Efficient Streaming PCA in Biological Neural Networks: An ODE-Inspired Convergence Framework

Chris Junchi Li[◇]

Department of Electrical Engineering and Computer Sciences[◇]
University of California, Berkeley

October 7, 2024

Abstract

This paper provides a new analysis of the biological version of Oja's rule for solving the Streaming Principal Component Analysis (PCA) problem. Motivated by the efficient coding principle in neural systems, we offer the first convergence rate guarantees that match the theoretical lower bounds for streaming PCA. Our approach uses an ODE-inspired framework to analyze the stochastic dynamics of the biological Oja's rule, showing that it efficiently handles high-dimensional streaming data with biologically realistic time scales. The convergence rate is nearly optimal and dimension-independent when the initial weight vector is close to the top eigenvector. This work advances our understanding of the biological plausibility of Oja's rule in real neural systems, particularly in the retina-optical nerve pathway.

Keywords: Streaming PCA, Oja's Rule, Biological Neural Networks, Convergence Rate, Stochastic Dynamics, Ordinary Differential Equations (ODE)

1 Introduction

Understanding how biological systems, particularly neural networks, process high-dimensional sensory data is a fundamental challenge in both neuroscience and machine learning. The retina, for example, receives and processes vast amounts of information, and studies have suggested that it may implement some form of Principal Component Analysis (PCA) to achieve dimensionality reduction. However, while this idea has been explored in various theoretical frameworks, the underlying learning dynamics have remained elusive for decades.

One of the most prominent mathematical models in this area is Oja's rule, originally proposed to describe how synaptic weights in a neural network could adapt to solve the streaming PCA problem. In its biological variant, Oja's rule not only ensures local updates but also maintains synaptic normalization, which makes it particularly appealing as a model of real neural processes. Despite its relevance, existing analyses of this rule have primarily focused on convergence guarantees in the limit, without addressing biologically realistic time scales.

In this work, we present the first detailed convergence rate analysis of the biological Oja's rule, showing that it solves streaming PCA with nearly optimal efficiency. Our approach utilizes an ODE-inspired framework, which allows us to handle the stochastic nature of neural inputs while maintaining rigorous control over the dynamics. By examining the learning process as a whole, rather than step-by-step, we achieve tighter bounds on the convergence time and demonstrate that the rule functions in a biologically plausible manner, independent of the input dimension.

Backgrounds Brains processes high dimensional visual inputs constantly. In our eyes, 100 millions photoreceptors in the retina receive gigabytes of information per second [Wan95, SLS77]. In

addition, the retina is a highly convergent pathway: 100 million photoreceptors converges the visual information onto one million retina ganglion cells in optical nerves [GS12]. Therefore, it is important to understand a neural implementation of the dimensionality reduction in the retina. Furthermore, many works in theoretical neuroscience [AR90, AR92, HvH98] demonstrated from the efficient coding principle that the retina might implement Principal Component Analysis (PCA). Specifically, they showed that under natural image statistics, PCA-like solution recovers the center-surround receptive fields in the retina. 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 [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 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 retina can change its receptive field to adapt to environments with different illumination [SEC84], contrast [SEC84, BM02, SBW$^+$97], spatial frequency [SBW$^+$97, HBM05], orientation and temporal correlation [HBM05] in the time scale of seconds. 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 (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

[1]Also 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 subsection 1.4 for more discussions on their differences.

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 subsection 1.1 and state the main results and their biological relevance in subsection 1.2. In subsection 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 subsection 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.

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

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 (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 $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T \overset{i.i.d.}{\sim} \mathcal{D}$ 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 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ 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 $\Pr\left[\frac{\langle \mathbf{w}_T, \mathbf{v}_1 \rangle^2}{\|\mathbf{w}\|_2^2} \geq 1 - \epsilon\right] \geq 1 - \delta$ 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 $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$ 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 (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*

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \eta_t y_t \left(\mathbf{x}_t - y_t \mathbf{w}_{t-1}\right) \tag{2}$$

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 weight and the firing rates of the two end neurons. Second, with some simple calculations (*e.g.,* Lemma 12), 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 subsection 1.2 for more discussions.

[2] In 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.

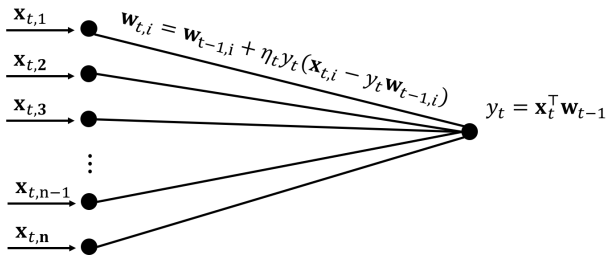


Figure 1. A neural network that uses biological Oja's rule to solve streaming PCA. The firing rate vector $\mathbf{x}_t$ is the input and the weight vector $\mathbf{w}_t$ is the output at time t.

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 normalization[3] to solve the streaming PCA problem.

$$\mathbf{w}_t = \frac{\left(I + \eta_t \mathbf{x}_t \mathbf{x}_t^\top\right) \mathbf{w}_{t-1}}{\| \left(I + \eta_t \mathbf{x}_t \mathbf{x}_t^\top\right) \mathbf{w}_{t-1} \|_2} \tag{3}$$

However, the normalization term $\| \left(I + \eta_t \mathbf{x}_t^\top \mathbf{x}_t\right) \mathbf{w}_{t-1} \|_2^{-1}$ 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 2). 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.

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

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 2 ([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 $\lim_{t\to\infty} \langle \mathbf{w}_t, \mathbf{v}_1 \rangle^2 = 1$ 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.

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

[3]This 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 subsection 1.4 for more discussion.

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

in the traditional step-by-step analysis and thus naive analysis would not work. We explain the difficulty further in subsection 1.3 and Appendix C. Given this situation, natural questions on the frontier would then be: ***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 subsection 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 3 (Global and local convergence)**.** *With the setting in Problem 1 and dynamic in Definition 1, let* $\mathsf{gap} := \lambda_1 - \lambda_2 > 0$*. For any* $\epsilon, \delta \in (0,1)$*, we have the following results.*

• *(Local Convergence) Suppose* $\frac{\langle \mathbf{w}_0, \mathbf{v}_1 \rangle^2}{\|\mathbf{w}_0\|_2^2} = \Omega(1)$*. For any* $n \in \mathbb{N}$*,* $\delta, \epsilon \in (0,1)$*), let*

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

Then, we have

$$\Pr\left[\frac{\langle \mathbf{w}_T, \mathbf{v}_1 \rangle^2}{\|\mathbf{w}_T\|_2^2} < 1 - \epsilon\right] < \delta$$

• *(Global Convergence) Suppose* $\mathbf{w}_0$ *is uniformly sampled from the unit sphere of* $\mathbb{R}^n$*. For any* $n \in \mathbb{N}$*,* $\delta, \epsilon \in (0,1)$*, let*

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

Then, we have

$$\Pr\left[\frac{\langle \mathbf{w}_T, \mathbf{v}_1 \rangle^2}{\|\mathbf{w}_T\|_2^2} < 1 - \epsilon\right] < \delta$$

The notation $a \wedge b$ *stands for* $\min\{a, b\}$ *and* $\widetilde{\Theta}$ *hides the poly-logrithmic factors with respect to* $\epsilon^{-1}, \delta^{-1}, \mathsf{gap}^{-1}, n$*.*

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 4 (For-all-time guarantee with slowly diminishing rate)**.** *With the setting in Problem 1 and dynamic in Definition 1, let* $\mathsf{gap} := \lambda_1 - \lambda_2 > 0$*. 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* $t \in \mathbb{N}$*, there exists* $\eta_t \geq \Theta\left(\frac{\epsilon \cdot \mathsf{gap}}{\lambda_1 \log(t/\delta)}\right)$ *such that*

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

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 4). Because our learning rate is slowly diminishing, when the environment changes, the learning rate is still large enough to do efficient learning. This allows the sensory system to continuously adapt to changing environments without taking a long time to adapt or reset the learning rate. 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 $\eta_t = \widetilde{O}(1/t)$ [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.,* $A = \mathsf{diag}(\lambda_1 \dots, \lambda_n)$ with $\lambda_1 > \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_n \geq 0$. 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 $t \to \infty$. A reduction from the general case to the diagonal case is provided in subsection 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

$$\frac{d\mathbf{w}_{t,1}}{dt} \geq (\lambda_1 - \lambda_2)\mathbf{w}_{t,1}(1 - \mathbf{w}_{t,1}^2) \quad \text{and} \quad \|\mathbf{w}_t\|_2 = 1 \tag{4}$$

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.

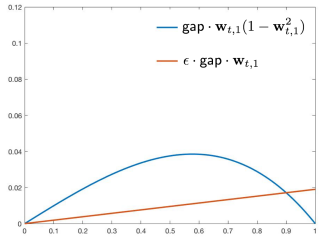
Note that in a discrete stochastic dynamic, there are two sources of the noises: (i) the intrinsic stochasticity from its continuous analog and (ii) the noise due to discretization. Thus, Equation 4 suggests that the noise in the biological Oja's rule only comes from discretization since the continuous Oja's rule is deterministic.

In addition to the limiting behavior, one can also read out finer structures of the continuous dynamic from Equation 4 by solving the differential equation using standard tools from dynamical system. The right hand side (RHS) of the inequality in Equation 4 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 4, 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.

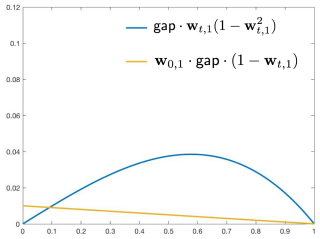
For example, we can linearize at 0 by lower bounding the RHS of Equation 4 by $\epsilon(\lambda_1 - \lambda_2)\mathbf{w}_{t,1}$ for any $\mathbf{w}_{t,1} \in [0, \sqrt{1-\epsilon}]$ (see Figure 2a). Similarly, we can linearize at 1 by using $\mathbf{w}_{0,1}(\lambda_1 - \lambda_2)(1 - \mathbf{w}_{t,1})$ 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).

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.

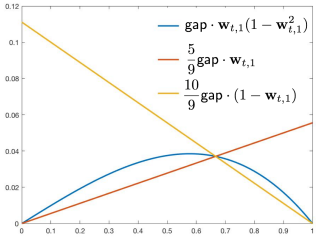
Another key inspiration from the continuous dynamic is the *ODE trick* which provides a close form characterization of the dynamic in terms of the drifting term captured by the continuous dynamic and the noise term originated from the linearization and discretization. The ODE trick



(a) Linearization only at 0.



(b) Linearization only at 1.



(c) Linearization at both 0 and 1.

Figure 2: Comparison between one-sided linearization and two-sided linearization.

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 linearization[5]. By the standard tool for solving linear ODE, the solution of $y(t)$ at $t = T$ is

$$y(T) = e^{aT} \cdot \left(y(0) + \int_0^\top e^{-at} b(t) dt \right) \tag{5}$$

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 $\int_0^\top e^{-at} b(t) dt$. 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 $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t$. For any $t \in \mathbb{N}$, we have

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

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

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

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 through careful variable substitution, but as a result they incur unnecessary logrithmic costs. 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 example[6], we have

$$\mathbf{w}_{T,1} = H^\top \cdot \left(\mathbf{w}_{0,1} + \sum_{t=1}^{T} \frac{N_t}{H^t} \right) \tag{6}$$

where $H > 1$ is the multiplier term and $\{N_t\}$ is the noise term which forms a martingale on the natural filtration. See Corollary 3 and Corollary 2 for a precise formulation of H and $\{N_t\}$ in our analysis. Intuitively, one should think of $H^\top \mathbf{w}_{0,1}$ 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 6 is small, Azuma's inequality (see Lemma 1) 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 case[7] 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

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

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

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 6: 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 $M_t = \sum_{t'=1}^{\top} H^{-t'} N_{t'}$ be the noise term in the ODE trick (*i.e.,* Equation 6) 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) which gives the following stronger guarantee than the traditional Azuma's inequality.

$$\Pr\left[\sup_{1\leq t\leq T} |M_{t\wedge\tau} - M_0| \geq a\right] < \delta \tag{7}$$

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 $t \in [T]$,

$$\Pr\left[\tau \geq t+1 \;\middle|\; \sup_{1\leq t'\leq t} |M_{t'} - M_0| < a\right] = 1 \tag{8}$$

Namely, if the bad event has not happened, then the martingale would not stop immediately. Intuitively, Equation 8 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 7 and Equation 8), we are able to show in the pull-out lemma that

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

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 $\{\sup_{1\le t\le T}|M_t - M_0| \ge a\}$ in to two parts P_1 and P_2 with the following properties. In P_1, we can show that

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

As for P_2, we use the chain condition in Equation 8 to show that the probability of error event is 0 based on a *diagonal argument*. Thus, we have

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

See section 6 and Figure 4 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

Related theory work on biological Oja's variants. Computational neuroscientists have proposed several variants of the biological Oja's rule to solve streaming PCA [Oja82, Oja92, San89, Fö89, Lee91, RT89, KDT94, PHC15]. In the single neuron case, Oja used stochastic approximation theory [KC78] to prove the global convergence in the limit [Oja82]. In the multi-neurons case, Hornik and Kuan similarly demonstrated the connection between the discrete dynamics and the associated ODE [HK92] from Kushner-Clark theorem [KC78]. However, most existing analysis on the multi-neurons case shows only local convergence [San89, Fö89, Lee91, RT89, KDT94, PHC15]. Even for the convergence in the limit, the global convergence for most networks in the multi-neurons case is difficult to show. Yan et al. provided the only global analysis on Oja's multi-neurons subspace network [Oja92, YHM94, Yan98]. Previously there is no work showing the convergence rate on the discrete dynamics. This paper shows the first convergence rate bound on the biological Oja's rule.

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 $t = 0, 1, \ldots, T$, the update rule is

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

and the output is $\mathbf{w}_T/\|\mathbf{w}_T\|_2$. Note that the above update rule is equivalent to Equation 3, *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 subsection 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 $\mathsf{gap} := \lambda_1 - \lambda_2 \in (0,1)$, 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 $\Omega(\lambda_1 \mathsf{gap}^{-2}\epsilon^{-1}\log(\delta^{-1}))$ 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.

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 [LMP17c, LMP17a, LMP17b, LM18, SCL19], the efficiency of spiking neural networks in storing temporal information [LW19, HP19], assemblies [LMPV18, PV18], spiking neural

*Let $f(\log n, \log(1/\epsilon), \log(1/\delta), \log(1/\mathsf{gap}))$ 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 $n, 1/\epsilon, 1/\delta, 1/\mathsf{gap}$ 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 $\Omega(\frac{\lambda_1}{\mathsf{gap}} \cdot \frac{1}{\epsilon})$ 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.

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	$\widetilde{O}\left(\frac{\lambda_1}{\mathsf{gap}^2}\cdot\frac{1}{\epsilon\wedge\delta^2}\right)$	3	$\widetilde{O}\left(\frac{\lambda_1}{\mathsf{gap}^2}\cdot\frac{1}{\epsilon}\right)$	2
ML Oja's Rule	[DSOR15]	N	$\widetilde{O}\left(\frac{n}{\mathsf{gap}^2}\cdot\frac{1}{\epsilon}\right)$ †	- ‡	$\widetilde{O}\left(\frac{n}{\mathsf{gap}^2}\cdot\frac{1}{\epsilon}\right)$ †	- ‡
	[Sha16]	Y	$\widetilde{O}\left(\frac{n}{\mathsf{gap}^2}\cdot\frac{1}{\epsilon}\right)$ †	- ‡	$\widetilde{O}\left(\frac{n}{\mathsf{gap}^2}\cdot\frac{1}{\epsilon}\right)$ †	- ‡
	[LWLZ18]	N	$\widetilde{O}\left(\frac{\lambda_1 n}{\mathsf{gap}^2}\cdot\frac{1}{\epsilon\delta^6}\right)$	- ‡	$\widetilde{O}\left(\frac{\lambda_1 n}{\mathsf{gap}^2}\cdot\frac{1}{\epsilon\delta^4}\right)$	- ‡
	[JJK+16]	Y	$\widetilde{O}\left(\frac{\lambda_1}{\mathsf{gap}^2}\cdot\frac{1}{\epsilon\delta^3}\right)$	2	$\widetilde{O}\left(\frac{\lambda_1}{\mathsf{gap}^2}\cdot\frac{1}{\epsilon\delta^3}\right)$	2
	[AZL17]	Y	$\widetilde{O}\left(\frac{\lambda_1}{\mathsf{gap}^2}\cdot\frac{1}{\epsilon\wedge\delta^2}\right)$	≥ 4	$\widetilde{O}\left(\frac{\lambda_1}{\mathsf{gap}^2}\cdot\frac{1}{\epsilon}\right)$	≥ 3
Any Algorithm	[AZL17]	$\Omega\left(\frac{\lambda_1}{\mathsf{gap}^2}\cdot\frac{\log\frac{1}{\delta}}{\epsilon}\right)$ §				

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.

networks in solving optimization problems [CCL19, Peh19] and learning hierarchically structured concepts [LMT19]. Under this context, this work provides an algorithmic insight in a plasticity learning rule that solves streaming PCA.

2 Preliminaries

In this section, we introduce the mathematical notations and tools that we use in this work.

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 $\widetilde{O}$ (similarly, $\widetilde{\Omega}$ and $\widetilde{\Theta}$) 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 $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ are the eigenvalues of the covariance matrix A. $\mathsf{diag}(\lambda)$ denotes the diagonal matrix with λ on the diagonal. We will follow the convention of stochastic process and denote $\min\{a, b\}$ as $a \wedge b$. We say an event happens *almost surely* if it happens with probability one.

2.2 Probability toolbox

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 (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 $X_0, X_1, \ldots, X_t$ 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 $X_0, X_1, \ldots, X_t$. One of the most common adapted processes is the martingale.

Definition 3 (Martingale)**.** *Let $\{M_t\}_{t\in\mathbb{N}_{\geq 0}}$ be a sequence of random variables and let $\{\mathcal{F}_t\}_{\mathbb{N}}$ 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.

Lemma 1 (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 $t = 1, 2, \ldots, T$, $|M_t - M_{t-1}| \leq c$ almost surely, then we have*

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

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 (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 $t = 1, 2, \ldots, T$, $|M_t - M_{t-1}| \leq c$ 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 strengthen by considering the conditional variance. This is known as the Freedman's inequality.

Lemma 3 (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 $t = 1, 2, \ldots, T$, $|M_t - M_{t-1}| \leq c$ almost surely and $\mathsf{Var}[M_t - M_{t-1} \mid \mathcal{F}_{t-1}] \leq \sigma_t^2$, 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 1. *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 $t = 1, 2, \ldots, T$, $|M_t - M_{t-1}| \leq c$ almost surely, $\mathsf{Var}[M_t - M_{t-1} \mid \mathcal{F}_{t-1}] \leq \sigma_t^2$, and $|\mathbb{E}[M_t - M_{t-1} \mid \mathcal{F}_{t-1}]| \leq \mu_t$, then we have*

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

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

Definition 4 (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 $\{M_t \geq a\}$".

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.

We have the following identity for the stopped process.

Lemma 4 (The difference of a stopped process)**.** *Given a stochastic process M_t and a stopping time τ. We have*

$$M_{t\wedge\tau} - M_{(t-1)\wedge\tau} = \mathbf{1}_{\tau\geq t}(M_t - M_{t-1})$$

Proof. We have

$$\begin{aligned} M_{t\wedge\tau} - M_{(t-1)\wedge\tau} &= \mathbf{1}_{\tau\geq t}M_t + \mathbf{1}_{\tau<t}M_\tau - \mathbf{1}_{\tau\geq t-1}M_{t-1} - \mathbf{1}_{\tau<t-1}M_\tau \\ &= \mathbf{1}_{\tau\geq t}M_t - \mathbf{1}_{\tau\geq t-1}M_{t-1} + \mathbf{1}_{\tau=t-1}M_\tau \end{aligned}$$

Since $\tau = t - 1$ at the last term, we can combine the last two terms to have

$$\begin{aligned} &= \mathbf{1}_{\tau\geq t}M_t - \mathbf{1}_{\tau\geq t}M_{t-1} \\ &= \mathbf{1}_{\tau\geq t}(M_t - M_{t-1}) \end{aligned}$$

as desired. □

Here we also define a *shifted stopping process* which is an useful variant used in more complicated situations.

Definition 5 (The shifted stopped process)**.** *Given an adapted stochastic process M_t with respect to filtration $\mathcal{F}_t$ and a stopping time τ, we define a new adapted process $M_{t\star\tau}$ with respect to $\mathcal{F}_t$ to be*

$$M_{t\star\tau} = \mathbf{1}_{\tau>t}M_t + \mathbf{1}_{\tau\leq t}M_{\tau-1}$$

Given $t \in \mathbb{N}$, we define a random variable $t \star \tau$ as

$$t \star \tau = \mathbf{1}_{\tau>t}t + \mathbf{1}_{\tau\leq t}(\tau - 1)$$

Intuitively, a shifted stopped process is the original process which moves one step back if the stopping time stops.

