki Oja. Principal components, minor components, and linear neural networks. *Neural networks*, 5(6):927–935, 1992.
- [OK85] Erkki Oja and Juha Karhunen. On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of mathematical analysis and applications*, 106(1):69–84, 1985.
- [Pea01] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572, 1901.
- [Peh19] Cengiz Pehlevan. A spiking neural network with local learning rules derived from nonnegative similarity matching. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 7958–7962, 2019.
- [PHC15] Cengiz Pehlevan, Tao Hu, and Dmitri B Chklovskii. A hebbian/anti-hebbian neural network for linear subspace learning: A derivation from multidimensional scaling of streaming data. *Neural computation*, 27(7):1461–1495, 2015.
- [Plu95] Mark D Plumbly. Lyapunov functions for convergence of principal component algorithms. *Neural Networks*, 8(1):11–23, 1995.
- [PV18] Christos H Papadimitriou and Santosh S Vempala. Random projection in the brain and computation with assemblies of neurons. In *10th Innovations in Theoretical Computer Science Conference (ITCS 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.
- [SA06] Christian D Swinehart and LF Abbott. Dimensional reduction for reward-based learning. *Network: Computation in Neural Systems*, 17(3):235–252, 2006.
- [San89] Terence D Sanger. Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural networks*, 2(6):459–473, 1989.
- [SCL19] Lili Su, Chia-Jung Chang, and Nancy Lynch. Spike-based winner-take-all computation: Fundamental limits and order-optimal circuits. *arXiv preprint arXiv:1904.10399*, 2019.
- [Sej77] T. Sejnowski. Storing covariance with nonlinearly interacting neurons. *J. Mathematical Biology*, 4:303–321, 1977.
- [Sha16] Ohad Shamir. Convergence of stochastic gradient descent for pca. In *International Conference on Machine Learning*, pages 257–265, 2016.

- [SLY06] Lifeng Shang, Jian Cheng Lv, and Zhang Yi. Rigid medical image registration using pca neural network. *Neurocomputing*, 69(13-15):1717–1722, 2006.
- [WBD77] Snyder A W, Laughlin S B, and Stavenga D. Information capacity of the eye. *Vision Res.*, 17:1163–75, 1977.
- [WC72] Hugh R Wilson and Jack D Cowan. Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical journal*, 12(1):1–24, 1972.
- [WC73] Hugh R Wilson and Jack D Cowan. A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetik*, 13(2):55–80, 1973.
- [Wil91] David Williams. *Probability with martingales*. Cambridge university press, 1991.
- [WL19] Mien Brabeeba Wang and Nancy Lynch. Integrating temporal information to spatial information in a neural circuit. *arXiv preprint arXiv:1903.01217*, 2019.
- [XOS92] Lei Xu, Erkki Oja, and Ching Y Suen. Modified hebbian learning for curve and surface fitting. *Neural Networks*, 5(3):441–457, 1992.
- [YYLT05] Zhang Yi, Mao Ye, Jian Cheng Lv, and Kok Kiong Tan. Convergence analysis of a deterministic discrete time system of oja’s pca learning algorithm. *IEEE Transactions on Neural Networks*, 16(6):1318–1328, 2005.
- [Zuf02] Pedro J Zufiria. On the discrete-time dynamics of the basic hebbian neural network node. *IEEE Transactions on Neural Networks*, 13(6):1342–1352, 2002.