Lemma 5 (The difference of a shifted stopped process)**.** *Given a stochastic process M_t and a stopping time τ. We have*

$$M_{t\star\tau} - M_{(t-1)\star\tau} = \mathbf{1}_{\tau>t}(M_t - M_{t-1})$$

Proof. We have

$$\begin{aligned} M_{t\star\tau} - M_{(t-1)\star\tau} &= \mathbf{1}_{\tau>t}M_t + \mathbf{1}_{\tau\leq t}M_{\tau-1} - \mathbf{1}_{\tau>t-1}M_{t-1} - \mathbf{1}_{\tau\leq t-1}M_{\tau-1} \\ &= \mathbf{1}_{\tau>t}M_t - \mathbf{1}_{\tau>t-1}M_{t-1} + \mathbf{1}_{\tau=t}M_{\tau-1} \end{aligned}$$

Since $\tau = t$ at the last term, we can combine the last two terms to have

$$\begin{aligned} &= \mathbf{1}_{\tau>t}M_t - \mathbf{1}_{\tau>t}M_{t-1} \\ &= \mathbf{1}_{\tau>t}(M_t - M_{t-1}) \end{aligned}$$

as desired. □

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 6 (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.*

- *$\beta_0 = 0$ and β_t is almost surely continuous.*
- *For any t_1, t_2, t_3, t_4 such that $0 \leq t_1 < t_2 \leq t_3 < t_4$, $\beta_{t_2} - \beta_{t_1}$ is independent from $\beta_{t_4} - \beta_{t_3}$.*
- *For any t_1, t_2 such that $0 \leq t_1 < t_2$, $\beta_{t_2} - \beta_{t_1} \sim N(0, t_2 - t_1)$.*

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 $d\beta_t = \lim_{\Delta\to 0} \beta_{t+\Delta} - \beta_t$. 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 for the first time. This is the *quadratic variation* of Brownian motion as stated below.

Lemma 6 (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.*

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

We omit the proof of Lemma 6 here and refer the interested readers to standard textbook such as [LG16] for more details on Brownian motion.

2.3 ODE toolbox

Lemma 7 (ODE trick for scalar)**.** *Let $\{X_t\}_{t\geq\mathbb{N}_{\geq 0}}$, $\{A_t\}_{t\in\mathbb{N}}$, and $\{H_t\}_{t\in\mathbb{N}}$ be sequences of random variables with the following dynamic*

$$X_t = H_t X_{t-1} + A_t \tag{9}$$

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 7. For each $t_0 < i \leq t$, dividing Equation 9 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. □

Lemma 8 (ODE trick for vector)**.** *Let $\{X_t\}_{t\in\mathbb{N}_{\geq 0}}$, $\{A_t\}_{t\in\mathbb{N}}$ be sequences of m_t-dimensional random variables and $\{H_t\}_{t\in\mathbb{N}}$ be a sequence of random $m_t \times m_{t-1}$ matrices with the following dynamic*

$$X_t = H_t X_{t-1} + A_t \tag{10}$$

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

$$X_t = \prod_{i=t_0+1}^{T} H_i X_{t_0} + \sum_{i=t_0+1}^{T} \prod_{j=i+1}^{T} H_{t-j} A_i$$

Proof of Lemma 8. The proof is a direct induction. □

2.4 Approximation toolbox

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

Lemma 9. *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 10. *For any $x \in (0, 0.5)$ and $t \in \mathbb{N}$,*

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

Lemma 11. *For any $\epsilon \in (0, 1)$, we have*

$$\left(\frac{\epsilon}{8}\right)^{1-\frac{1}{\log\frac{8}{\epsilon}}} = \frac{\epsilon}{4}$$

Proof. Rewrite the expression as the follows.

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

It suffices to show that $\left(\frac{8}{\epsilon}\right)^{\frac{1}{\log\frac{8}{\epsilon}}} \cdot \frac{1}{8} = \frac{1}{4}$. Consider the logarithm of the term, we have

$$\log\left(\left(\frac{8}{\epsilon}\right)^{\frac{1}{\log\frac{8}{\epsilon}}} \cdot \frac{1}{8}\right) = \frac{1}{\log\frac{8}{\epsilon}}\left(3 + \log\frac{1}{\epsilon}\right) - 3 = 1 - 3 = -2$$

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 $A = \mathsf{diag}(\lambda)$ 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 subsection 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,

$$\frac{1}{\Delta}\mathbb{E}\left[\left(B_{t+\Delta} - B_t\right)\left(B_{t+\Delta} - B_t\right)^\top\right] = \mathsf{diag}(\lambda) \tag{11}$$

Thus, by discretizing B_t into intervals of length $\Delta > 0$, $\left\{\frac{1}{\sqrt{\Delta}}\left(B_{j\cdot\Delta} - B_{(j-1)\cdot\Delta}\right)\right\}_{j\in\mathbb{N}}$ naturally forms a stream of i.i.d. input[8] with covariance matrix being A. To put this into the context of biological

[8]Though 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.

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 $\Delta B_j = \left(B_{j\cdot\Delta} - B_{(j-1)\cdot\Delta}\right)$ [9]. Then, we get the following dynamic.

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

The above dynamics becomes continuous once we let $\Delta \to 0$. Formally, we replace[10] $B_{t+\Delta} - B_t$ with dB_t and index the weight vector by $t \geq 0$, *i.e.,* $(\mathbf{w}_t)_{\geq 0}$. The above dynamic becomes the following SDE.

$$d\mathbf{w}_t = dB_t^\top \mathbf{w}_t dB_t - \left(dB_t^\top \mathbf{w}_t\right)^2 \mathbf{w}_t \tag{12}$$

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 12. 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 $(\beta_t)_{\geq 0}$ be a Brownian motion, we have $d\beta_t^2 = dt$ almost surely (see Lemma 6). 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 12 can be rewritten as the following *deterministic* process almost surely.

$$d\mathbf{w}_t = \left[\mathsf{diag}(\lambda)\mathbf{w}_t - \mathbf{w}_t^\top \mathsf{diag}(\lambda)\mathbf{w}_t \mathbf{w}_t\right] dt \tag{13}$$

With the continuous Oja's rule being deterministic as in Equation 13, 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 13 by linearizing the dynamic at 0 and 1 respectively and get two incomparable convergence rates (Theorem 5 and Theorem 6).

Theorem 5 (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 6 (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* $\mathbf{w}_{t,1}^2 > 1 - \epsilon$*.*

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

[10] This 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 5 and Theorem 6 are based on applying Taylor's expansion on Equation 13 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 $O(\frac{\log n}{\epsilon(\lambda_1-\lambda_2)})$ and $O(\frac{\sqrt{n}\log(1/\epsilon)}{\lambda_1-\lambda_2})$ 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 5 with the error parameter being 0.5, then for some $t_1 = O(\frac{\log n}{\lambda_1-\lambda_2})$, we have $\mathbf{w}_{t_1,1} > 0.5$. Next, we invoke Theorem 6 starting from $\mathbf{w}_{t_1}$ and with the error parameter being ϵ, then for some $t_2 = O(\frac{\log(1/\epsilon)}{\lambda_1-\lambda_2})$, 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 7 (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 and Definition 1. Thus, for simplicity, we would not repeat the setup in their statements.

In Theorem 8, we show that both the local and the global convergence of Oja's rule are efficient. We remind readers that in the local convergence setting, the weight vector is correlated with the top eigenvector by a constant while in the global convergence setting, the weight vector is randomly initiated. In Theorem 9 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 without decreasing the learning rate too much. This demonstrates the capacity of Oja's rule as a continual learning mechanism in a living system.

Theorem 8 (Main Theorem)**.** *We have the following results on the local and global convergence of Oja's rule.*

• *(Local Convergence) Let* $n \in \mathbb{N}$*,* $\delta \in (0,1), \epsilon \in (0, \frac{1}{8})$*. Suppose* $\frac{\langle \mathbf{w}_0, \mathbf{v}_1\rangle^2}{\|\mathbf{w}_0\|_2^2} \geq 2/3$*. Let*

$$\eta = \Theta\left(\frac{\epsilon(\lambda_1-\lambda_2)}{\lambda_1 \log \frac{\log\log \frac{1}{\epsilon}}{\delta}}\right),\ T = \Theta\left(\frac{\log \frac{1}{\epsilon}}{\eta(\lambda_1-\lambda_2)}\right)$$

Then, we have

$$\Pr\left[\frac{\langle \mathbf{w}_T, \mathbf{v}_1\rangle^2}{\|\mathbf{w}_T\|_2^2} < 1-\epsilon\right] < \delta$$

Namely, the convergence rate is of order

$$\Theta\left(\frac{\lambda_1 \log \frac{1}{\epsilon}\left(\log\log\log \frac{1}{\epsilon} + \log \frac{1}{\delta}\right)}{\epsilon(\lambda_1-\lambda_2)^2}\right)$$

with probability at least $1-\delta$.

• *(Global Convergence) Let* $n \in \mathbb{N}$, $\delta \in (0,1), \epsilon \in (0, \frac{1}{4})$. *Suppose* $\mathbf{w}_0$ *is uniformly sampled from the unit sphere of* $\mathbb{R}^n$. *Let*

$$\eta = \Theta\left(\frac{\lambda_1-\lambda_2}{\lambda_1} \cdot \left(\frac{\epsilon}{\log \frac{\log \frac{n}{\epsilon}}{\delta}} \bigwedge \frac{\delta^2}{\log^2 \frac{\lambda_1 n}{\delta(\lambda_1-\lambda_2)^2}}\right)\right), T = \Theta\left(\frac{\log \frac{1}{\epsilon} + \log \frac{n}{\delta}}{\eta(\lambda_1-\lambda_2)}\right)$$

Then, we have

$$\Pr\left[\frac{\langle \mathbf{w}_T, \mathbf{v}_1\rangle^2}{\|\mathbf{w}_T\|_2^2} < 1-\epsilon\right] < \delta$$

Namely, the convergence rate is of order

$$\Theta\left(\frac{\lambda_1\left(\log \frac{1}{\epsilon} + \log \frac{n}{\delta}\right)}{(\lambda_1-\lambda_2)^2} \cdot \max\left\{\frac{\log \frac{\log \frac{n}{\epsilon}}{\delta}}{\epsilon}, \frac{\log^2 \frac{\lambda_1 n}{\delta(\lambda_1-\lambda_2)^2}}{\delta^2}\right\}\right)$$

with probability at least $1-\delta$.

Proof structure of Theorem 8. To prove Theorem 8, we first reduce the general setting where the covariance matrix A is PSD to the special case where $A = \mathsf{diag}(\lambda)$ in section 5. For local convergence, we show that starting from constant correlation, Oja's rule can efficiently converge to the top eigenvector up to arbitrarily small error in Theorem 10 of section 6. For global convergence, we show that starting from random initialization, Oja's rule can efficiently converge to the top eigenvector up to arbitrarily small error in Theorem 13 of section 7. To get tight analysis for global convergence, we need to take an extra care on a *cross term* in Theorem 14 of section 8. See Figure 3 for the proof structure of these theorems.

Theorem 9 (Continual Learning)**.** *We have the following results on the continual learning aspects of Oja's rule.*

• *(Finite continual learning) Let* $n, l \in \mathbb{N}$, $\epsilon, \delta \in (0,1)$. *Suppose* $\frac{\langle \mathbf{w}_0, \mathbf{v}_1\rangle^2}{\|\mathbf{w}_0\|_2^2} \geq 1 - \frac{\epsilon}{2}$. *Let*

$$\eta = \Theta\left(\frac{\epsilon(\lambda_1-\lambda_2)}{\lambda_1 \log \frac{l}{\delta}}\right)$$

Then

$$\Pr\left[\exists 1 \leq t \leq \Theta\left(\frac{l}{\eta(\lambda_1-\lambda_2)}\right), \frac{\langle \mathbf{w}_T, \mathbf{v}_1\rangle^2}{\|\mathbf{w}_T\|_2^2} < 1-\epsilon\right] < \delta$$

• *(For-all-time continual learning) Let* $n, t_0 \in \mathbb{N}$, $\epsilon, \delta \in (0,1)$. *Suppose* $\frac{\langle \mathbf{w}_0, \mathbf{v}_1\rangle^2}{\|\mathbf{w}_0\|_2^2} \geq 1 - \frac{\epsilon}{2}$. *Then there is*

$$\eta_t \geq \Theta\left(\frac{\epsilon(\lambda_1-\lambda_2)}{\lambda_1 \log \frac{t}{\delta}}\right)$$

such that

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

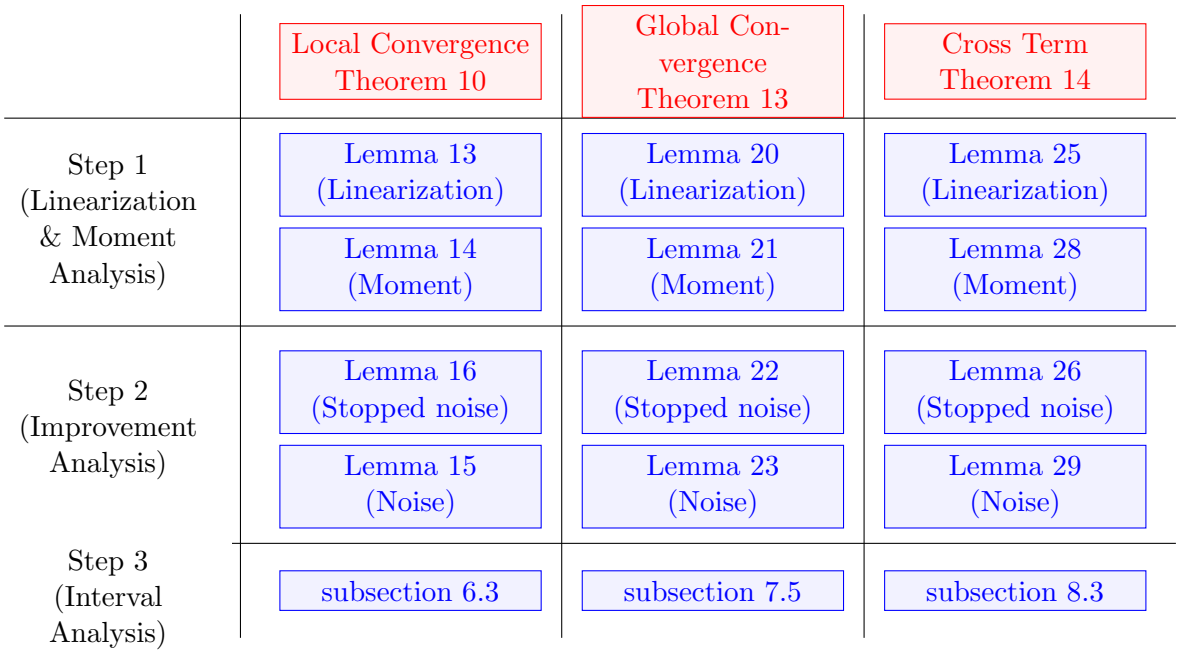


Figure 3. The proof structure of key theorems. Here we present the structure of the three main theorems using the three-step framework described in a follow-up paper [CWY20].

Proof structure of Theorem 9. We first reduce the general setting to the special case where $A = \mathsf{diag}(\lambda)$ in section 5. The proof of finite continual learning is then a direct application of techniques developed in local convergence. By repetitively applying finite continual learning, we can show for-all-time continual learning. The results will be proven in subsection 6.4.

5 Preprocessing

Before the main analysis of the biological Oja's rule, we provide two useful observations on the dynamic in this section. Specifically, we show in subsection 5.1 that considering the covariance matrix being *diagonal* is sufficient for the analysis and in subsection 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.,* $1 \geq \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$, 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 $\widetilde{\mathbf{w}}_t = U\mathbf{w}_t$ and $\widetilde{\mathbf{x}}_t = U\mathbf{x}_t$. As $U^\top U = UU^\top = I$, we have $\widetilde{\mathbf{x}}_t^\top \widetilde{\mathbf{w}} = \mathbf{x}_t^\top \mathbf{w}$ and $\mathbb{E}[\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}^\top] = D$. Let $\mathbf{v}_1$ be the top eigenvector of A (*i.e.,* the first row of U), we also have

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

where $\mathbf{e}_1$ is the indicator vector for the first coordinate. Namely, it suffices to analyze how fast

does $\widetilde{\mathbf{w}}_t$ converge to $\mathbf{e}_1$. Thus, we without loss of generality consider the diagonal case where the goal would be showing that $\mathbf{w}_{t,1}^2 \geq 1 - \epsilon$.

5.2 Bounded conditions of Oja's rule

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

Lemma 12. *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 of Lemma 12. Here we prove only 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

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

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. □

6 Local Convergence: Starting With Correlated Weights

For the local convergence result, the synaptic weight $\mathbf{w}_0$ is correlated with the top eigenvector by a constant. To be precise, we suppose that $\mathbf{w}_{0,1}^2 \geq \frac{2}{3}$. The goal of this section is to show that $1 - \mathbf{w}_{t,1}^2 \leq \epsilon$ for some $t = O(\frac{\lambda_1 \log(1/\epsilon)(\log\log\log(1/\epsilon) + \log(1/\delta))}{\epsilon(\lambda_1 - \lambda_2)^2})$ for any small $\epsilon > 0$. Let us first state the main theorem of this section as follows.

Theorem 10 (Local convergence of the diagonal case)**.** *Suppose $\mathbf{w}_{0,1}^2 \geq 2/3$. For any $n \in \mathbb{N}$, $\delta \in (0, 1), \epsilon \in (0, \frac{1}{8})$, let*

$$\eta = \Theta\left(\frac{\epsilon(\lambda_1 - \lambda_2)}{\lambda_1 \log \frac{\log\log \frac{1}{\epsilon}}{\delta}}\right),\ T = \Theta\left(\frac{\log \frac{1}{\epsilon}}{\eta(\lambda_1 - \lambda_2)}\right)$$

Then

$$\Pr\left[\mathbf{w}_{T,1}^2 < 1 - \epsilon\right] < \delta$$

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

Proof overview and organization. First note that by applying the diagonal reduction argument in subsection 5.1, Theorem 10 implies the local convergence part of Theorem 8 as a corollary. The proof structure of Theorem 10 is as follows. First, in subsection 6.1 we derive a linearization of the dynamic using a center at 1 instead of 0 based on the intuition from the continuous dynamic in section 3. Furthermore, we use the ODE trick to write down the dynamic in a closed form with respect to the linearization.

Next, in subsection 6.2, we want to show that the noise term is small. However, the difficulty here is that $\mathbf{w}_{t,1}$ might *go back* to the small region (*e.g.*, $\mathbf{w}_{t,1}^2 < \frac{1}{3}$) and thus the bounded difference

might become too large to effectively bound the noise with Freedman's inequality. To deal with this issue, we consider a stopping time where $\mathbf{w}_{t,1}^2 < 1 - a$ to give good control on the bounded difference and subsequently bound the stopped version of the noise term in Lemma 16. After we show that the stopped noise term is small, we want to pull out the stopping time to show the concentration on the original noise term. In general, pulling out the stopping time is impossible without introducing extra failure probability; however, by exploiting the structure of the dynamic, we are able to pull out the stopping time without additional cost in Lemma 17.

Finally in subsection 6.3, by combining the small noise and the ODE trick, we are able to prove Theorem 10 with an interval analysis. As a corollary of Lemma 15 in the local convergence, we show that biological Oja's rule has the continual learning capacity in subsection 6.4. In a biological system, it is important to function for a long period of time instead of at one time point. In this section, we prove two theorems on continual learning. Theorem 11 guarantees Oja's rule can maintain the convergence for any finite time length efficiently while Theorem 12 guarantees Oja's rule can function for all time without sacrificing too much efficiency to adapt to a new environment.

6.1 Linearization and ODE trick centered at 1

In this section, we derive the linearization of Oja's rule with a center at 1 in Lemma 13 and the closed form solution of Oja' rule in Corollary 2. In addition, we show that the bounded differences and moments of the noise can be controlled in Lemma 14.

In the analysis of the local convergence, we use the linearization with a 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 $\widetilde{\mathbf{w}}_{t,1} = \mathbf{w}_{t,1} - 1$ and the goal becomes to show that $\widetilde{\mathbf{w}}_{t_0+t_2,1} > -\epsilon$ with probability at least $1 - \delta$. The following lemma states the linearization for $\widetilde{\mathbf{w}}_{t,1}$.

Lemma 13 (Linearization at 1)**.** *Let* $\widetilde{\mathbf{w}}_t = \mathbf{w}_{t,1}^2 - 1$ *and* $\mathbf{z}_t = \mathbf{x}_t y_t - y_t^2 \mathbf{w}_{t-1}$*. For any* $t \in \mathbb{N}_{\geq 0}$ *and* $\eta \in (0, 1)$*, we have*

$$\widetilde{\mathbf{w}}_t \geq H \cdot \widetilde{\mathbf{w}}_{t-1} + A_t + B_t$$

almost surely, where

$$\begin{aligned}
H &= 1 - \frac{2}{3}(\lambda_1 - \lambda_2)\eta\,, \\
A_t &= 2\eta \mathbf{z}_{t,1}\mathbf{w}_{t-1,1} + \eta^2 \mathbf{z}_{t,1}^2 - \mathbb{E}\left[2\eta \mathbf{z}_{t,1}\mathbf{w}_{t-1,1} \,|\, \mathbf{w}_{t-1}\right] + 2\eta\lambda_2(1 - \|\mathcal{F}_{t-1}\|^2)\mathbf{w}_{t-1,1}^2\,, \\
B_t &= -2\eta(\lambda_1 - \lambda_2)\widetilde{\mathbf{w}}_{t,1}(\frac{2}{3} + \widetilde{\mathbf{w}}_{t,1})
\end{aligned}$$

Proof of Lemma 13. By expanding $\mathbf{w}_{t,1}^2$ with the Oja's rule (Equation 2), we have

$$\mathbf{w}_{t,1}^2 = \mathbf{w}_{t-1,1}^2 + 2\eta \mathbf{z}_{t,1}\mathbf{w}_{t-1,1} + \eta^2 \mathbf{z}_t^2\,.$$

Add and subtract $\mathbb{E}\left[2\eta \mathbf{z}_{t,1}\mathbf{w}_{t-1,1} \,|\, \mathcal{F}_{t-1}\right] - 2\eta\lambda_2(1 - \|\mathbf{w}_{t-1}\|^2)\mathbf{w}_{t-1}^2$. We have

$$= \mathbf{w}_{t-1,1}^2 + 2\eta(\lambda_1 \mathbf{w}_{t-1,1}^2 - \sum_{i=1}^{n} \lambda_i \mathbf{w}_{t-1,i}^2 \mathbf{w}_{t-1,1}^2 - \lambda_2(1 - \|\mathbf{w}_{t-1}\|^2)\mathbf{w}_{t-1,1}^2) + A_t$$

Upper bound $\sum_{i=2}^{n} \lambda_i \mathbf{w}_{t-1,i}^2 \mathbf{w}_{t-1,1}^2$ by $\lambda_2 \sum_{i=2}^{n} \mathbf{w}_{t-1,i}^2 \mathbf{w}_{t-1,1}^2$, we then have

$$\geq \mathbf{w}_{t-1,1}^2 + 2\eta(\lambda_1(\mathbf{w}_{t-1,1}^2 - \mathbf{w}_{t-1,1}^4) - \lambda_2(\mathbf{w}_{t-1,1}^2 - \mathbf{w}_{t-1,1}^4)) + A_t$$

$$= \mathbf{w}_{t-1,1}^2 + 2\eta(\lambda_1 - \lambda_2)\mathbf{w}_{t-1,1}^2(1 - \mathbf{w}_{t-1,1}^2) + A_t \,. \tag{14}$$

Based on the intuition from the continuous dynamic in section 3, since we want to converge from constant error to ϵ error, we want to linearize at 1. Hence we rewrite Equation 14 in terms of $\widetilde{\mathbf{w}}_{t,1} = \mathbf{w}_{t,1}^2 - 1$ and get

$$\begin{aligned}\widetilde{\mathbf{w}_t} &\geq \widetilde{\mathbf{w}}_{t-1} - 2\eta(\lambda_1 - \lambda_2)\widetilde{\mathbf{w}}_{t-1}(1 + \widetilde{\mathbf{w}}_{t-1}) + A_t \\ &= H \cdot \widetilde{\mathbf{w}}_{t-1} + A_t + B_t\end{aligned}$$

as desired. □

We apply the ODE trick (see Lemma 7) on Lemma 13 and get the following corollary.

Corollary 2 (ODE trick)**.** *For any $t_0 \in \mathbb{N}_{\geq 0}$, $t \in \mathbb{N}$, and $\eta \in (0, 1)$, we have*

$$\widetilde{\mathbf{w}}_{t_0+t} \geq H^t \cdot \left(\widetilde{\mathbf{w}}_{t_0} + \sum_{i=t_0+1}^{t_0+t} \frac{A_i + B_i}{H^{i-t_0}} \right)$$

To control the noise term, we need to have bounds on the bounded differences and the moments of A_i, B_i.

Lemma 14. *Let A_t, B_t be defined as in Lemma 13. For any $t \in \mathbb{N}$, we have A_t, B_t satisfy the following properties:*

- *(Bounded difference) $|A_t| = O(\eta|\widetilde{\mathbf{w}}_{t-1}| + \eta|\widetilde{\mathbf{w}}_{t-1}|^{\frac{1}{2}} + \eta^{\frac{3}{2}})$ almost surely. If $\widetilde{\mathbf{w}}_{t-1,1} \geq -\frac{2}{3}$, then $B_t \geq -O(\eta^2)$ almost surely.*
- *(Conditional expectation) $\mathbb{E}[A_t \mid \mathcal{F}_{t-1}] = O(\eta^2\lambda_1)$.*
- *(Conditional variance) $\mathsf{Var}[A_t \mid \mathcal{F}_{t-1}] = O\left(\eta^2\lambda_1\left(|\widetilde{\mathbf{w}}_{t-1}|^2 + |\widetilde{\mathbf{w}}_{t-1}| + \eta\right)\right)$.*

Proof. First by Lemma 12, we have $|\mathbf{w}_{t,1}|, |y_t| < \sqrt{1 + 10\eta} < 1 + 10\eta < 2$. Now let's bound $|\mathbf{z}_{t,1}|$ first. By expanding $|\mathbf{z}_{t,1}|$, we have

$$\begin{aligned}|\mathbf{z}_{t,1}| &= |y_t(\mathbf{x}_{t,1} - y_t\mathbf{w}_{t-1,1})| \\ &= \left| y_t \left(\mathbf{x}_{t,1}(1 - \mathbf{w}_{t-1,1}^2) - \sum_{i=2}^{n} \mathbf{x}_{t,i}\mathbf{w}_{t-1,i}\mathbf{w}_{t-1,1} \right) \right| \\ &\leq |y_t| \cdot \left(|\mathbf{x}_{t,1}\widetilde{\mathbf{w}}_{t-1}| + \left| \sum_{i=2}^{n} \mathbf{x}_{t,i}\mathbf{w}_{t-1,i}\mathbf{w}_{t-1,1} \right| \right).\end{aligned}$$

By Cauchy-Schwarz and the fact that $\|\mathbf{x}\|_2 = 1$, we have

$$\leq |y_t| \cdot \left(|\widetilde{\mathbf{w}}_{t-1}| + \left| \sqrt{\left(\sum_{i=2}^{n} \mathbf{x}_{t,i}^2 \right) \left(\sum_{i=2}^{n} \mathbf{w}_{t-1,i}^2 \right)} \mathbf{w}_{t-1,1} \right| \right).$$

By Lemma 12 and the definition of $\widetilde{\mathbf{w}}_{t-1}$, we have

$$\leq |y_t| \cdot \left(|\widetilde{\mathbf{w}}_{t-1}| + \left| \sqrt{-\widetilde{\mathbf{w}}_{t-1} + 10\eta} \right| \right)$$

$$\leq |y_t| \cdot \left(|\widetilde{\mathbf{w}}_{t-1}| + \sqrt{|\widetilde{\mathbf{w}}_{t-1}|} + \sqrt{10\eta} \right). \tag{15}$$

Since $|y_t| \leq 2$, we have

$$\leq 2 \left(|\widetilde{\mathbf{w}}_{t-1}| + \sqrt{|\widetilde{\mathbf{w}}_{t-1}|} + \sqrt{10\eta} \right).$$

Combining above, Lemma 12 and the fact that $\mathbf{z}_{t,1} = O(1)$, we have

$$|A_t| = O\left(\eta |\widetilde{\mathbf{w}}_{t-1}| + \eta |\widetilde{\mathbf{w}}_{t-1}|^{\frac{1}{2}} + \eta^{\frac{3}{2}} \right)$$

and for $\widetilde{\mathbf{w}}_{t-1} \geq -\frac{2}{3}$, we have $B_t \geq -O(\eta^2)$ because $\frac{2}{3} + \widetilde{\mathbf{w}}_{t-1} > 0$ and $\widetilde{\mathbf{w}}_{t-1} \leq O(\eta)$.

For conditional expectation, notice that $\mathbb{E}[y_t^2 | \mathcal{F}_{t-1}] = \mathbf{w}_{t-1}^\top \mathsf{diag}(\lambda) \mathbf{w}_{t-1} = O(\lambda_1)$. This implies that $\mathbb{E}[\mathbf{z}_{t,1}^2 | \mathcal{F}_{t-1}] = O(\lambda_1)$ and hence $\mathbb{E}[A_t | \mathcal{F}_{t-1}] = O(\eta^2 \lambda_1)$. Now the conditional variance is

$$\mathsf{Var}[A_t | \mathcal{F}_{t-1}] = O\left(\eta^2 \mathbb{E}[\mathbf{z}_{t,1}^2 \,|\, \mathcal{F}_{t-1}] \mathbf{w}_{t-1,1}^2 + \lambda_1 \eta^4 \right).$$

By Equation 15, we have

$$\begin{aligned} &= O\left(\eta^2 \mathbb{E}[y_t^2 | \mathcal{F}_{t-1}] \left(|\widetilde{\mathbf{w}}_{t-1}| + \sqrt{|\widetilde{\mathbf{w}}_{t-1}|} + \sqrt{10\eta} \right)^2 + \lambda_1 \eta^4 \right) \\ &= O\left(\eta^2 \lambda_1 \left(|\widetilde{\mathbf{w}}_{t-1}|^2 + |\widetilde{\mathbf{w}}_{t-1}| + \eta \right) \right) \end{aligned}$$

as desired. □

6.2 Concentration of noise and pulling out the stopping time

In this subsection, we want to show that the noise term in Corollary 2 is small. Specifically, we prove the following lemma.

Lemma 15 (Concentration of the noise term in local convergence)**.** *Let $\epsilon, \delta \in (0,1), T \in \mathbb{N}_{\geq 0}$. Suppose given $t_0 \in \mathbb{N}$, $v_0 \in (-\frac{1}{3}, 0)$ and $a \in [0,1]$, we have $\widetilde{\mathbf{w}}_{t_0} \geq v_0$ and $v_0 = -\Theta(\epsilon^{1-a})$. Let $\eta = \Theta\left(\frac{\epsilon(\lambda_1 - \lambda_2)}{\lambda_1 \log \frac{1}{\delta}} \right)$. If $H^{-T} = \Theta(\epsilon^{-\frac{a}{2}})$, then*

$$\Pr\left[\min_{1 \leq t \leq T} \sum_{i=t_0+1}^{t_0+t} \frac{A_i + B_i}{H^{i-t_0}} \leq v_0 \right] < \delta$$

The most natural way to prove such a statement is using a martingale concentration inequality. However, the difficulty here is that $\widetilde{\mathbf{w}}_t$ might *go back* to the small region (*e.g.,* $\widetilde{\mathbf{w}}_t < -2/3$) and thus the bounded difference might become too large to bound the noise effectively with Freedman's inequality. Nevertheless, the continuous dynamic (see section 3) suggests that this situation should happen with only a small probability because the $\mathbf{w}_1$ term in the continuous dynamic increases monotonically to 1. To enforce the analysis, we consider a *stopped process* where the dynamic stops once $\widetilde{\mathbf{w}}_t$ is too small. This stopped process satisfies good bounded difference conditions by its construction and thus we can apply Freedman's inequality on it. See Lemma 16 for a formal statement of the above intuition.

After obtaining good control of the noise term in the stopped process, we want to remove the stopping time and show the concentration of the original non-stopped process in order to

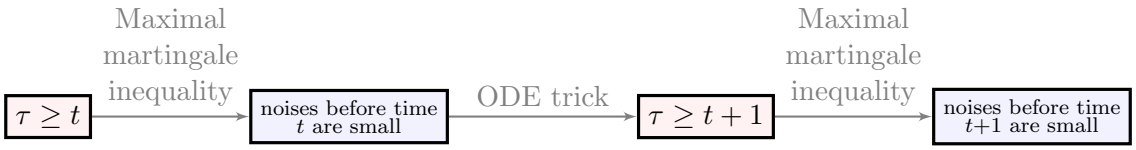


Figure 4: Intuition on why it is possible to pull out stopping time in Phase 2.

prove Lemma 15. This can be done by Lemma 17 which *pulls out* the stopping time from the concentration inequality for the stopped process. In general, pulling out the stopping time is impossible without introducing additional failure probability; however, the following structure of the stochastic process we are looking at allows us to pull out the stopping time. Intuitively, given a stopping time τ with $\tau \geq t$ for some t, with high probability all the noise terms before time t are small (using a maximal martingale inequality). Next, the noise being small at time t would further imply that $\tau \geq t+1$ (using the ODE trick). The above argument forms a chain of implications as pictured in Figure 4.

With the above *chain* structure in the noise terms, we are then able to pull out the stopping time in Lemma 16 by introducing another stopping time to help us properly partition the probability space. The rest of this subsection is devoted to formalizing the above intuition and completing the proof for Lemma 15.

First, let us show the concentration of the stopped process.

Lemma 16 (Concentration of stopped noise in an interval)**.** *Let $\epsilon, \delta \in (0,1), T \in \mathbb{N}_{\geq 0}$. Suppose given $t_0 \in \mathbb{N}$, $v_0 \in (-\frac{1}{3}, 0)$ and $a \in [0,1]$, we have $\widetilde{\mathbf{w}}_{t_0} \geq v_0$ and $v_0 = -\Theta(\epsilon^{1-a})$. Let τ_{v_0} to be the stopping time $\{\widetilde{\mathbf{w}}_t < 2v_0\}$ such that $t > t_0$. Let $\eta = \Theta\left(\frac{\epsilon(\lambda_1-\lambda_2)}{\lambda_1 \log \frac{1}{\delta}}\right)$. If $H^{-T} = \Theta(\epsilon^{-\frac{a}{2}})$, then*

$$\Pr\left[\min_{1\leq t\leq T} \sum_{i=t_0+1}^{(t_0+t)\wedge\tau_{v_0}} \frac{A_i + B_i}{H^{i-t_0}} \leq v_0\right] < \delta$$

Proof of Lemma 16. We are going to apply Freedman's inequality Corollary 1 on the stopped process $\sum_{i=t_0+1}^{(t_0+t)\wedge\tau_{v_0}} \frac{A_i}{H^{i-t_0}}$. First notice that given a stopping time τ and an adapted stochastic process M_t, the difference of the stopped process can be described as

$$M_{t\wedge\tau} - M_{(t-1)\wedge\tau} = \mathbf{1}_{\tau\geq t}(M_t - M_{t-1})$$

For notational convenience, we denote $\mathbf{1}_{\tau_{v_0}\geq(t_0+t)}A_t$ as $\overline{A}_t$. Now by Lemma 14 and geometric series, *i.e.*, $\sum_{i=1}^{T} H^{-i} \leq O(\frac{H^{-T}}{\eta(\lambda_1-\lambda_2)})$, we have

$$\forall 1 \leq t \leq T,\ \left|\frac{\overline{A}_{t_0+t}}{H^{\top}}\right| \leq O\left(\eta\epsilon^{\frac{1-a}{2}}\right)$$

$$\left|\sum_{i=t_0+1}^{t_0+T} \mathbb{E}\left[\frac{\overline{A}_i}{H^{i-t_0}} \,\middle|\, \mathcal{F}_{i-1}\right]\right| \leq O\left(\eta^2 \frac{H^{-T}}{\eta(\lambda_1-\lambda_2)}\right) = O\left(\frac{\eta\lambda_1\epsilon^{-\frac{a}{2}}}{\lambda_1-\lambda_2}\right), \text{ and}$$

$$\left|\sum_{i=t_0+1}^{t_0+T} \mathsf{Var}\left[\frac{\overline{A}_i}{H^{i-t_0}} \,\middle|\, \mathcal{F}_{i-1}\right]\right| \leq O\left(\eta^2\lambda_1\epsilon^{1-a} \frac{H^{-2T}}{\eta(\lambda_1-\lambda_2)}\right) = O\left(\frac{\eta\lambda_1\epsilon^{1-2a}}{\lambda_1-\lambda_2}\right)$$

By applying the above bounds to Lemma 3, we have

$$\Pr\left[\max_{0\le t\le T}\left|\sum_{i=t_0+1}^{(t_0+t)\wedge\tau_{v_0}}\frac{A_i}{H^{i-t_0}}\right|\ge\frac{|v_0|}{2}\right]<\delta$$

because the deviation term is $O(\sqrt{\frac{\log\frac{1}{\delta}\eta\lambda_1\epsilon^{1-2a}}{\lambda_1-\lambda_2}}) = O\left(\epsilon^{1-a}\right) \le \frac{|v_0|}{4}$ and the summation of the conditional expectation terms is $O(\frac{\eta\lambda_1\epsilon^{-\frac{a}{2}}}{\lambda_1-\lambda_2}) = O\left(\epsilon^{1-a}\right) \le \frac{|v_0|}{4}$. By stopping time and Lemma 14, we have

$$\sum_{i=t_0+1}^{(t_0+T)\wedge\tau_{t_0}}\frac{B_i}{H^{i-t_0}}\ge -O\left(\eta^2\frac{\epsilon^{-\frac{a}{2}}}{\eta(\lambda_1-\lambda_2)}\right)\ge -O(\epsilon^{1-\frac{a}{2}})\ge -\frac{v_0}{2}$$

By combining both inequalities, we get

$$\Pr\left[\min_{1\le t\le T}\sum_{i=t_0+1}^{(t_0+t)\wedge\tau_{v_0}}\frac{A_i+B_i}{H^{i-t_0}}\le v_0\right]<\delta$$

□

We are going to pull out the stopping time τ_{t_0} in Lemma 16. The following lemma shows that under a certain *chain* condition, it is possible to pull out the stopping time without introducing additional failure probability.

Lemma 17 (Pulling out the stopping time using a chain condition)**.** *Let $\{M_t\}_{t\in\mathbb{N}_{\ge 0}}$ be an adapted stochastic process and τ be a stopping time. Let $\{M_t^*\}_{t\in\mathbb{N}_{\ge 0}}$ be the maximal process of $\{M_t\}_{t\in\mathbb{N}_{\ge 0}}$ where $M_t^* = \max_{1\le t'\le t} M_t$. For any $t\in\mathbb{N}$, $a\in\mathbb{R}$, and $\delta\in(0,1)$, suppose*

1. *$\Pr[M_{t\wedge\tau}^*\ge a]<\delta$ and*
2. *For any $1\le t'<t$, $\Pr[\tau\ge t'+1 \mid M_{t'}^*<a]=1$.*

Then, we have

$$\Pr[M_t^*\ge a]<\delta$$

Proof of Lemma 17. 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 $\{M_{t\wedge\tau}^*\ge a\}$. The following claim shows that if τ stopped before time t, then τ' should stop earlier than τ.

Claim 16. *Let τ and τ' be stopping times as defined above. Suppose the conditions in Lemma 17 hold. Then we have*

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

Proof of Claim 16. 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

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

which is a contradiction. □

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

$$\Pr[M_t^* \geq a] = \Pr[M_t^* \geq a, \tau \geq t] + \Pr[M_t^* \geq a, \tau < t, \tau' \leq \tau] + \Pr[M_t^* \geq a, \tau < t, \tau' > \tau]$$

By Claim 16, we have $\Pr[M_t^* \geq a, \tau < t, \tau' > \tau] = 0$. We have

$$= \Pr[M_t^* \geq a, \tau \geq t] + \Pr[M_t^* \geq a, \tau < t, \tau' \leq \tau]$$

For the first term, 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, the equation becomes

$$\begin{aligned}&= \Pr[M_{t\wedge\tau}^* \geq a,\ \tau \geq t] + \Pr[M_{t\wedge\tau}^* \geq a,\ \tau < t,\ \tau' \leq \tau] \\ &\leq \Pr[M_{t\wedge\tau}^* \geq a]\end{aligned}$$

Thus, we conclude that $\Pr[M_t^* \geq a] \leq \Pr[M_{t\wedge\tau}^* \geq a] < \delta$ as desired. □

By applying the above Lemma 17 on Lemma 16, we can pull out the stopping time and show concentration on the original process in Lemma 15.

Lemma 15 (Concentration of the noise term in local convergence)**.** *Let $\epsilon, \delta \in (0,1), T \in \mathbb{N}_{\geq 0}$. Suppose given $t_0 \in \mathbb{N}$, $v_0 \in (-\frac{1}{3}, 0)$ and $a \in [0,1]$, we have $\widetilde{\mathbf{w}}_{t_0} \geq v_0$ and $v_0 = -\Theta(\epsilon^{1-a})$. Let $\eta = \Theta\left(\frac{\epsilon(\lambda_1-\lambda_2)}{\lambda_1 \log\frac{1}{\delta}}\right)$. If $H^{-T} = \Theta(\epsilon^{-\frac{a}{2}})$, then*

$$\Pr\left[\min_{1\leq t\leq T} \sum_{i=t_0+1}^{t_0+t} \frac{A_i + B_i}{H^{i-t_0}} \leq v_0\right] < \delta$$

Proof of Lemma 15. Let τ_{v_0} be the stopping time $\{\widetilde{\mathbf{w}}_t < 2v_0\}$ such that $t > t_0$. We want to apply Lemma 17 with $M_t = -\sum_{i=t_0+1}^{t_0+t} \frac{A_i+B_i}{H^{i-t_0}}$, $a = -v_0$ and $\tau = \tau_{v_0} - t_0$. First condition is satisfied by Lemma 16. So it is suffice to check that

$$\Pr\left[\tau_{v_0} \geq t' + t_0 + 1 \middle| \min_{1\leq t\leq t'} \sum_{i=t_0+1}^{t_0+t} \frac{A_i + B_i}{H^{i-t_0}} \leq v_0\right] = 1$$

And indeed we have by Corollary 2

$$\widetilde{\mathbf{w}}_{t_0+t'} \geq H^{t'} \cdot \left(\widetilde{\mathbf{w}}_{t_0} + \sum_{i=t_0+1}^{t_0+t'} \frac{A_i + B_i}{H^{i-t_0}}\right) > H^{t'} \cdot (v_0 + v_0) \geq 2v_0$$

This implies that $\tau_{v_0} \geq t' + t_0 + 1$ as desired. □

6.3 Interval Analysis

Given $\epsilon \in (0,1)$, let $\widetilde{\epsilon} = \frac{\epsilon}{8}$. The goal of this section is to prove the local convergence of Oja's rule (Theorem 10) with the following interval scheme that shows the improvement of $\widetilde{\mathbf{w}}_t$

$$-\frac{1}{3} \to -\widetilde{\epsilon}^{1-\frac{1}{2}} \to -\widetilde{\epsilon}^{1-\frac{1}{4}} \to \cdots \to -\widetilde{\epsilon}^{1-\frac{1}{\log\frac{1}{\epsilon}}}$$

Proof of Theorem 10. Let $\widetilde{\epsilon} = \frac{\epsilon}{8}$ and $v_0 = -\frac{1}{3}, l = \log\log\frac{1}{\widetilde{\epsilon}}$. For $1 \le i \le l$, choose $T_i \in \mathbb{N}$ such that $\frac{1}{2}\widetilde{\epsilon}^{\frac{1}{2^i}} \ge H^{T_i} \ge \frac{1}{4}\widetilde{\epsilon}^{\frac{1}{2^i}}$ and $v_i = -\widetilde{\epsilon}^{1-\frac{1}{2^i}}$. Let $S_j = \sum_{i=1}^{j} T_i$ and let $T = S_l$. Notice that by Lemma 11, we have $v_l = -\frac{\epsilon}{4}$.

We are going to show that for all $1 \le j \le l$, we have

$$\Pr\left[\widetilde{\mathbf{w}}_{S_j} \le v_j \middle| \widetilde{\mathbf{w}}_{S_{j-1}} \ge v_{j-1}\right] < \frac{\delta}{l} \tag{17}$$

Then by union bounding over j, we have $\Pr\left[\widetilde{\mathbf{w}}_T \le -\frac{\epsilon}{4}\right] < \delta$ and

$$\frac{1}{4}^l \frac{\epsilon}{4} \le H^T \le \frac{1}{2}^l \frac{\epsilon}{4} \Rightarrow T = \Theta\left(\frac{\log\log\frac{1}{\epsilon} + \log\frac{1}{\epsilon}}{\eta(\lambda_1 - \lambda_2)}\right) = \Theta\left(\frac{\log\frac{1}{\epsilon}}{\eta(\lambda_1 - \lambda_2)}\right)$$

as desired. What remains to be shown is Equation 17. Now by Lemma 15, for $\eta = \Theta\left(\frac{\epsilon(\lambda_1 - \lambda_2)}{\lambda_1 \log\frac{\log\log\frac{1}{\epsilon}}{\delta}}\right)$, we have for all $1 \le j \le l$

$$\Pr\left[\min_{1\le t\le S_j} \sum_{i=S_{j-1}+1}^{S_{j-1}+t} \frac{A_i + B_i}{H^{i-S_{j-1}}} \le v_j \middle| \widetilde{\mathbf{w}}_{S_{j-1}} \ge v_{j-1}\right] < \frac{\delta}{l}$$

Now by Corollary 2, the following is true with probability $1 - \delta$

$$\widetilde{\mathbf{w}}_{S_j} \ge H^{T_j} \cdot \left(\widetilde{\mathbf{w}}_{S_{j-1}} + \sum_{i=S_{j-1}+1}^{S_{j-1}+t} \frac{A_i + B_i}{H^{i-S_{j-1}}}\right) \ge \frac{1}{2}\widetilde{\epsilon}^{\frac{1}{2^j}} \cdot 2v_{j-1} \ge v_j$$

This shows that

$$\Pr\left[\mathbf{w}_{T,1}^2 \le 1 - \epsilon\right] \le \Pr\left[\widetilde{\mathbf{w}}_T \le -\frac{\epsilon}{4}\right] < \delta$$

as desired. □

6.4 Continual Learning

One of the most remarkable aspects of the biological learning system is its ability to function indefinitely and continuously adapt. In previous sections, we have only been looking at the convergence of Oja's rule at a time point. However, the sensory system needs to function for a long period of time or even for all time. In this section, we explore the capacity of Oja's rule for continual learning as an application of the previous techniques. In Theorem 11, we show that Oja's rule can maintain its convergence for any finite time while in Theorem 12, we show that Oja's rule can maintain its convergence for all time with a slowly diminishing learning rate that scales like $\Omega(\frac{1}{\log t})$. This shows that even if the animal switches to a new environment after a period of time, the learning rate is still large enough to allow efficient continual learning. Notice that the Kushner-Clark theorem requires $\sum_t \eta_t^2 < \infty$ where the learning rate is commonly set as $\eta_t = O(\frac{1}{t})$. In comparison, our slowly diminishing learning rate can achieve $\sum_t \eta_t^2 = \infty$ and thus enables efficient continual learning.

First, we have the following finite continual learning theorem. By applying the diagonal reduction argument in subsection 5.1, we prove the finite continual learning part of Theorem 9 as a corollary.

Theorem 11 (Finite continual learning)**.** *Let* $n, l \in \mathbb{N}$, $\epsilon, \delta \in (0,1)$. *Suppose* $\mathbf{w}_{0,1}^2 \geq 1 - \frac{\epsilon}{2}$. *Let*

$$\eta = \Theta\left(\frac{\epsilon(\lambda_1 - \lambda_2)}{\lambda_1 \log \frac{l}{\delta}}\right)$$

Choose t' *such that* $\frac{1}{4} \geq H^{t'} \geq \frac{1}{8}$. *Then*

$$\Pr\left[\exists 1 \leq t \leq lt',\ \mathbf{w}_{t,1}^2 < 1 - \epsilon\right] < \delta$$

Proof of Theorem 11. Given any $1 \leq j \leq l$, by Lemma 15, we have

$$\Pr\left[\min_{1\leq t\leq t'} \sum\nolimits_{i=(j-1)t'+1}^{jt'+t} \frac{A_i + B_i}{H^{i-(j-1)t'}} \leq -\frac{\epsilon}{2} \,\middle|\, \widetilde{\mathbf{w}}_{(j-1)t'} \geq -\frac{\epsilon}{2}\right] < \frac{\delta}{l}$$

Notice conditioned on $\widetilde{\mathbf{w}}_{(j-1)t'} \geq -\frac{\epsilon}{2}$ and $\min_{1\leq t\leq t'} \sum_{i=(j-1)t'+1}^{jt'+t} \frac{A_i+B_i}{H^{i-(j-1)t'}} > -\frac{\epsilon}{2}$, we have for $1 \leq t \leq t'$ by Corollary 2

$$\widetilde{\mathbf{w}}_{(j-1)t'+t} \geq H^t \cdot \left(\widetilde{\mathbf{w}}_{(j-1)t'} + \sum\nolimits_{i=(j-1)t'+1}^{(j-1)t'+t} \frac{A_i + B_i}{H^{i-(j-1)t'}}\right) \geq H^t(-\frac{\epsilon}{2} - \frac{\epsilon}{2}) \geq -H^t\epsilon$$

In particular, $\widetilde{\mathbf{w}}_{jt'} \geq -\frac{\epsilon}{2}$. This implies that

$$\Pr\left[\left(\exists 0 \leq t \leq t',\ \widetilde{\mathbf{w}}_{(j-1)t'+t} < -\epsilon\right) \cup \left(\widetilde{\mathbf{w}}_{jt'} < -\frac{\epsilon}{2}\right) \,\middle|\, \widetilde{\mathbf{w}}_{(j-1)t'} \geq -\frac{\epsilon}{2}\right] < \frac{\delta}{l}$$

Union bound over $1 \leq j \leq l$, we get

$$\Pr\left[\exists 1 \leq t \leq lt',\ \mathbf{w}_{t,1}^2 < 1 - \epsilon\right] < \delta$$

as desired. □

As a corollary of the above finite continual learning theorem, we can obtain the following for-all-time continual learning theorem. By applying the diagonal reduction argument in subsection 5.1, we prove the for-all-time continual learning part of Theorem 9 as a corollary.

Theorem 12 (For-all-time continual learning)**.** *Let* $n, t_0 \in \mathbb{N}$, $\epsilon, \delta \in (0,1)$. *Suppose* $\mathbf{w}_{0,1}^2 \geq 1 - \frac{\epsilon}{2}$. *There is*

$$\eta_t \geq \Theta\left(\frac{\epsilon(\lambda_1 - \lambda_2)}{\lambda_1 \log \frac{t}{\delta}}\right)$$

such that

$$\Pr\left[\exists t \in \mathbb{N},\ \mathbf{w}_{t,1}^2 < 1 - \epsilon\right] < \delta$$

Proof of Theorem 12. The proof proceeds by recursively choosing η_t in intervals and apply Theorem 11 repetitively. Let $\delta_i = \frac{\delta}{2i^2}$. Then notice that $\sum_{i=1}^{\infty} \delta_i < \delta$. Now apply Theorem 11 with $t_0 = 1$ with failure probability δ_1 to get the corresponding η, t' and denote them as $\eta_{(1)}, t'_{(1)}$. Now for $1 \leq j \leq t'_{(1)}$, define $\eta_j = \eta_{(1)}$. By Theorem 11, this shows that

$$\Pr\left[\exists 1 \leq t \leq t'_{(1)},\ \mathbf{w}_{t,1}^2 < 1 - \epsilon\right] < \delta$$

For the ith interval, we apply Theorem 11 with $t_0 = 1$ with failure probability δ_i to get the corresponding η, t' and denote them as $\eta_{(i)}, t'_{(i)}$. Now for $t'_{(i-1)} \leq j \leq t'_{(i)}$, define $\eta_j = \eta_{(i)}$. Notice that the above recursive scheme ensures that $\eta_t \geq \Theta(\frac{\epsilon(\lambda_1-\lambda_2)}{\lambda_1 \log \frac{t}{\delta}})$. And by union bound, we get

$$\Pr\left[\exists t \in \mathbb{N},\ \mathbf{w}_{t,1}^2 < 1 - \epsilon\right] < \delta$$

□

7 Global Convergence: Starting From Random Initialization

For the global convergence result, the synaptic weight $\mathbf{w}_0$ starts from a random initialization. Specifically, we suppose that $\mathbf{w}_0$ is uniformly sampled from the unit sphere of $\mathbb{R}^n$. The main theorem in this section states the convergence of Oja's rule starting from a random initialization for the diagonal case. By applying the diagonal reduction argument in subsection 5.1, we prove the global convergence part of Theorem 8 as a corollary. The following theorem is the main theorem of this section.

Theorem 13 (Global convergence of the diagonal case)**.** *Suppose* $\mathbf{w}_0$ *is uniformly sampled from the unit sphere of* $\mathbb{R}^n$*. For any* $n \in \mathbb{N}$*,* $\delta \in (0,1), \epsilon \in (0, \frac{1}{4})$*, let*

$$\eta = \Theta\left(\frac{\lambda_1 - \lambda_2}{\lambda_1} \cdot \left(\frac{\epsilon}{\log \frac{\log \frac{n}{\epsilon}}{\delta}} \bigwedge \frac{\delta^2}{\log^2 \frac{\lambda_1 n}{\delta(\lambda_1 - \lambda_2)^2}}\right)\right),\ T = \Theta\left(\frac{\log \frac{1}{\epsilon} + \log \frac{n}{\delta}}{\eta(\lambda_1 - \lambda_2)}\right)$$

Then

$$\Pr\left[\mathbf{w}_{T,1}^2 < 1 - \epsilon\right] < \delta$$

Proof overview and organization. The main difficulty in the global convergence is that at the beginning the bounded differences of the noise in the linearization (see Lemma 21) cannot be controlled directly. To be precise, the *cross term* $|y_t|$ at the worst case needs to be bounded by $O(\sqrt{n}|\mathbf{w}_{t,1}|)$. This will introduce a polynomial dependency on n, which makes the convergence inefficient. To deal with this issue, in subsection 7.1, we provide an initialization lemma and the definition of the stopping time $\xi_{p,\delta}$ that controls the bounded difference of $|y_t|$. Next, in subsection 7.2, we show that we can control the bounded difference of $|y_t|$ by showing that the stopping time $\xi_{p,\delta}$ is large with high probability in Theorem 14. The details of the proof is delayed to section 8.

Next, in subsection 7.3 we derive a linearization for $\mathbf{w}_{t,1}^2$ using a center at 0 instead of 1 based on the intuition from the continuous dynamic in section 3. Furthermore, we use the ODE trick to write down the dynamic in a closed form with respect to the linearization.

Similar to the local convergence, in subsection 7.4, we show that the noise from the ODE trick can be controlled with the stopping time and we can pull out the stopping time carefully to bound the original noise. In subsection 7.5, we prove that $\mathbf{w}_{t,1}^2$ is greater than $2/3$ efficiently with high probability in an interval analysis in Theorem 15. Finally, in subsection 7.6, by combining Theorem 15, the local convergence Theorem 10 and the finite continual learning Theorem 11, we prove the efficient global convergence in Theorem 13.

7.1 Initialization and the main stopping time

In this section, we begin with Definition 7, which introduces the auxiliary stochastic processes that we study in subsection 7.2 for controlling $|y_t|$. Then we give an initialization lemma for the auxiliary processes in Lemma 18, which guarantees that the processes perform well with good probability at the first time step.

Definition 7 (Auxiliary processes)**.** *For each* $2 \le j \le n$*,* $t \in [T]$*, and* $\mathbf{w} \in \mathbb{R}^n$*, define*

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

We cite the initialization lemma in [AZL17, Lemma 5.1] with some straightforward modification below.

Lemma 18 (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 $p, \delta \in (0,1)$, there exists*

$$\Lambda_{p,\delta} = \Theta\left(\frac{1}{p}\sqrt{\log\frac{nT}{\delta}}\right), \Lambda'_p = \Theta\left(\frac{n}{p^2}\log\frac{n}{p}\right)$$

such that let $\mathcal{C}_0^{p,\delta}(\mathbf{w}_0)$ denote the event $\{\exists j \in [n],\ t \in [T],\ |f_{t,j}(\mathbf{w}_0)| > \Lambda_{p,\delta}\}$ and let $\mathcal{C}_{init}^{p,\delta}$ denote the event $\{\mathbf{w}_{0,1}^2 \geq \frac{1}{\Lambda'_p}\} \cap \{\Pr_{\mathbf{x}_1,\ldots,\mathbf{x}_T\sim\mathcal{D}}[\mathcal{C}_0^{p,\delta}(\mathbf{w}_0)] < \delta\}$, we have

$$\Pr_{\mathbf{w}_0}\left[\mathcal{C}_{init}^{p,\delta}\right] \geq 1 - p - \delta$$

In particular, by the definition of $\mathcal{C}_{init}^{p,\delta}$ we have $\Pr[\mathcal{C}_0^{p,\delta} \mid \mathcal{C}_{init}^{p,\delta}] < \delta$. For convenience, we denote $\mathcal{C}_0^{p,\delta}(\mathbf{w}_0)$ as $\mathcal{C}_0^{p,\delta}$ when there is no confusion.

7.2 $|y_t|$ is small with high probability

As we said at the beginning of the section, in order to keep the bounded differences of the noise in the global convergence small, we need to make $|y_t|$ small with high probability. To achieve this, we introduce the following stopping time.

Definition 8 (Stopping times for controlling $|y_t|$)**.** *Given $p, \delta \in (0,1)$ in Lemma 18, we define the stopping time $\psi_{p,\delta}$ to be the first time t such that $\mathbf{w}_{t,1}^2 < 1/(2\Lambda'_{p,\delta})$ and the stopping time $\xi_{p,\delta}$ to be the first time t such that $|f_{t,n}(\mathbf{w}_{(t-1)\wedge\psi_{p,\delta}})| > 2\Lambda_{p,\delta}$.*

When there is no confusion, we will abbreviate $\psi_{p,\delta}, \xi_{p,\delta}, \Lambda_{p,\delta}, \Lambda'_p$ as $\psi, \xi, \Lambda, \Lambda'$ respectively. Intuitively, bounded difference and moments in the noise of the auxiliary process will be small if ψ and ξ are large with high probability.

To control $|y_t|$, we would like to show that show that $\xi_{p,\delta}$ is large with high probability. Specifically, we want to show the following Theorem 14.

Theorem 14 (ξ is large with high probability)**.** *Let $T \in \mathbb{N}$ and $p, \delta \in (0,1)$. Let $\eta = \Theta\left(\frac{(\lambda_1-\lambda_2)}{\lambda_1\Lambda^2_{p,\delta/4n^2T}\log\frac{nT}{\delta}}\right)$. If we have $T = \Omega\left(\frac{1}{\eta\lambda_1}\right)$ and $p \leq \delta$, then we have*

$$\forall t \in [T],\ \Pr\left[\xi = t \;\middle|\; \mathcal{C}_{init}^{p,\frac{\delta}{4n^2T}}\right] < \frac{\delta}{2n^2T}$$

In particular we have

$$\forall t \in [T],\ \Pr\left[\xi = t \;\middle|\; \xi \geq t, \mathcal{C}_{init}^{p,\frac{\delta}{4nT}}\right] \leq \frac{\delta}{n^2T}$$

and

$$\Pr\left[\xi \leq T \;\middle|\; \mathcal{C}_{init}^{p,\frac{\delta}{4nT}}\right] < \frac{\delta}{2n^2}$$

The proof of Theorem 14 requires a careful analysis using stopping time and pull-out technique on $(f_{t,2}, \ldots, f_{t,n})$. We describe the rough strategy as follows and defer the details to section 8. For all $2 \le i \le n$, we consider the processes $\{f_{t,i}(\mathbf{w}_s)\}_{s\in[t-1]}$ and, specifically, the shifted stopped process $f_{t,i}(\mathbf{w}_{s\wedge\psi\star\xi})$ to make sure the bounded differece of the noise is small. Now Lemma 18 makes sure that $|f_{t,i}(\mathbf{w}_0)|$ is under controlled with high probability and by our framework to analyze stochastic processes we can control $|f_{t,i}(\mathbf{w}_{t-1\wedge\psi\star\xi})|$ with high probability and therefore show that ξ is large.

Notice that here we use the shifted stopped process in Definition 5 with stopping time ξ. Recall the definition of the shifted stopped process.

Definition 5 (The shifted stopped process)**.** *Given an adapted stochastic process M_t with respect to filtration $\mathcal{F}_t$ and a stopping time τ, we define a new adapted process $M_{t\star\tau}$ with respect to $\mathcal{F}_t$ to be*

$$M_{t\star\tau} = \mathbf{1}_{\tau>t} M_t + \mathbf{1}_{\tau\le t} M_{\tau-1}$$

Given $t \in \mathbb{N}$, we define a random variable $t \star \tau$ as

$$t \star \tau = \mathbf{1}_{\tau>t} t + \mathbf{1}_{\tau\le t}(\tau - 1)$$

Intuitively, we can consider the shifted stopped process as an original process which moves one step back if it stops. When we consider to apply the stopping time ξ, we should use the shifted stopped process instead of the normal stopped process. To given an intuition, we have by Lemma 4

$$\mathbf{w}_{t\wedge\xi} - \mathbf{w}_{(t-1)\wedge\xi} = \mathbf{1}_{\xi\ge t}\eta y_t(\mathbf{x}_t - y_t\mathbf{w}_{t-1})$$

However, $\mathbf{1}_{\xi\ge t}$ only ensures $f_{t-1,n}(\mathbf{w}_{t-2})$ is bounded and hence y_{t-1} is bounded, but we need y_t to be bounded instead. Instead if we use the shifted stopped process, by Lemma 5,

$$\mathbf{w}_{t\star\xi} - \mathbf{w}_{(t-1)\star\xi} = \mathbf{1}_{\xi>t}\eta y_t(\mathbf{x}_t - y_t\mathbf{w}_{t-1})$$

This makes sure that we have a good control on $|y_t|$ because $\xi > t$.

However, the shifted star process creates one caveat when we calculate the moments for the differences term. To be precise, the difference of a shifted star process contains $\mathbf{1}_{\xi>s}$ and $\xi > s$ is not $\mathcal{F}_{s-1}$ measurable anymore. So we need to do an addition conditioning. Observe that when we condition on the event $\xi > s$, the conditional expectation might slightly change. The following lemma shows that this is not significant.

Lemma 19 (Conditional expectation on $\xi > s$)**.** *Let $T \in \mathbb{N}$, ξ be the stopping time specified before, $\delta' \in (0, 0.5)$, and $t \in [T]$. For $s < t$, suppose*

$$\Pr\left[\xi = s \;\middle|\; \xi \ge s, \mathcal{F}_{s-1}^{(t)}\right] < \delta'$$

we have

$$\left|\mathbb{E}\left[\mathbf{x}_{s,i}\mathbf{x}_{s,j} \;\middle|\; \mathcal{F}_{s-1}^{(t)}, \xi > s\right] - \mathbb{E}\left[\mathbf{x}_{s,i}\mathbf{x}_{s,j} \;\middle|\; \mathcal{F}_{s-1}^{(t)}\right]\right| < \frac{2\delta'}{1-\delta'}$$

and furthermore

$$\left\|\mathbb{E}\left[\mathbf{x}_s\mathbf{x}_s^\top \;\middle|\; \mathcal{F}_{s-1}^{(t)}, \xi > s\right] - \mathbb{E}\left[\mathbf{x}_s\mathbf{x}_s^\top \;\middle|\; \mathcal{F}_{s-1}^{(t)}\right]\right\|_2 < \frac{2\delta'}{1-\delta'}$$

Proof of Lemma 19. By the laws of total expectation and rearranging the terms we have

$$\begin{aligned}\mathbb{E}[\mathbf{x}_{s,i}\mathbf{x}_{s,j} \,|\, \mathcal{F}_{s-1}^{(t)}] &= \mathbb{E}[\mathbf{x}_{s,i}\mathbf{x}_{s,j} \,|\, \mathcal{F}_{s-1}^{(t)},\ \xi \geq s] \\ &= \Pr[\xi > s \,|\, \mathcal{F}_{s-1}^{(t)}] \cdot \mathbb{E}[\mathbf{x}_{s,i}\mathbf{x}_{s,j} \,|\, \mathcal{F}_{s-1}^{(t)},\ \xi > s] \\ &+ \Pr[\xi = s \,|\, \mathcal{F}_{s-1}^{(t)},\ \xi \geq s] \cdot \mathbb{E}[\mathbf{x}_{s,i}\mathbf{x}_{s,j} \,|\, \mathcal{F}_{s-1}^{(t)},\ \xi = s]\end{aligned}$$

The first equality is due to the fact that $\{\xi \geq s\} \in \mathcal{F}_{s-1}^{(t)}$. Now, the first inequality are then immediate consequences of the condition of the lemma and the fact that $|\mathbf{x}_{s,i}\mathbf{x}_{s,j}| \leq 1$ almost surety. The second inequality can be obtained similarly. □

7.3 Linearization and ODE trick centered at 0

In the analysis of the global convergence, we use a linearization with a center at 0 instead of 1. The idea is inspired from the analysis of the continuous dynamic as explained in section 3. However, unlike in the local convergence case, the bounded differences here can only be controlled after applying the stopping time $\xi_{p,\delta}$ from the last section. For the rest of the section, we set a stopping time and an initialization event from subsection 7.2 to be $\xi_T = \xi_{\delta/4,\delta/8n^2T}$ and $\mathcal{C}_{init}^T = \mathcal{C}_{init}^{\delta/4,\delta/8n^2T}$. In particular by Lemma 18 and Theorem 14, we have $\forall t \in [T]$,

$$\Pr[\mathcal{C}_{init}^T] \geq 1 - \frac{\delta}{2},\ \Pr[\xi_T < T \mid \mathcal{C}_{init}^T] < \frac{\delta}{4n^2},\ \Pr[\xi_T = t \,|\, \xi_T \geq t, \mathcal{C}_{init}^T] \leq \frac{\delta}{2n^2T} \tag{18}$$

We abbreviate the corresponding $\Lambda_{\delta/4,\delta/8n^2T}, \Lambda'_{\delta/4}$ as Λ, Λ' for the rest of the section. As in the local convergence, we will use stopping times, ψ, ξ_T, to make sure the bounded difference of the noise in the global convergence is small. As explained in the subsection 7.2, we need to use the shifted stopped process with ξ_T. Specifically, we are going to look at $\mathbf{w}_{t\wedge\psi\star\xi_T,1}^2$.

We first derive the linearization with a center at 0.

Lemma 20 (Linearization at 0)**.** *Let $\mathbf{z}_t = \mathbf{x}_t y_t - y_t^2 \mathbf{w}_{t-1}$. For any $t \in \mathbb{N}$ and $\eta \in (0,1)$, we have*

$$\mathbf{w}_{t,1}^2 \geq H \cdot \mathbf{w}_{t-1,1}^2 + A_t + B_t$$

almost surely, where

$$\begin{aligned}H &= 1 + \frac{2}{3}(\lambda_1 - \lambda_2)\eta\,, \\ A_t &= 2\eta\mathbf{z}_{t,1}\mathbf{w}_{t-1,1} + \eta^2\mathbf{z}_{t,1}^2 - \mathbb{E}\left[2\eta\mathbf{z}_{t,1}\mathbf{w}_{t-1,1} \,|\, \mathcal{F}_{t-1}\right] + 2\eta\lambda_2(1 - \|\mathbf{w}_{t-1}\|^2)\mathbf{w}_{t-1,1}^2\,, \text{ and} \\ B_t &= 2\eta(\lambda_1 - \lambda_2)\mathbf{w}_{t-1,1}^2(1 - \mathbf{w}_{t-1,1}^2 - \frac{1}{3})\end{aligned}$$

Proof of Lemma 20. By Equation 14, we have

$$\begin{aligned}\mathbf{w}_{t,1}^2 &\geq \mathbf{w}_{t-1,1}^2 + 2\eta(\lambda_1 - \lambda_2)\mathbf{w}_{t-1,1}^2(1 - \mathbf{w}_{t-1,1}^2) + A_t \\ &= \mathbf{w}_{t-1,1}^2 + H \cdot \mathbf{w}_{t-1,1}^2 + A_t + B_t\end{aligned}$$

as desired. □

We apply the ODE trick (see Lemma 7) on Lemma 20 and get the following corollary.

Corollary 3 (ODE trick)**.** *For any $t_0 \in \mathbb{N}_{\geq 0}$, $t \in \mathbb{N}$, and $\eta \in (0,1)$, we have*

$$\mathbf{w}_{t_0+t,1}^2 \geq H^t \cdot \left(\mathbf{w}_{t_0,1}^2 + \sum_{i=t_0+1}^{t_0+t} \frac{A_i + B_i}{H^{i-t_0}} \right)$$

To control the noise term, we need to have bounds on the bounded differences and the moments of A_i, B_i.

Lemma 21. *Let A_t, B_t be defined as in Lemma 13. Let $\eta = O\left(\frac{\lambda_1 - \lambda_2}{\lambda_1 \Lambda^2 \log \frac{nT}{\delta}}\right)$. If $T = \Omega(\frac{1}{\eta \lambda_1})$, then for any $t \in [T]$ we have A_t, B_t satisfy the following properties:*

- *(Bounded difference) $|\mathbf{1}_{\xi_T > t, \psi \geq t} A_t| = O\left(\eta \Lambda \mathbf{w}_{t-1,1}^2\right)$ almost surely. If $\mathbf{w}_{t-1,1}^2 \leq \frac{2}{3}$, then $B_t \geq 0$ almost surely.*
- *(Conditional expectation) $\mathbb{E}[\mathbf{1}_{\xi_T > t, \psi \geq t} A_t \mid \mathcal{F}_{t-1}, \mathcal{C}_{init}^T] = O\left(\lambda_1 \eta^2 \Lambda^2 \mathbf{w}_{t-1,1}^2\right)$.*
- *(Conditional variance) $\mathsf{Var}\left[\mathbf{1}_{\xi_T > t, \psi \geq t} A_t \mid \mathcal{F}_{t-1}, \mathcal{C}_{init}^T\right] = O\left(\lambda_1 \eta^2 \Lambda^2 \mathbf{w}_{t-1,1}^4\right)$.*

Proof of Lemma 21. First by the definition of ξ_T, we have $|\mathbf{1}_{\xi_T > t, \psi \geq t} y_t| = O(\Lambda |\mathbf{w}_{t-1,1}|)$ and $|\mathbf{1}_{\xi_T > t, \psi \geq t} \mathbf{z}_{t\star,1}| = O(\Lambda |\mathbf{w}_{t-1,1}|)$. Combining above and Lemma 12, we have

$$|\mathbf{1}_{\xi_T > t, \psi \geq t} A_t| = O\left(\left(\eta \Lambda + \eta^2 \Lambda^2 + \eta^2\right) \mathbf{w}_{t-1,1}^2\right) = O(\eta \Lambda \mathbf{w}_{t-1,1}^2)$$

And for $\mathbf{w}_{t-1,1}^2 \leq \frac{2}{3}$, we have $B_t \geq 0$ because $1 - \mathbf{w}_{t-1,1}^2 - \frac{1}{3} > 0$. For the conditional expectation, we have

$$\mathbb{E}\left[\mathbf{1}_{\xi_T > t, \psi \geq t} \mathbf{z}_{t,1}^2 \;\middle|\; \mathcal{F}_{t-1}, \mathcal{C}_{init}^T\right] = \mathbb{E}\left[\mathbf{1}_{\xi_T > t, \psi \geq t} y_t^2 (x_{t,1} - y_t \mathbf{w}_{t-1,1})^2 \;\middle|\; \mathcal{F}_{t-1}, \mathcal{C}_{init}^T\right]$$

By Lemma 19, Theorem 14 and definition of ξ_T, we have

$$\leq O(\lambda_1 \Lambda^2 \mathbf{w}_{t-1,1}^2)$$

Given a random variable v, we denote $\mathbb{E}[\mathbf{1}_{\xi_T > t, \psi \geq t} v | \mathcal{F}_{t-1}, \mathcal{C}_{init}^T] - \mathbb{E}[v \mid \mathcal{F}_{t-1}]$ as $\overline{v}$. Now we also have

$$\overline{\mathbf{z}}_{t,1} = \mathbf{w}_{t-1}^\top \overline{\mathbf{x}_t \mathbf{x}_{t,1}} - \mathbf{w}_{t-1}^\top \overline{\mathbf{x}_t \mathbf{x}_t^\top} \mathbf{w}_{t-1} \mathbf{w}_{t-1,1}$$

By applying Lemma 19 with Equation 18 and Cauchy-Schwarz, we have

$$= O\left(\|\mathbf{w}_{t-1}\|_2 \frac{\sqrt{n}}{n^2 T} + \|\mathbf{w}_{t-1}\|_2^3 \frac{1}{n^2 T} \right)$$

Conditioning on $\psi \geq t$ we have $\frac{1}{n} = O(\Lambda^2 \mathbf{w}_{t-1,1}^2)$ by the definition of Λ, Λ'. We have

$$= O(\eta \lambda_1 \Lambda^2 \mathbf{w}_{t-1,1}^2)$$

So combining above we have

$$\mathbb{E}[\mathbf{1}_{\xi_T > t, \psi \geq t} A_t | \mathcal{F}_{t-1}] = O(\eta^2 \lambda_1 \Lambda^2 \mathbf{w}_{t-1,1}^2)$$

And similarly applying Lemma 19, we obtain that the conditional variance is

$$\begin{aligned} \mathsf{Var}\left[\mathbf{1}_{\xi_T > t, \psi \geq t} A_t | \mathcal{F}_{t-1}, \mathcal{C}_{init}^T\right] &= O\left(\eta^2 \mathbb{E}\left[\mathbf{1}_{\xi_T > t, \psi \geq t} \mathbf{z}_{t,1}^2 | \mathcal{F}_{t-1}, \mathcal{C}_{init}^T\right] \mathbf{w}_{t-1,1}^2 + \eta^4 \mathbb{E}\left[\mathbf{1}_{\xi_T > t, \psi \geq t} \mathbf{z}_{t,1}^4 | \mathcal{F}_{t-1}, \mathcal{C}_{init}^T\right]\right) \\ &= O(\eta^2 \lambda_1 \Lambda^2 \mathbf{w}_{t-1,1}^4) \end{aligned}$$

as desired. □

7.4 Concentration of noise in an interval

In this subsection, we want to show that the noise term in Corollary 3 is small. As in the local analysis, we are going to use a stopping time to control good bounded differences. Specifically, we have the following stopped concentration follow from Lemma 21.

Lemma 22 (Concentration of the stopped noise term in an interval)**.** *Let $t_0, T, t' \in \mathbb{N}, \delta, \delta' \in (0,1)$ and $a \in (0, \frac{2}{3})$. Suppose $\mathbf{w}^2_{t_0 \wedge \psi \star \xi_T, 1} \geq \frac{a}{2}$. Let τ_a be the stopping time $\{\mathbf{w}^2_{t \wedge \psi \star \xi_T, 1} \geq a\}$. Let $\eta = \Theta\left(\frac{(\lambda_1 - \lambda_2)}{\lambda_1 \Lambda^2 \log \frac{1}{\delta'}}\right)$. If $\delta' = O(\frac{\delta}{nT})$, $8 \geq H^{t'} \geq 4$ and $T = \Omega(\frac{1}{\eta \lambda_1})$, then*

$$\Pr\left[\min_{1 \leq t \leq t'} \sum\nolimits_{i=t_0+1}^{(t_0+t) \wedge \psi \star \xi_T \wedge \tau_a} \frac{A_i + B_i}{H^{i-t_0}} \leq -\frac{a}{2} \;\middle|\; \mathcal{C}^T_{init}\right] < \delta'$$

Proof of Lemma 22. For notational convenience, we denote $\mathbf{1}_{\tau_a, \psi \geq t, \xi > t} A_t$ as $\overline{A}_t$. The proof is based on upper bounding the three moment quantities of the stopped noise term and applying martingale concentration inequality (*i.e.,* Lemma 3). This is very similar to the proof of Lemma 16. To analyze the moment quantities, first recall from Lemma 4 and Lemma 5 that the martingale difference of the stopped noise term at time t is $\mathbf{1}_{\tau_a, \psi \geq t, \xi > t} \frac{A_t + B_t}{H^{t-t_0}}$. Next, by the definition of τ_a, we could think of $\mathbf{w}^2_{t-1,1}$ as $\mathbf{w}^2_{t-1,1} \leq a$. Now by using Lemma 21 to properly bound the moment quantities of $\mathbf{1}_{\tau_a, \psi \geq t, \xi > t} \frac{A_t}{H^{t-t_0}}$ and apply the bounds to Lemma 3, we have

$$\Pr\left[\max_{1 \leq t \leq t'} \left| \sum_{i=t_0+1}^{(t_0+t) \wedge \psi \star \xi_T \wedge \tau_a} \frac{A_i}{H^{i-t_0}} \right| \geq \frac{a}{2} \;\middle|\; \mathcal{C}^T_{init}\right] < \delta'$$

because the deviation term is $\sqrt{\frac{\log \frac{1}{\delta'} \eta \lambda_1 \Lambda^2 a^2}{\lambda_1 - \lambda_2}} = O(a) \leq \frac{a}{4}$ and the summation of conditional expectation term is $\frac{\lambda_1 \eta \Lambda^2 a}{\lambda_1 - \lambda_2} = O(a) \leq \frac{a}{4}$. By stopping time τ_a and Lemma 21, we have

$$\sum_{i=t_0+1}^{(t_0+t) \wedge \psi \star \xi_T \wedge \tau_a} \frac{B_i}{H^{i-t_0}} \geq 0$$

Combine the above two inequalities we get the desiring concentration of the stopped noise term. □

Now we will pull out the stopping time ψ, τ_a and ξ_T together to show that $\mathbf{w}^2_{t,1}$ doubles itself efficiently with high probability.

Lemma 23 (Pull out stopping time in an interval)**.** *Let $t_0, T, t' \in \mathbb{N}, \delta, \delta' \in (0,1)$ and $a \in (0, \frac{2}{3})$. Suppose $\mathbf{w}^2_{t_0 \wedge \psi \star \xi_T, 1} \geq \frac{a}{2}$. Let τ be the stopping time of $\{\mathbf{w}^2_{t,1} \geq a\}$. Let $\eta = \Theta\left(\frac{(\lambda_1 - \lambda_2)}{\lambda_1 \Lambda^2 \log \frac{1}{\delta'}}\right)$. If $\delta' = O(\frac{\delta}{nT})$, $8 \geq H^{t'} \geq 4$, $t_0 + t' \leq T$ and $T = \Omega(\frac{1}{\eta \lambda_1})$, then*

$$\Pr[\tau > t_0 + t' \,|\, \xi > T, \mathcal{C}^T_{init}] < \delta'$$

Proof of Lemma 23. Let τ_a be the stopping time $\{\mathbf{w}^2_{t \wedge \psi \star \xi_T, 1} \geq a\}$. Notice that now we only have controls on $\mathbf{w}^2_{t \wedge \psi \star \xi_T \wedge \tau_a, 1}$ via the moment information in Lemma 21. To conclude a statement about

τ, we need to pull out ψ, τ_a, ξ_T from $\mathbf{w}^2_{t,1}$. ξ_T will be pull out by paying union bounds in the conditioning. τ_a and ψ will be pulled out similar to Lemma 17.

The main goal is to upper bound the probability of the event $\tau > t_0 + t'$. First, observe that this event implies $\tau_a > t_0 + t'$ and thus we have

$$\Pr[\tau > t_0 + t' \,|\, \xi > T, \mathcal{C}^T_{init}] \le \Pr[\tau_a > t_0 + t' \,|\, \xi > T, \mathcal{C}^T_{init}]$$

Next, we further partition the probability space with the event $\psi > t_0 + t'$ and its complement. Namely,

$$\begin{aligned}\Pr[\tau_a > t_0 + t' \,|\, \xi > T, \mathcal{C}^T_{init}] &\le \Pr[\tau_a > t_0 + t' \,,\ \psi > t_0 + t' \,|\, \xi > T, \mathcal{C}^T_{init}] \\ &\quad + \Pr[\tau_a > t_0 + t' \,,\ \psi \le t_0 + t' \,|\, \xi > T, \mathcal{C}^T_{init}] \\ &= \text{Ⓐ} + \text{Ⓑ}\end{aligned}$$

In the following, we show that both Ⓐ and Ⓑ are at most $\delta'/2$ and thus complete the proof. Let us start with defining two error events on the (stopped) noise term.

$$\mathcal{A} := \left\{ \min_{1 \le t \le t'} \sum_{i=t_0+1}^{(t_0+t)\wedge\psi\star\xi\wedge\tau_a} \frac{A_i + B_i}{H^{i-t_0}} \le -\frac{a}{2} \right\} \text{ and } \mathcal{B} := \left\{ \min_{1 \le t \le t'} \sum_{i=t_0+1}^{t_0+t} \frac{A_i + B_i}{H^{i-t_0}} \le -\frac{a}{2} \right\}$$

Now, we partition the probability space again using $\mathcal{A}, \mathcal{B}$ and their complements as follows.

$$\begin{aligned}\text{Ⓐ} &= \Pr\left[\tau_a, \psi > t_0 + t', \mathcal{A} \;\middle|\; \xi_T > T, \mathcal{C}^T_{init}\right] + \Pr\left[\tau_a, \psi > t_0 + t', \neg\mathcal{A} \;\middle|\; \xi_T > T, \mathcal{C}^T_{init}\right], \\ \text{Ⓑ} &= \Pr\left[\tau_a > t_0 + t', \psi \le t_0 + t', \mathcal{B} \middle| \xi_T > T, \mathcal{C}^T_{init}\right] + \Pr\left[\tau_a > t_0 + t', \psi \le t_0 + t', \neg\mathcal{B} \middle| \xi_T > T, \mathcal{C}^T_{init}\right]\end{aligned}$$

We show that for both Ⓐ and Ⓑ, the first term is small and the second term is 0.

Claim 19 (The first term is small)**.**

$$\Pr[\mathcal{A} \,|\, \mathcal{C}^T_{init}] < \frac{\delta'}{4} \text{ and } \Pr\left[\mathcal{B}, \tau_a > t_0 + t', \xi_T > T \;\middle|\; \mathcal{C}^T_{init}\right] < \frac{\delta'}{4}$$

Proof of Claim 19. Note that by Lemma 16 we know that $\mathcal{A}$ happens with small probability when conditioning on $\mathcal{C}^T_{init}$ as desired.

As for $\mathcal{B}$, we need to apply the pull-out lemma (see Lemma 17) as follows. It suffices to check that if $\neg\mathcal{B}$ is true, then we have $\psi > t_0 + t$. By Corollary 3 we have for all $t \in [t']$

$$\mathbf{w}^2_{t_0+t,1} \ge H^t \left(\mathbf{w}^2_{t_0,1} + \sum_{i=t_0+1}^{t_0+t} \frac{A_i + B_i}{H^{i-t_0}} \right) \ge H^t (a - \frac{a}{2}) \ge \frac{a}{2}$$

as desired. By Lemma 17, this shows that $\Pr\left[\mathcal{B}, \tau_a > t_0 + t', \xi_T > T \;\middle|\; \mathcal{C}^T_{init}\right] < \delta'/4$ as desired. □

Claim 20 (The second term is 0)**.**

$$\Pr[\tau_a > t_0 + t', \xi_T > T, \psi > t_0 + t', \neg\mathcal{A}] = 0 \text{ and } \Pr[\tau_a > t_0 + t', \xi_T > T, \psi \le t_0 + t', \neg\mathcal{B}] = 0$$

Proof of Claim 20. For the first equality, for the sake of contradiction assuming all the events $\tau_a > t_0 + t', \xi_T > T, \psi > t_0 + t'$ and $\neg\mathcal{A}$ are happening. By Corollary 3, we have

$$\begin{aligned}\mathbf{w}^2_{(t_0+t')\wedge\psi\star\xi,1} &\geq H^{t'\wedge(\psi-t_0)\star(\xi-t_0)}\left(\mathbf{w}^2_{t_0,1} + \sum_{i=t_0+1}^{(t_0+t)\wedge\psi\star\xi}\frac{A_i+B_i}{H^{i-t_0}}\right)\\ &= H^{t'}\left(\mathbf{w}^2_{t_0,1} + \sum_{i=t_0+1}^{(t_0+t)\wedge\psi\star\xi\wedge\tau_a}\frac{A_i+B_i}{H^{i-t_0}}\right) \geq 4(a-\frac{a}{2}) = 2a\end{aligned}$$

which contradicts to $\tau_a > t_0 + t'$. Thus, these events cannot happen simultaneously.

Similarly, for the second equality, notice that when all the events $\tau_a > t_0 + t', \xi_T > T, \tau \leq t_0 + t'$ and $\neg\mathcal{B}$ are happening, by the second condition of Lemma 17 we checked in Claim 19, we have $\tau > t_0 + t'$ which contradicts to $\tau \leq t_0 + t'$. □

To wrap up, by Claim 20 we know that the second term of Ⓐ vanishes and thus

$$Ⓐ = \Pr\left[\tau_a, \psi > t_0 + t', \mathcal{A} \;\middle|\; \xi_T > T, \mathcal{C}^T_{init}\right] \leq \Pr\left[\mathcal{A} \;\middle|\; \xi_T > T, \mathcal{C}^T_{init}\right] \leq \frac{\Pr[\mathcal{A}\,|\,\mathcal{C}^T_{init}]}{\Pr[\xi_T > T\,|\,\mathcal{C}^T_{init}]}$$

As $\Pr[\mathcal{A}\,|\,\mathcal{C}^T_{init}] \leq \delta'/4$ by Claim 19 and $\Pr\left[\xi_T > T \;\middle|\; \mathcal{C}^T_{init}\right] \geq \frac{1}{2}$ from Equation 18, we have $Ⓐ \leq \delta'/2$ as desired. Similarly, as the second term of Ⓑ vanishes by Claim 20, we have

$$Ⓑ = \Pr\left[\tau_a > t_0 + t', \psi \leq t_0 + t', \mathcal{B} \middle| \xi_T > T, \mathcal{C}^T_{init}\right] = \frac{\Pr\left[\tau_a > t_0 + t', \xi_T > T, \psi \leq t_0 + t', \mathcal{B} \middle| \mathcal{C}^T_{init}\right]}{\Pr[\xi_T > T\,|\,\mathcal{C}^T_{init}]}$$

As $\Pr\left[\mathcal{B}, \tau_a > t_0 + t', \xi_T > T \;\middle|\; \mathcal{C}^T_{init}\right] < \delta'/4$ by Claim 19 and $\Pr\left[\xi_T > T \;\middle|\; \mathcal{C}^T_{init}\right] \geq \frac{1}{2}$ from Equation 18, we have $Ⓑ \leq \delta'/2$ as desired. In conclusion, we have

$$\Pr[\tau > t_0 + t'\,|\,\xi > T, \mathcal{C}^T_{init}] \leq Ⓐ + Ⓑ < \delta'$$

□

7.5 Interval Analysis: From Global to Local

In this section, we proceed with the following interval scheme to show the improvement of $\mathbf{w}^2_{t,1}$ from the global regime to the local regime

$$\frac{1}{\Lambda'} \to 2\frac{1}{\Lambda'} \to \cdots \to 2^{\lfloor \log \frac{2\Lambda'}{3} \rfloor}\frac{1}{\Lambda'} \to \frac{2}{3}$$

We first show in Lemma 24 on how to choose the learning rate without dependency on T and then show that $\mathbf{w}^2_{t,1}$ is going to reach $2/3$ efficiently.

Lemma 24 (Choice of parameters in global convergence)**.** *Given t' such that $8 \geq H^{t'} \geq 4$, there exists*

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

such that

$$\eta = \Theta\left(\frac{\lambda_1-\lambda_2}{\lambda_1\Lambda_T^2\log\frac{nT}{\delta}}\right),\ T \geq t'\log\Lambda'$$

Proof of Lemma 24. Since $8 \geq H^{t'} \geq 4$, we have that $t' = \Theta(1/\eta(\lambda_1 - \lambda_2))$. Now

$$t' \log \Lambda' = \Theta\left(\frac{\lambda_1 \log \Lambda' \log^2 \frac{nT}{\delta}}{\delta^2(\lambda_1 - \lambda_2)^2}\right)$$

For notational convenience we let $A = \frac{\lambda_1 \log \Lambda'}{\delta^2(\lambda_1-\lambda_2)^2}$. Then we need $T \geq A \log^2 \frac{nT}{\delta}$ and

$$T = \Theta(A \log^2 nA) = \Theta\left(\frac{\lambda_1 \log \frac{n}{\delta} \log^2 \frac{n\lambda_1}{\delta(\lambda_1-\lambda_2)^2}}{\delta^2(\lambda_1 - \lambda_2)^2}\right)$$

satisfied the requirement as desired. □

Theorem 15 (From global to local)**.** *Let* $n \in \mathbb{N}, \epsilon, \delta \in (0,1)$. *Let* $T = \Theta\left(\frac{\lambda_1 \log \frac{n}{\delta} \log^2 \frac{n\lambda_1}{\delta(\lambda_1-\lambda_2)^2}}{\delta^2(\lambda_1-\lambda_2)^2}\right)$. *Let* τ *be the stopping time of* $\mathbf{w}_{t,1}^2 \geq \frac{2}{3}$. *Let*

$$\eta = O\left(\frac{(\lambda_1 - \lambda_2)}{\lambda_1 \Lambda^2 \log \frac{nT}{\delta}}\right), T_0 = \frac{\lceil \log \frac{2}{3\Lambda'} \rceil + 1}{\eta(\lambda_1 - \lambda_2)}$$

Then we have

$$\Pr\left[\tau > T_0\right] < \delta$$

Proof of Theorem 15. Choose t' such that $8 \geq H^{t'} \geq 4$ and let $m = \lceil \log \frac{2}{3\Lambda'} \rceil + 1$, $v_i = \frac{1}{\Lambda'} 2^i$ for $i = 0, \cdots, m-1$ and $v_m = \frac{2}{3}$. Let τ_{v_i} be the stopping time of $\{\mathbf{w}_{t,1}^2 \geq v_i\}$ and let $T_0 = mt'$. We will apply Lemma 23 with $\delta' = \frac{\delta}{4m}$. Notice that since $\log \Lambda' = O(nT)$, we can choose δ' this way. Now we have

$$\begin{aligned}
\Pr[\tau > T_0] &= \Pr[\tau_{v_m} > mt'] \\
&\leq \Pr[\tau_{v_m} > mt' \,|\, \xi_T > T, \mathcal{C}_{init}^T] + \Pr[\xi \leq T \,|\, \mathcal{C}_{init}^T] + \Pr[\neg \mathcal{C}_{init}^T]
\end{aligned}$$

By Equation 18 and union bound, we have

$$\begin{aligned}
&< \sum_{i=1}^{m} \Pr[\tau_{v_i} > it', \tau_{v_{i-1}} \leq (i-1)t' \,|\, \xi_T > T, \mathcal{C}_{init}^T] + \frac{\delta}{4n^2} + \frac{\delta}{2} \\
&\leq \sum_{i=1}^{m} \Pr[\tau_{v_i} > \tau_{v_{i-1}} + t' \,|\, \xi_T > T, \mathcal{C}_{init}^T] + \frac{\delta}{4n^2} + \frac{\delta}{2}
\end{aligned}$$

By Lemma 23, each summand can be bounded by $\frac{\delta}{4m}$, we have

$$< \frac{\delta m}{4m} + \frac{\delta}{4n^2} + \frac{\delta}{2} \leq \delta$$

as desired. □

7.6 Combining Theorem 15 with the local analysis

In this section, since we have shown that $\mathbf{w}_{t,1}^2$ efficiently reaches $2/3$ in Theorem 15, by combining Theorem 15, the local convergence (Theorem 10) and the finite continual learning (Theorem 11), we derive Theorem 13.

Proof of Theorem 13. Let τ to be the hitting time of $\mathbf{w}_{t,1}^2 > 1 - \frac{\epsilon}{2}$. With

$$\eta = \Theta\left(\frac{\lambda_1 - \lambda_2}{\lambda_1} \cdot \left(\frac{\epsilon}{\log \frac{\log \frac{n}{\epsilon}}{\delta}} \bigwedge \frac{\delta^2}{\log^2 \frac{\lambda_1 n}{\delta(\lambda_1 - \lambda_2)^2}}\right)\right)$$

we can apply Theorem 15, Theorem 10 to get that

$$\Pr[\tau > T] < \frac{\delta}{2}$$

where $T = \Theta(\frac{\log \frac{1}{\epsilon} + \log \Lambda'}{\eta(\lambda_1 - \lambda_2)}) = \Theta(\frac{\log \frac{1}{\epsilon} + \log \frac{n}{\delta}}{\eta(\lambda_1 - \lambda_2)})$. Now we initialize Theorem 11 with $t_0 = \Theta(\log \frac{1}{\epsilon} + \log \frac{n}{\delta})$ with failure probability $\frac{\delta}{2}$ to get

$$\Pr[\exists 1 \le t \le T, \mathbf{w}_{\tau+t,1}^2 < 1 - \epsilon] < \frac{\delta}{2}$$

Since $T \in [\tau, \tau + T]$ if $\tau \le T$, now by union bounding two inequalities, we have

$$\Pr[\mathbf{w}_{T,1}^2 < 1 - \epsilon] < \delta$$

□

8 The Cross Term is Small

In this section, we will prove Theorem 14 to finish the proof of the global convergence. We recall the motivation again. In order to keep the bounded differences of the noise in the global convergence small, we need to make $f_{t,n}(\mathbf{w}_{t-1})$ small with high probability. Concretely, the stopping time $\xi_{p,\delta}$ stops whenever $|f_{t,n}(\mathbf{w}_{(t-1)\wedge\psi_{p,\delta}})| > 2\Lambda_{p,\delta}$ (see Definition 8). Therefore, the main goal of this section is to show that $\xi_{p,\delta}$ is large with high probability in Theorem 14.

Theorem 14 (ξ is large with high probability)**.** *Let $T \in \mathbb{N}$ and $p, \delta \in (0,1)$. Let $\eta = \Theta\left(\frac{(\lambda_1 - \lambda_2)}{\lambda_1 \Lambda_{p,\delta/4n^2T}^2 \log \frac{nT}{\delta}}\right)$. If we have $T = \Omega\left(\frac{1}{\eta\lambda_1}\right)$ and $p \le \delta$, then we have*

$$\forall t \in [T],\ \Pr\left[\xi = t \;\middle|\; \mathcal{C}_{init}^{p,\frac{\delta}{4n^2T}}\right] < \frac{\delta}{2n^2T}$$

In particular we have

$$\forall t \in [T],\ \Pr\left[\xi = t \;\middle|\; \xi \ge t, \mathcal{C}_{init}^{p,\frac{\delta}{4nT}}\right] \le \frac{\delta}{n^2T}$$

and

$$\Pr\left[\xi \le T \;\middle|\; \mathcal{C}_{init}^{p,\frac{\delta}{4nT}}\right] < \frac{\delta}{2n^2}$$

In order to prove Theorem 14, we consider the auxiliary processes $(f_{t,2}, \dots, f_{t,n})$ (see Definition 7) as a vector and use a vector linearization as well as the ODE trick in Corollary 4 and Corollary 5. We then obtain the concentration on the stopped processes in Lemma 26. Finally, to prove the main theorem by induction, we prove the induction step in Lemma 29 by carefully pulling out the stopping time to finish the proof.

To bound the stopping time ξ, we need to show the concentration of $f_{t,j}(\mathbf{w}_{(t-1)\wedge\psi})$ and as before the linearization and the ODE trick would be our main tools.

8.1 Linearization and ODE trick for controlling $|y_t|$

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

Lemma 25 (Linearization)**.** *Let $t \in [T]$, $s \in [t-1]$. Let $\mathbf{w}_s = \mathbf{w}_{s-1} + \eta \mathbf{z}_s$ where $\mathbf{z}_s = y_s(\mathbf{x}_s - y_s \mathbf{w}_{s-1})$. Then there exists $\overline{\mathbf{w}}_{s-1} = \mathbf{w}_{s-1} + c\eta\mathbf{z}_s$ for some $c \in [0,1]$ such that for all j, $2 \le j \le n$,*

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

where

$$A_{s,j}^{(t)} = \eta \nabla f_{t,j}(\mathbf{w}_{s-1})^\top (\mathbf{z}_s - \mathbb{E}[\mathbf{z}_s \mid \mathcal{F}_{s-1}]) + \eta^2 \mathbf{z}_s^\top \nabla^2 f_{t,j}(\overline{\mathbf{w}}_{s-1}) \mathbf{z}_s$$

Proof of Lemma 25. This is a direct application of Taylor expansion. Concretely, there exists $\overline{\mathbf{w}}_{s-1} = \mathbf{w}_{s-1} + c\eta\mathbf{z}_s$ for some $c \in [0,1]$ such that

$$f_{t,j}(\mathbf{w}_s) = f_{t,j}(\mathbf{w}_{s-1}) + \eta \nabla f_{t,j}(\mathbf{w}_{s-1})^\top \mathbf{z}_s + \eta^2 \mathbf{z}_s^\top \nabla^2 f_{t,j}(\overline{\mathbf{w}}_{s-1}) \mathbf{z}_s$$

Note that $\frac{\partial f_{t,j}(\mathbf{w})}{\partial \mathbf{w}_1} = -\frac{f_{t,j}(\mathbf{w})}{\mathbf{w}_1}$ and $\frac{\partial f_{t,j}(\mathbf{w})}{\partial \mathbf{w}_i} = \frac{\mathbf{1}_{i \le j} \cdot \mathbf{x}_{t,i}}{\mathbf{w}_1}$ for $i = 2, \dots, n$.We have

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

Next, recall that $\mathbb{E}[\mathbf{z}_{s,i} \mid \mathcal{F}_{s-1}] = (\lambda_i - \mathbf{w}_{s-1}^\top \mathsf{diag}(\lambda) \mathbf{w}_{s-1}) \cdot \mathbf{w}_{s-1,i}$. By adding and subtracting the expectations, the equation becomes

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

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

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

as desired. □

We can write the above lemma in a vector form. For any $t \in [T]$, let $\mathbf{f}_t(\mathbf{w}), \mathbf{A}_s^{(t)} \in \mathbb{R}^{n-1}$ be $(n-1)$-dimensional vectors where the i^{th} coordinates of them are $f_{t,i+1}(\mathbf{w})$, $A_{s,i+1}^{(t)}$ respectively. The following is an immediate corollary of Lemma 25 by rewriting everything into a vector form.

Corollary 4 (Linearization in a vector form)**.** *For any $t \in [T]$ and $s \in [t-1]$, we have*

$$\mathbf{f}_t(\mathbf{w}_s) = H\mathbf{f}_t(\mathbf{w}_{s-1}) + \mathbf{A}_s^{(t)}$$

where

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

By the ODE trick for vector (see Lemma 8), we immediately have the following corollary for a closed form solution to $\mathbf{f}_t(\mathbf{w}_s)$.

Corollary 5 (ODE trick)**.** *For any $t \in [T]$, $s \in [t-1]$, we have*

$$\mathbf{f}_t(\mathbf{w}_s) = H^s\mathbf{f}_t(\mathbf{w}_0) + \sum_{s'=1}^{s} H^{s-s'}\mathbf{A}_{s'}^{(t)}$$

8.2 Concentration of the noise term

We want to control the noise term in Corollary 5. However, same as the situation before, we cannot get the concentration for the noise terms of the ODE trick directly. As a consequence, we have to introduce a new stopping time τ_t to make sure the bounded difference of the stopped processes are small enough for the martingale concentration inequality.

For a fixed $t \in [T]$, we define a stopping time τ_t for the noise terms from $s = 1, 2, \ldots, t-1$ 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 $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_s\} \cup \{\mathbf{x}_t\}$. Note that $\{\mathcal{F}_s^{(t)}\}_{s\in[t-1]}$ is well-defined and $\{A_{t,s,j}\}_{s\in[t-1]}$ is an adapted random process with respect to $\{\mathcal{F}_s^{(t)}\}_{s\in[t-1]}$, *i.e.,* $A_{t,s,j}$ lies 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 $\{\mathcal{F}_s^{(t)}\}$. The following is the definition of the stopping time τ_t.

Definition 9 (Stopping time for $\mathbf{f}_t$)**.** *Let $T \in \mathbb{N}_{\geq 0}, p, \delta \in (0,1)$. For every $t \in [T]$, let $\Lambda_{p,\delta}$ be the parameter specified in Lemma 18 and ξ, ψ be the stopping times specified in Definition 8. Define τ_t to be the stopping time for the first s such that $\{\|\mathbf{f}_t(\mathbf{w}_{s\wedge\psi\star\xi})\|_\infty > 2\Lambda_{p,\delta}\}$.*

The goal of this subsection is to show the concentration of the stopped noise vector as follows.

Lemma 26 (Concentration for the stopped noise term)**.** *Let $T \in \mathbb{N}_{\geq 0}, p, \delta, \delta' \in (0,1), t \in [T]$. Let $\Lambda_{p,\delta}$ be the parameter specified before and ξ, ψ, τ_t be the stopping times as chosen before. Let $\eta = \Theta\left(\frac{(\lambda_1-\lambda_2)}{\lambda_1\Lambda_{p,\delta}^2 \log\frac{1}{\delta'}}\right)$. If $T = \Omega(\frac{1}{\eta\lambda_1}), p \leq \delta$ and the following condition is true*

$$\forall 1 \leq t' \leq t-1,\ \Pr[\xi = t' \,|\, \xi \geq t', \mathcal{C}_{init}^{p,\delta}] \leq \frac{1}{n^2 T}\,,$$

then for all $\overline{s} \in [t-1]$,

$$\Pr\left[\exists i \in [n-1],\ \sum_{s=1}^{\overline{s}\wedge\psi\star\xi_{p,\delta}\wedge\tau_t} \left(H^{\overline{s}-s}\mathbf{A}_{t,s}\right)_i \geq \Lambda_{p,\delta} \mid \mathcal{C}_{init}^{p,\delta}\right] < n\delta'$$

To enable martingale concentration, we have to bound the three moment quantities of the martingale difference of $M_{t,s} \in \mathbb{R}^{n-1}$ where $M_{t,s,j}$ is the j^{th} entry of $\sum_{s'=1}^{s\wedge\psi\star\xi\wedge\tau_t} H^{\overline{s}-s'}\mathbf{A}_{t,s'}$ for every $s \in [\overline{s}]$ and $j \in [n-1]$. By Lemma 4 and Lemma 5, the difference can be rewritten as

$$M_{t,s} - M_{t,s-1} = \mathbf{1}_{\tau_t \geq s\wedge\psi\star\xi,\xi>s\wedge\psi,\psi\geq s} H^{\overline{s}-s}\mathbf{A}_{t,s} = \mathbf{1}_{\tau_t,\psi\geq s,\xi>s} H^{\overline{s}-s}\mathbf{A}_{t,s}$$

Before we bound the bounded differences and the moments for $\mathbf{1}_{\tau_t,\psi\geq s,\xi>s}H^{\overline{s}-s}\mathbf{A}_{t,s}$, recall that in Lemma 19 we showed that the conditional expectation of $\mathbf{x}_{s,i}\mathbf{x}_{s,j}$ would not change by too much when conditioning on the event $\xi > s$. Let us start with bounding the three moment quantities of the stopped process of $\mathbf{A}_{t,s}$ in Lemma 27 and then extend to that of $H^{\overline{s}-s}\mathbf{A}_{t,s}$ in Lemma 28.

Lemma 27 (Structure of the stopped $\mathbf{A}_{t,s}$)**.** *Let* $T \in \mathbb{N}, \eta \in (0,1), t \in [T]$ *and* $s \in [t-1]$*. Let* Λ *be the parameter specified before and* ξ, ψ, τ_t *be the stopping times as chosen before. If* $\eta = O\left(\frac{1}{\Lambda}\right), T = \Omega(\frac{1}{\eta\lambda_1}), p \leq \delta$ *and the following condition holds*

$$\forall 1 \leq t' \leq t-1,\ \Pr[\xi = t' \mid \xi \geq t', \mathcal{C}_{init}^{p,\delta}] \leq \frac{1}{n^2 T} \tag{21}$$

then the following holds almost surely.

- *(Bounded difference) We have*

$$\left\|\mathbf{1}_{\tau_t,\psi\geq s,\xi>s}\mathbf{A}_s^{(t)}\right\|_\infty = O(\eta\Lambda^2)$$

- *(Conditional expectation) We have*

$$\left\|\mathbb{E}[\mathbf{1}_{\tau_t,\psi\geq s,\xi>s}\mathbf{A}_s^{(t)} \mid \mathcal{F}_{s-1}^{(t)}, \mathcal{C}_{init}^{p,\delta}]\right\|_\infty = O(\eta^2\lambda_1\Lambda^3)$$

- *(Conditional variance) We have*

$$\left\|\mathbb{E}[\mathbf{1}_{\tau_t,\psi\geq s,\xi>s}\mathbf{A}_s^{(t)}\mathbf{A}_s^{(t)\top} \mid \mathcal{F}_{s-1}^{(t)}, \mathcal{C}_{init}^{p,\delta}]\right\|_{max} = O(\eta^2\lambda_1\Lambda^4)$$

where the $\|\cdot\|_{max}$ *is the entrywise maximum of a matrix.*

Proof of Lemma 27. The proof is basically direct verification using the definition of stopping time and Lemma 19. We postpone the proof to Appendix D. □

Now note that given $\overline{s} \in [t-1]$ the stopped process $\left\{\sum_{s'=1}^{s\wedge\psi\star\xi\wedge\tau_t} H^{\overline{s}-s'}\mathbf{A}_{t,s'}\right\}_{s\in[\overline{s}]}$ is an adapted stochastic process with respect to $\{\mathcal{F}_s^{(t)}\}_{s\in[\overline{s}]}$. Furthermore, it has small bounded difference and moments. Concretely we have the following.

Lemma 28 (Structure of the stopped $H^{\overline{s}-s}\mathbf{A}_{t,s}$)**.** *Let $T \in \mathbb{N}, \eta, \delta \in (0,1), t \in [T], \overline{s} \in [t-1]$. Let Λ be the parameter specified before and ξ, τ_t be the stopping times as chosen before. For any $s \in [\overline{s}]$ and $j \in [n-1]$, let $M_{t,s,j}$ be the j^{th} entry of $\sum_{s'=1}^{s\wedge\psi\star\xi\wedge\tau_t} H^{\overline{s}-s'}\mathbf{A}_{t,s'}$. If $\eta = O\left(\frac{1}{\Lambda}\right), T = \Omega\left(\frac{1}{\eta\lambda_1}\right), p \leq \delta$ and the following condition is true*

$$\forall 1 \leq t' \leq t-1,\ \Pr[\xi = t' \,|\, \xi \geq t', \mathcal{C}_{init}^{p,\delta}] \leq \frac{1}{n^2 T},$$

then the following holds.

- *(Bounded difference) For any $j \in [n-1]$, we have*

$$\max_{s\in[\overline{s}]} |M_{t,s,j} - M_{t,s-1,j}| = O(\eta\Lambda^2) \text{ almost surely.}$$

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

$$\sum_{s=1}^{\overline{s}} \mathbb{E}\left[M_{t,s,j} - M_{t,s-1,j} \;\middle|\; \mathcal{F}_{s-1}^{(t)}, \mathcal{C}_{init}^{p,\delta}\right] = O\left(\frac{\eta\lambda_1\Lambda^3}{\lambda_1-\lambda_2}\right)$$

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

$$\sum_{s=1}^{\overline{s}} \mathsf{Var}\left[M_{t,s,j} - M_{t,s-1,j} \;\middle|\; \mathcal{F}_{s-1}^{(t)}, \mathcal{C}_{init}^{p,\delta}\right] = O\left(\frac{\eta\lambda_1\Lambda^4}{\lambda_1-\lambda_2}\right)$$

Proof of Lemma 28. For notational convenience, given a matrix A, we will denote its j^{th} row as $A_{(j)}$ for the rest of the proof. First, notice that the multiplier matrix $H = VDV^{-1}$ is invertible where

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

and $D = \mathsf{diag}(d_1, d_2, \ldots, d_{n-1})$ such that $d_i = 1 - \eta(\lambda_1 - \lambda_{i+1})$ for every $i = 1, 2, \ldots, n-1$. To see the above, observe that for any diagonal matrix $D' = \mathsf{diag}(d'_1, d'_2, \ldots, d'_{n-1})$, 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 $d'_1 \geq d'_2 \geq \cdots \geq d'_{n-1} \geq 0$, then we have

$$\| \left(VD'V^{-1}\right)_{(i)} \|_1 = d'_i + \sum_{j=1}^{i-1} d'_j - d'_{j+1} = d'_1 \tag{22}$$

- (Bounded difference) Fixed $j \in [n]$. First we have for all $s \in [\overline{s}]$,

$$|M_{t,s,j} - M_{t,s-1,j}| = |\mathbf{1}_{\tau_t,\psi \geq s,\xi > s} H^{\overline{s}-s} \mathbf{A}_{t,s}| \leq O(\eta \Lambda^2)$$

by Equation 22.

- (Conditional expectation) Similarly, we have

$$\begin{aligned}\mathbb{E}\left[M_{t,s,j} - M_{t,s-1,j} \;\middle|\; \mathcal{F}_{s-1}^{(t)}, \mathcal{C}_{init}^{p,\delta}\right] &= \mathbb{E}\left[\mathbf{1}_{\tau_t,\psi \geq s,\xi > s} H^{\overline{s}-s} \mathbf{A}_{t,s} \;\middle|\; \mathcal{F}_{s-1}^{(t)}, \mathcal{C}_{init}^{p,\delta}\right] \\ &\leq \left\|H_{(j)}^{\overline{s}-s}\right\|_1 \left\|\mathbb{E}\left[\mathbf{1}_{\tau_t,\psi \geq s,\xi > s} \mathbf{A}_{t,s} \;\middle|\; \mathcal{F}_{s-1}^{(t)}, \mathcal{C}_{init}^{p,\delta}\right]\right\|_\infty\end{aligned}$$

By Equation 22 and Lemma 27, we have

$$\leq (1 - \eta(\lambda_1 - \lambda_2))^{\overline{s}-s} \cdot O(\eta^2 \lambda_1 \Lambda^3)$$

So by geometric series, we have

$$\sum_{s=1}^{\overline{s}} \mathbb{E}\left[M_{t,s,j} - M_{t,s-1,j} \;\middle|\; \mathcal{F}_{s-1}^{(t)}, \mathcal{C}_{init}^{p,\delta}\right] = O\left(\frac{\eta \lambda_1 \Lambda^3}{\lambda_1 - \lambda_2}\right)$$

- (Conditional variance) Similarly, we have

$$\begin{aligned}&\mathsf{Var}\left[M_{t,s,j} - M_{t,s-1,j} \;\middle|\; \mathcal{F}_{s-1}^{(t)}, \mathcal{C}_{init}^{p,\delta}\right] \\ &= \mathbb{E}\left[\mathbf{1}_{\tau_t,\psi \geq s,\xi > s} (H^{\overline{s}-s} \mathbf{A}_{t,s})^2 \;\middle|\; \mathcal{F}_{s-1}^{(t)}, \mathcal{C}_{init}^{p,\delta}\right] \\ &= (H_{(j)}^{\overline{s}-s})^\top \mathbb{E}\left[\mathbf{1}_{\tau_t,\psi \geq s,\xi > s} \mathbf{A}_{t,s} \mathbf{A}_{t,s}^\top \;\middle|\; \mathcal{F}_{s-1}^{(t)}, \mathcal{C}_{init}^{p,\delta}\right] (H_{(j)}^{\overline{s}-s}) \\ &\leq \left\|H_{(j)}^{\overline{s}-s}\right\|_1 \cdot \left\|\mathbb{E}\left[\mathbf{1}_{\tau_t,\psi \geq s,\xi > s} \mathbf{A}_{t,s} \mathbf{A}_{t,s}^\top \;\middle|\; \mathcal{F}_{s-1}^{(t)}, \mathcal{C}_{init}^{p,\delta}\right]\right\|_{max} \cdot \left\|H_{(j)}^{\overline{s}-s}\right\|_1\end{aligned}$$

By Equation 22 and Lemma 27, we have

$$\leq (1 - \eta(\lambda_1 - \lambda_2))^{2(\overline{s}-s)} \cdot O(\eta^2 \lambda_1 \Lambda^4)$$

So by geometric series, we have

$$\sum_{s=1}^{\overline{s}} \mathsf{Var}\left[M_{t,s,j} - M_{t,s-1,j} \;\middle|\; \mathcal{F}_{s-1}^{(t)}, \mathcal{C}_{init}^{p,\delta}\right] = O\left(\frac{\eta \lambda_1 \Lambda^4}{\lambda_1 - \lambda_2}\right)$$

□

As a consequence of Lemma 28, we are able to prove the concentration for the stopped noise term.

Proof of Lemma 26. The proof is based on applying the corollary of Freedman's inequality (see Corollary 1) on each coordinate using Lemma 28. We have

$$\Pr\left[\sum\nolimits_{s=1}^{\overline{s} \wedge \psi \star \xi_{p,\delta} \wedge \tau_t} \left(H^{\overline{s}-s} \mathbf{A}_{t,s}\right)_i \geq \Lambda_{p,\delta} \mid \mathcal{C}_{init}^{p,\delta}\right] < \delta'$$

by noticing that the deviation term is $O(\sqrt{\frac{\eta \lambda_1 \Lambda_{p,\delta}^4 \log \frac{1}{\delta'}}{\lambda_1 - \lambda_2}}) < \frac{\Lambda_{p,\delta}}{2}$ and the sum of conditional expectation term is $O(\frac{\eta \lambda_1 \Lambda_{p,\delta}^3}{\lambda_1 - \lambda_2}) < \frac{\Lambda_{p,\delta}}{2}$. Now we obtain the desired inequality by union bounding over $i \in [n-1]$. □

8.3 Wrap up

First fix δ, δ' in the Lemma 26 as $\frac{\delta}{4n^2T}$, $\frac{\delta}{4n^3T^2}$ respectively. The following lemma proves the inductive step toward the main theorem.

Lemma 29 (Inductive step)**.** *Let $T \in \mathbb{N}_{\geq 0}, p, \delta \in (0,1)$ be the parameters and ξ be the stopping times as chosen before. Let $\eta = \Theta\left(\frac{(\lambda_1-\lambda_2)}{\lambda_1 \Lambda^2_{p,\delta/4n^2T} \log \frac{nT}{\delta}}\right)$. If $T = \Omega(\frac{1}{\eta\lambda_1})$, $t \in [T]$, $p \leq \delta$ and for every $t' \in [t-1]$,*

$$\Pr\left[\xi = t' \,\middle|\, \mathcal{C}_{init}^{p,\frac{\delta}{4n^2T}}\right] < \frac{\delta}{2n^2T}$$

then

$$\Pr\left[\xi = t \,\middle|\, \mathcal{C}_{init}^{p,\frac{\delta}{4n^2T}}\right] < \frac{\delta}{2n^2T}$$

Proof of Lemma 29. Fix $T \in \mathbb{N}$ such that $T = \Omega(\frac{1}{\eta\lambda_1})$ and $t \in [T]$, we have

$$\begin{aligned}\Pr\left[\xi = t \,\middle|\, \mathcal{C}_{init}^{p,\frac{\delta}{4n^2T}}\right] &= \Pr\left[|f_{t,n}(\mathbf{w}_{(t-1)\wedge\psi\star\xi})| > 2\Lambda, \xi \geq t \,\middle|\, \mathcal{C}_{init}^{p,\frac{\delta}{4n^2T}}\right] \\ &\leq \Pr\left[\tau_t < t \,\middle|\, \mathcal{C}_{init}^{p,\frac{\delta}{4n^2T}}\right]\end{aligned}$$

where the last inequality is due to Definition 9. Intuitively, when the noise term of the linearization for $\mathbf{f}_t$ is small, then we expect τ_t would not stop by the ODE trick. Formally, denote the event where the noise term is large as $\mathcal{A}_{s_0} = \{\exists i \in [n], \sum_{s=1}^{s_0\wedge\psi\star\xi} (H^{s_0-s}\mathbf{A}_{t,s})_i \geq \Lambda\}$ and denote its stopped version as $\mathcal{A}_{s_0}^{\tau_t} = \{\exists i \in [n], \sum_{s=1}^{s_0\wedge\psi\star\xi\wedge\tau_t} (H^{s_0-s}\mathbf{A}_{t,s})_i \geq \Lambda\}$. Recall from Lemma 18 that $\mathcal{C}_0^{p,\frac{\delta}{4n^2T}}$ is the event

$$\{\exists j \in [n],\ t \in [T],\ |f_{t,j}(\mathbf{w}_0)| > \Lambda_{p,\delta}\}$$

Now, we partition the probability space and rewrite the error probability as follows.

$$\begin{aligned}\Pr\left[\tau_t < t \,\middle|\, \mathcal{C}_{init}^{p,\frac{\delta}{4n^2T}}\right] &\leq \Pr\left[\tau_t < t, \neg\mathcal{C}_0^{p,\frac{\delta}{4n^2T}}, \neg\mathcal{A}_{t-1} \,\middle|\, \mathcal{C}_{init}^{p,\frac{\delta}{4n^2T}}\right] + \Pr\left[\mathcal{A}_{t-1} \cup \mathcal{C}_0^{p,\frac{\delta}{4n^2T}} \,\middle|\, \mathcal{C}_{init}^{p,\frac{\delta}{4n^2T}}\right] \\ &= \text{Ⓐ} + \text{Ⓑ}\end{aligned}$$

Notice that the concentration analysis in the previous subsection only works for the *stopped version*, *i.e.*, $\mathcal{A}_{s_0}^{\tau_t}$. To go from $\mathcal{A}_{s_0}$ to $\mathcal{A}_{s_0}^{\tau_t}$, observe that the events $\neg\mathcal{C}_0^{p,\frac{\delta}{4n^2T}}$ and $\neg\mathcal{A}_{s_0}$ imply the noise term to be small and thus the ODE trick (see Corollary 5) gives the following useful equality for every $1 \leq s_0 \leq t-1$

$$\Pr\left[\tau_t \geq s_0 + 1 \,\middle|\, \neg\mathcal{C}_0^{p,\frac{\delta}{4n^2T}}, \neg\mathcal{A}_{s_0}\right] = 1 \tag{23}$$

From Equation 23, we immediately have Ⓐ $= 0$. As for Ⓑ, by the chain rule of expectation, we have

$$\text{Ⓑ} = \sum_{s=1}^{t-1} \Pr\left[\mathcal{A}_s, \neg\mathcal{A}_{s-1}, \neg\mathcal{C}_0^{p,\frac{\delta}{4n^2T}} \,\middle|\, \mathcal{C}_{init}^{p,\frac{\delta}{4n^2T}}\right] + \Pr\left[\mathcal{C}_0^{p,\frac{\delta}{4n^2T}} \,\middle|\, \mathcal{C}_{init}^{p,\frac{\delta}{4n^2T}}\right]$$

Notice that the last term of the above equation can be upper bounded by $\delta/(4n^2T)$ due to Lemma 18. As for each term in the summation, by Equation 23, we have

$$\Pr\left[\mathcal{A}_s, \neg\mathcal{A}_{s-1}, \neg\mathcal{C}_0^{p,\frac{\delta}{4n^2T}} \;\middle|\; \mathcal{C}_{init}^{p,\frac{\delta}{4n^2T}}\right] = \Pr\left[\mathcal{A}_s^{\tau_t}, \neg\mathcal{A}_{s-1}, \neg\mathcal{C}_0^{p,\frac{\delta}{4n^2T}} \;\middle|\; \mathcal{C}_{init}^{p,\frac{\delta}{4n^2T}}\right] \leq \Pr\left[\mathcal{A}_s^{\tau_t} \;\middle|\; \mathcal{C}_{init}^{p,\frac{\delta}{4n^2T}}\right]$$

Finally, to invoke the concentration we proved in the previous subsection (*i.e.* Lemma 26), we have to verify the condition as follows. For every $1 \leq t' \leq t-1$,

$$\Pr[\xi = t' \,|\, \xi \geq t', \mathcal{C}_{init}^{p,\frac{\delta}{4n^2T}}] \leq \frac{\Pr[\xi = t' \,|\, \mathcal{C}_{init}^{p,\frac{\delta}{2nT}}]}{1 - \Pr[\xi < t' \,|\, \mathcal{C}_{init}^{p,\frac{\delta}{4n^2T}}]} < \frac{\frac{\delta}{4n^2T}}{1 - \frac{(t-1)\delta}{2nT}} \leq \frac{1}{n^2T}$$

where the inequalities hold due to the induction hypothesis in the lemma statement. As a result, Lemma 26 gives $\Pr[\mathcal{A}_s^{\tau_t} \,|\, \mathcal{C}_{init}^{p,\frac{\delta}{4n^2T}}] < (t-1)n\delta/(4n^3T^2)$. This gives us

$$\text{Ⓑ} \leq \sum_{s=1}^{t-1} \frac{(t-1)n\delta}{4n^3T^2} + \frac{\delta}{4n^2T} < \frac{\delta}{2n^2T}$$

Recall that Equation 23 implies $\text{Ⓐ} = 0$, we conclude that

$$\Pr\left[\xi = t \,\middle|\, \mathcal{C}_{init}^{p,\frac{\delta}{4n^2T}}\right] \leq \text{Ⓐ} + \text{Ⓑ} < \frac{\delta}{2n^2T}$$

as desired.

□

Now the main theorem can be derived as a corollary.

Theorem 14 (ξ is large with high probability)**.** *Let $T \in \mathbb{N}$ and $p, \delta \in (0,1)$. Let $\eta = \Theta\left(\frac{(\lambda_1-\lambda_2)}{\lambda_1\Lambda^2_{p,\delta/4n^2T}\log\frac{nT}{\delta}}\right)$. If we have $T = \Omega\left(\frac{1}{\eta\lambda_1}\right)$ and $p \leq \delta$, then we have*

$$\forall t \in [T],\ \Pr\left[\xi = t \;\middle|\; \mathcal{C}_{init}^{p,\frac{\delta}{4n^2T}}\right] < \frac{\delta}{2n^2T}$$

In particular we have

$$\forall t \in [T],\ \Pr\left[\xi = t \;\middle|\; \xi \geq t, \mathcal{C}_{init}^{p,\frac{\delta}{4nT}}\right] \leq \frac{\delta}{n^2T}$$

and

$$\Pr\left[\xi \leq T \;\middle|\; \mathcal{C}_{init}^{p,\frac{\delta}{4nT}}\right] < \frac{\delta}{2n^2}$$

Proof of Theorem 14. The proof proceed by induction. For the base case, we have

$$\Pr[\xi = 1 \,|\, \mathcal{C}_{init}^{p,\frac{\delta}{4n^2T}}] = \Pr\left[|f_{1,j}(\mathbf{w}_0)| > 2\Lambda \;\middle|\; \mathcal{C}_{init}^{p,\frac{\delta}{4n^2T}}\right] < \frac{\delta}{2n^2T}$$

The induction step is exactly Lemma 29 and this gives us the first conclusion. The second conclusion can be obtained from union bounding over T. □

9 Conclusions and Future Directions

In this work, we provide the first convergence rate analysis for the biological version of Oja's rule in solving streaming principal component analysis (PCA). Our results show that the convergence rate is nearly optimal, matching the information-theoretic lower bound up to logarithmic factors, and outperforming the state-of-the-art for streaming PCA, including machine learning variants of Oja's rule. This offers strong theoretical evidence that the biological Oja's rule functions efficiently within biologically realistic time scales, particularly in neural systems like the retina-optical nerve pathway.

In addition to this key result, we introduce a novel ODE-inspired framework to analyze stochastic dynamics. Unlike traditional step-by-step methods, our one-shot approach provides a closed-form solution for the entire dynamic, allowing for precise control using stopping time and martingale techniques. We believe this framework captures the intrinsic behaviors of stochastic systems and has the potential to simplify and extend analysis to a broader class of problems involving stochastic dynamics.

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

9.1 Biological aspects

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 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 subsection 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 an 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.

Convergence rate analysis for biologically-plausible learning rules for online k-PCA In this work, we focus on 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]. As mentioned in subsection 1.4, computational neuroscientists have proposed several variants of biological Oja's rule to solve streaming k-PCA [Oja92, San89, Fö89, Lee91, RT89, KDT94, PHC15]. Some networks use feedforward connections only but the learning rules are not local [Oja92, San89] while some use Hebbian learning on the feedforward connection and use anti-Hebbian learning on the recurrent connection to decorrelate the outputs [Fö89, Lee91, RT89, KDT94, PHC15]. However, there is no convergence rate analysis for these networks and even the results on the global convergence in the limit are not known for most of these networks. Therefore, it will be interesting to apply our framework to derive a convergence rate analysis for these biologically-plausible learning rules to solve online k-PCA.

9.2 Algorithmic aspects

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 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 $O\left(\lambda_1\epsilon^{-1}\mathsf{gap}^{-2}\right)$ 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.

References

[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] Larry 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.

[AR90] Joseph J. Atick and A. Norman Redlich. Towards a theory of early visual processing. *Neural Computation*, 2:308–320, 1990.

[AR92] Joseph J. Atick and A. Norman Redlich. What does the retina know about natural scenes? *Neural Computation*, 4:196–210, 1992.

[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.

[BM02] Stephen A. Baccus and Markus Meister. Fast and slow contrast adaptation in retinal circuitry. *Neuron*, 36:909–919, 2002.

[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, Wlodzimierz Kasprzak, and Wladyslaw 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 online unsupervised learning machine for adaptive feature extraction. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 41(2):87–98, 1994.

[CWY20] Chi-Ning Chou, Mien Brabeeba Wang, and Tiancheng Yu. A general framework for analyzing stochastic dynamics in learning algorithms. *arXiv: 2006.06171*, 2020.

[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.

[Fö89] Peter Földiák. Adaptive network for optimal linear feature extraction. *International 1989 Joint Conference on Neural Networks*, pages 401–405, 1989.

[GS12] Surya Ganguli and Haim Sompolinsky. Compressed sensing, sparsity and neural data. *Annual review of neuroscience*, 35(1):463–483, 2012.

[HBM05] Toshihiko Hosoya, Stephen A. Baccus, and Markus Meister. Dynamic predictive coding by the retina. *Nature*, 436:71–77, 2005.

[Heb49] Donald O. Hebb. *The organization of behavior: A neuropsychological theory*. Wiley, New York, June 1949.

[HK92] Kurt Hornik and Chung-Ming Kuan. Convergence analysis of local feature extraction algorithms. *Neural Networks*, 5:229–240, 1992.

[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. *27th Annual European Symposium on Algorithms*, 2019.

[HvH98] M. Haft and J. Leo van Hemmen. 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.

[KC78] Harold. J. Kushner and Dean S. Clark. *Stochastic approximataon for constrained and unconstrained systems*. Springer, Berlin, 1978.

[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.

[Lee91] Tood K. Leen. Dynamics of learning in linear feature-discovery networks. *Network*, 2(1):85–105, 1991.

[LG16] Jean-François Le Gall. *Brownian motion, martingales, and stochastic calculus*, volume 274. Springer, 2016.

[LM18] Nancy A. Lynch and Cameron Musco. A basic compositional model for spiking neural networks. *arXiv preprint arXiv:1808.03884*, 2018.

[LMP17a] 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.

[LMP17b] 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.

[LMP17c] Nancy A. 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.

[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 A. 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.

[LW19] Nancy A. Lynch and Mien Brabeeba Wang. Integrating temporal information to spatial information in a neural circuit. *arXiv preprint arXiv:1903.01217*, 2019.

[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] Karl 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. Plumbley. 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.

[RT89] Jeanne Rubner and Paul Tavan. A self-organizing network for principal-component analysis. *Europhysics Letters (EPL)*, 10(7):693–698, 1989.

[SA06] Christian D. Swinehart and Larry F. 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.

[SBW+97] Stellos M. Smirnakis, Michael J. Berry, David K. Warland, William Bialek, and Markus Meister. Adaptation of retinal processing to image contrast and spatial scale. *Nature*, 386:69–73, 1997.

[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.

[SEC84] Robert Shapley and Christina Enroth-Cugell. Visual adaptation and retinal gain controls. *Progress in Retinal Research*, 3:263–346, 1984.

[Sej77] Terrence J. Sejnowski. Storing covariance with nonlinearly interacting neurons. *Journal of 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.

[SLS77] Allan W. Snyder, Simon B. Laughlin, and Doekele G. Stavenga. Information capacity of the eye. *Vision Research*, 17:1163—75, 1977.

[SLY06] Lifeng Shang, Jian Cheng Lv, and Zhang Yi. Rigid medical image registration using pca neural network. *Neurocomputing*, 69(13-15):1717–1722, 2006.

[Wan95] Brian A. Wandell. *Foundations of vision*. Sunderland, MA: Sinauer, 1995.

[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.

[XOS92] Lei Xu, Erkki Oja, and Ching Y. Suen. Modified hebbian learning for curve and surface fitting. *Neural Networks*, 5(3):441–457, 1992.

[Yan98] Wei-Yong Yan. Stability and convergence of principal component learning algorithms. *SIAM Journal on Matrix Analysis and Applications*, 19(4):933–955, 1998.

[YHM94] Wei-Yong Yan, Uwe Helmke, and John B. Moore. Global analysis of Oja's flow for neural networks. *IEEE Transactions on Neural Networks*, 5:674–683, 1994.

[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.

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.

$$\mathbf{w}_t = \frac{\left(I + \eta_t \mathbf{x}_t \mathbf{x}_t^\top\right) \mathbf{w}_{t-1}}{\| \left(I + \eta_t \mathbf{x}_t \mathbf{x}_t^\top\right) \mathbf{w}_{t-1} \|_2} \tag{24}$$

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

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

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

$$= \left(1 + 2\eta_t y_t^2 + O(\eta_t^2)\right)^{-1/2} = 1 - \eta_t y_t^2 + O(\eta_t^2) \tag{25}$$

Replace the denominator of Equation 24 with Equation 25 and truncate the $O(\eta_t^2)$ term, one gets exactly Equation 2.

B Details of the Linearizations in Continuous Oja's Rule

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

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

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 30 (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 30. In the following, everything holds almost surely so we would not mention this condition every time. First, consider

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

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. □

Now, we restate and prove Theorem 5 as follows.

Theorem 5 (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 5. Observe that for any $t \geq 0$ such that $\mathbf{w}_{t,1}^2 \leq 1 - \epsilon$, by the second item of Lemma 30, 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 30), 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 30. Thus, we conclude that for any $t = \Omega\left(\frac{\log(1/\mathbf{w}_{0,1}^2)}{\epsilon(\lambda_1 - \lambda_2)}\right)$, $\mathbf{w}_{t,1}^2 > 1 - \epsilon$. □

Now, we restate and prove Theorem 6 as follows.

Theorem 6 (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 6. Observe that for any $t \geq 0$, by the second item of Lemma 30, 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 30) and at most 1, we have

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

By solvign 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 $\mathbf{w}_{t,1}^2 > 1 - \epsilon$. □

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

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

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 27

Proof of Lemma 27. The proof is basically direct verification using the definition of ξ, τ_t, ψ and Lemma 19. Let's first describe $\nabla f_{t,j}(\mathbf{w}_{s-1})$ and $\nabla^2 f_{t,j}(\mathbf{w}_{s-1})$ and give their corresponding bounds. We have

$$(\nabla f_{t,j}(\mathbf{w}_{s-1}))_1 = \frac{-f_{t,j}(\mathbf{w}_{s-1})}{\mathbf{w}_{s-1,1}},\ \forall 1 < i \leq j, (\nabla f_{t,j}(\mathbf{w}_{s-1}))_i = \frac{\mathbf{x}_{t,i}}{\mathbf{w}_{s-1,1}}$$

and all other coordinates are zero. In particular, conditioning on $\tau_t, \psi \geq s$, we have

$$\|\nabla f_{t,j}(\mathbf{w}_{s-1})\|_2 = O(\frac{\Lambda}{\mathbf{w}_{s-1,1}^2}) = O(\sqrt{\Lambda'}\Lambda) \tag{26}$$

For $\nabla^2 f_{t,j}(\mathbf{w}_{s-1})$, we have

$$(\nabla^2 f_{t,j}(\mathbf{w}_{s-1}))_{1,1} = \frac{\sum_{i=2}^j \mathbf{x}_{t,i}\overline{\mathbf{w}}_{s-1,i}}{\overline{\mathbf{w}}_{s-1,1}^3}$$

$$\forall 1 < i \le j,\ (\nabla^2 f_{t,j}(\mathbf{w}_{s-1}))_{1,i} = (\nabla^2 f_{t,j}(\mathbf{w}_{s-1}))_{i,1} = -\frac{\mathbf{x}_{t,i}}{\overline{\mathbf{w}}_{s-1,1}^2}$$

and all other coordinates are zero. In particular, we can rewrite it as linear combination of three rank one matrices

$$\nabla^2 f_{t,j}(\mathbf{w}_{s-1}) = \alpha_1 \mathbf{x}_t^{(j,1)} \mathbf{x}_t^{(j,1)\top} + \alpha_2 \mathbf{x}_t^{(j,0)} \mathbf{x}_t^{(j,0)\top} + \alpha_3 e_1 e_1^\top$$

where

$$\alpha_1 = -\frac{1}{\overline{\mathbf{w}}_{s-1}^2},\ \alpha_2 = \frac{1}{\overline{\mathbf{w}}_{s-1}^2},\ \alpha_3 = \frac{\sum_{i=2}^{j} \mathbf{x}_{t,i} \overline{\mathbf{w}}_{s-1,i}}{\overline{\mathbf{w}}_{s-1}^3} + \frac{1}{\overline{\mathbf{w}}_{s-1}^2}, \text{ and}$$

e_1 is the basis vector of first coordinate and $\mathbf{x}_{t,i}^{(j,a)} = \mathbf{x}_{t,i}$ if $1 < i \le j$, $\mathbf{x}_{t,1}^{(j,a)} = a$ and it is zero at all other coordinates. Now we would like to bound the coefficient. Notice that since $\eta = O(\frac{1}{\Lambda})$,

$$\overline{\mathbf{w}}_{s-1,i} = \mathbf{w}_{s-1,i} + c\eta \mathbf{z}_{s,i} = \mathbf{w}_{s-1,i} + O(\mathbf{w}_{s-1,1}\mathbf{x}_{s,i} + \eta \mathbf{w}_{s-1,i})$$

In particular, $\overline{\mathbf{w}}_{s-1,i} = O(\mathbf{w}_{s-1,i} + \mathbf{w}_{s-1,1}\mathbf{x}_{s,i})$. Now we can bound the coefficient $|\alpha_1| = O(\frac{1}{\mathbf{w}_{s-1,1}^2})$, $|\alpha_2| = O(\frac{1}{\mathbf{w}_{s-1,1}^2})$ and $|\alpha_3| = O\left(\frac{\Lambda}{\mathbf{w}_{s-1,1}^2}\right)$. In particular, given any vector v, we have

$$\begin{aligned} |v^\top \nabla^2 f_{t,j}(\mathbf{w}_{s-1}) v| &= \left| \alpha_1 v^\top \mathbf{x}_t^{(j,1)} \mathbf{x}_t^{(j,1)\top} v + \alpha_2 v^\top \mathbf{x}_t^{(j,0)} \mathbf{x}_t^{(j,0)\top} v + \alpha_3 v^\top e_1 e_1^\top v \right| \\ &= \left| \alpha_1 \mathbf{x}_t^{(j,1)\top} v v^\top \mathbf{x}_t^{(j,1)} + \alpha_2 \mathbf{x}_t^{(j,0)\top} v v^\top \mathbf{x}_t^{(j,0)} + \alpha_3 e_1^\top v v^\top e_1 \right| \end{aligned}$$

By combining the bound $\alpha_i = O\left(\frac{\Lambda}{\mathbf{w}_{s-1,1}^2}\right)$, $\|vv^\top\|_2 = \|v\|_2^2$ and $\|e_1\|_2$, $\|\mathbf{x}_t^{(j,0)}\|_2$, $\|\mathbf{x}_t^{(j,1)}\|_2 \le 2$, we have

$$\le O\left(\frac{\Lambda}{\mathbf{w}_{s-1,1}^2} \|v\|_2^2\right) \tag{27}$$

Now we are ready to analyze the bounds on $\mathbf{A}_{s,j}^{(t)}$. For notational convenience, denote $\mathbf{z}_s - \mathbb{E}[\mathbf{z}_s \mid \mathcal{F}_{s-1}]$ as $\overline{\mathbf{z}}_s$ and separate $\mathbf{A}_{s,j}^{(t)}$ into two terms where $\mathbf{A}_{s,j}^{(t,1)} = \eta \nabla f_{t,j}(\mathbf{w}_{s-1})^\top \overline{\mathbf{z}}_s$ and $\mathbf{A}_{s,j}^{(t,2)} = \eta^2 \mathbf{z}_s^\top \nabla f_{t,j}^2(\overline{\mathbf{w}}_{s-1}) \cdot \mathbf{z}_s$. By Cauchy-Schwarz and Equation 26, We have

$$|\mathbf{A}_{s,j}^{(t,1)}| \le \eta \|\nabla f_{t,j}(\mathbf{w}_{s-1})\|_2 \|\overline{\mathbf{z}}_s\|_2 = O\left(\eta \cdot \frac{\Lambda}{\mathbf{w}_{s-1}} \cdot y_s\right) = O(\eta \Lambda^2)$$

We also have

$$|\mathbf{A}_{s,j}^{(t,2)}| = |\eta^2 \mathbf{z}_s^\top \nabla^2 f_{t,j}(\mathbf{w}_{s-1}) \mathbf{z}_s|$$

By Equation 27, we have

$$= O\left(\eta^2 \frac{\Lambda}{\mathbf{w}_{s-1,1}^2} \|\mathbf{z}_s\|_2^2\right)$$

Because $\|\mathbf{z}_s\|_2^2 = O(y_s^2)$, we have

$$= O\left(\eta^2 \frac{\Lambda}{\mathbf{w}_{s-1,1}^2} y_s^2\right)$$
$$= O(\eta^2 \Lambda^3) = O(\eta \Lambda^2)$$

This gives us $|\mathbf{1}_{\psi,\tau_t \geq s,\xi > s} \mathbf{A}_{s,j}^{(t)}| = O(\eta\Lambda^2)$. For conditional expectation, we have

$$\left|\mathbb{E}[\mathbf{1}_{\psi,\tau_t \geq s,\xi > s} \mathbf{A}_{s,j}^{(t,1)} \mid \mathcal{F}_{s-1}^{(t)}, \mathcal{C}_{init}^{p,\delta}]\right| = \left|\mathbb{E}[\mathbf{1}_{\psi,\tau_t \geq s,\xi > s} \eta \nabla f_{t,j}(\mathbf{w}_{s-1})^\top \overline{\mathbf{z}}_s | \mathcal{F}_{s-1}^{(t)}, \mathcal{C}_{init}^{p,\delta}]\right|$$

Notice that we have $\mathbb{E}[\mathbf{z}_s \mid \mathcal{F}_{s-1}^{(t)}, \mathcal{C}_{init}^{p,\delta}] = \mathbb{E}[\mathbf{x}_s \mathbf{x}_s^\top \mathbf{w}_{s-1} - \mathbf{w}_{s-1}^\top \mathbf{x}_s \mathbf{x}_s^\top \mathbf{w}_{s-1} \mathbf{w}_{s-1} \mid \mathcal{F}_{s-1}^{(t)}, \mathcal{C}_{init}^{p,\delta}]$. By Lemma 19 applying on $\mathbf{x}_s \mathbf{x}_s^\top$ and Cauchy-Schwarz, we have

$$\leq O\left(\eta \|\nabla f_{t,j}(\mathbf{w}_{s-1})\|_2 \frac{1}{nT} \|\mathbf{x}_s\|_2) + \eta \|\nabla f_{t,j}(\mathbf{w}_{s-1})\|_2 \|\mathbf{w}_{s-1}\|_2^3 \frac{1}{nT}\right)$$

By Equation 26 and $T = \Omega(\frac{1}{\eta\lambda_1})$, we have

$$\leq O\left(\frac{\eta^2 \lambda_1 \Lambda^2}{\sqrt{n}}\right) = O(\eta^2 \lambda_1 \Lambda^3)$$

For $\mathbf{A}_{s,j}^{(t,2)}$, we have

$$\left|\mathbb{E}[\mathbf{1}_{\psi,\tau_t \geq s,\xi > s} \mathbf{A}_{s}^{(t,2)} \mid \mathcal{F}_{s-1}^{(t)}, \mathcal{C}_{init}^{p,\delta}]\right| = \left|\mathbb{E}[\mathbf{1}_{\psi,\tau_t \geq s,\xi > s} \eta^2 \mathbf{z}_s^\top \nabla^2 f_{t,j}(\mathbf{w}_{s-1})^\top \mathbf{z}_s | \mathcal{F}_{s-1}^{(t)}, \mathcal{C}_{init}^{p,\delta}]\right|$$

Notice we have $\|\mathbb{E}[\mathbf{1}_{\psi,\tau_t \geq s,\xi > s} \mathbf{z}_s \mathbf{z}_s^\top \mid \mathcal{F}_{s-1}^{(t)}]\|_2 = O(y_s^2 \lambda_1)$ by Lemma 19. Again by Equation 27, we have

$$\leq O\left(\eta^2 y_s^2 \lambda_1 \frac{\Lambda}{\mathbf{w}_{s-1,1}^2}\right) = O(\eta^2 \lambda_1 \Lambda^3)$$

So we have

$$\left|\mathbb{E}[\mathbf{1}_{\psi,\tau_t \geq s,\xi > s} \mathbf{A}_{s}^{(t)} \mid \mathcal{F}_{s-1}^{(t)}, \mathcal{C}_{init}^{p,\delta}]\right| = O(\eta^2 \lambda_1 \Lambda^3)$$

For the last moment bound, fix $2 \leq j, j' \leq n$. Expanding the definition, we get

$$\mathbb{E}[\mathbf{1}_{\psi,\tau_t \geq s,\xi > s} \mathbf{A}_{s,j}^{(t,1)} \mathbf{A}_{s,j'}^{(t,1)} + \mathbf{A}_{s,j}^{(t,1)} \mathbf{A}_{s,j'}^{(t,2)} + \mathbf{A}_{s,j}^{(t,2)} \mathbf{A}_{s,j'}^{(t,1)} + \mathbf{A}_{s,j}^{(t,2)} \mathbf{A}_{s,j'}^{(t,2)} \mid \mathcal{F}_{s-1}, \mathcal{C}_{init}^{p,\delta}]$$

For the first term, we have

$$|\mathbb{E}[\mathbf{1}_{\psi,\tau_t \geq s,\xi > s} \mathbf{A}_{s,j}^{(t,1)} \mathbf{A}_{s,j'}^{(t,1)} \mid \mathcal{F}_{s-1}, \mathcal{C}_{init}^{p,\delta}]| = |\mathbb{E}[\mathbf{1}_{\psi,\tau_t \geq s,\xi > s} \eta^2 \nabla f_{t,j}(\mathbf{w}_{s-1})^\top \overline{\mathbf{z}}_s \overline{\mathbf{z}}_s^\top \nabla f_{t,j'}(\mathbf{w}_{s-1}) \mid \mathcal{F}_{s-1}, \mathcal{C}_{init}^{p,\delta}]|$$

Notice we have $\|\mathbb{E}[\mathbf{1}_{\psi,\tau_t \geq s,\xi > s} \overline{\mathbf{z}}_s \overline{\mathbf{z}}_s^\top \mid \mathcal{F}_{s-1}^{(t)}, \mathcal{C}_{init}^{p,\delta}]\|_2 = O(y_s^2 \lambda_1)$ by Lemma 19. We have

$$\leq O(\eta^2 \|\nabla f_{t,j}(\mathbf{w}_{s-1})\|_2 y_s^2 \lambda_1 \|\nabla f_{t,j'}(\mathbf{w}_{s-1})\|_2)$$

Since we know $\|\nabla f_{t,j}(\mathbf{w}_{s-1})\|_2 = O\left(\frac{\Lambda}{\mathbf{w}_{s-1,1}^2}\right)$, we have

$$= O(\eta^2 \lambda_1 \Lambda^4)$$

For the second and third term, since they are symmetric, we will only deal with the second term. We have

$$|\mathbb{E}[\mathbf{1}_{\psi,\tau_t\geq s,\xi>s}\mathbf{A}_{s,j}^{(t,1)}\mathbf{A}_{s,j'}^{(t,2)} \mid \mathcal{F}_{s-1},\mathcal{C}_{init}^{p,\delta}]| = |\mathbb{E}[\mathbf{1}_{\psi,\tau_t\geq s,\xi>s}\eta^3\nabla f_{t,j}(\mathbf{w}_{s-1})^\top\overline{\mathbf{z}}_s\mathbf{z}_s^\top\nabla^2 f_{t,j'}(\mathbf{w}_{s-1})\mathbf{z}_s^\top \mid \mathcal{F}_{s-1},\mathcal{C}_{init}^{p,\delta}]|$$

By taking the maximum of the $\mathbf{A}_{s,j}^{(t,1)}$ and combining with Equation 27, we have

$$\begin{aligned}&\leq O(\eta\Lambda^2\cdot\eta^2\lambda_1\Lambda^3)\\&= O(\eta^3\lambda_1\Lambda^5) = O(\eta^2\lambda_1\Lambda^4)\end{aligned}$$

For the last term, we can deal with it completely analogously. In particular we have

$$|\mathbb{E}[\mathbf{1}_{\psi,\tau_t\geq s,\xi>s}\mathbf{A}_{s,j}^{(t,2)}\mathbf{A}_{s,j'}^{(t,2)} \mid \mathcal{F}_{s-1},\mathcal{C}_{init}^{p,\delta}]| \leq O(\eta\Lambda^2\cdot\eta^2\lambda_1\Lambda^3\cdot) = O(\eta^2\lambda_1\Lambda^4)$$

Combining all the terms, we get

$$|\mathbb{E}[\mathbf{1}_{\psi,\tau_t\geq s,\xi>s}\mathbf{A}_{s,j}^{(t)}\mathbf{A}_{s,j'}^{(t)} \mid \mathcal{F}_{s-1},\mathcal{C}_{init}^{p,\delta}]| = O(\eta^2\lambda_1\Lambda^4)$$

□