Mathematical Analysis of Static and Plastic Biological Neural Circuits

by

Mien Brabeeba Wang

B.A. in Mathematics, Harvard University (2018)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2020

Author .....
Department of Electrical Engineering and Computer Science
May 15, 2020

Certified by .....
Nancy A. Lynch
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by .....
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science

Mathematical Analysis of Static and Plastic Biological Neural Circuits

by

Mien Brabeeba Wang

Submitted to the Department of Electrical Engineering and Computer Science
on May 15, 2020, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering and Computer Science

Abstract

In this thesis, I explain possible mathematical principles behind brain computations during the processing of temporal information and fast sensory adaptation using static and plastic neural circuits respectively. For the static part of the thesis, I investigate the possible computational principles behind how the brain can process temporal information over a long time range using neurons with transient activities. Specifically, I design static memoryless neural circuits that are capable of processing temporal sequences in either rate coding or temporal coding and prove that the networks are optimal in both the number of the neurons and the convergence time. For the plastic part of the thesis, I show how a sensory system can potentially adapt quickly under Barlow's efficient coding principle despite having high dimensional sensory inputs. Specifically, I use Oja's rule as an example of sensory adaptation under the efficient coding principle and give the first convergence rate analysis of Oja's rule in solving streaming principal component analysis (PCA). In particular, the convergence rate I obtain matches the information-theoretic lower bound up to logarithmic factors and outperforms the state-of-the-art analysis for other streaming PCA algorithms in the literature. I further demonstrate the capacity of Oja's rule for continual learning in a living system. Specifically, I prove that Oja's rule can continuously adapt to changing environments without sacrificing too much efficiency and remain functional throughout the process.

Thesis Supervisor: Nancy A. Lynch
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

I would like to thank my family and my friends for supporting me emotionally during the research process. I would like to thank Nancy Lynch for supervision and discussion on this thesis. I would like to thank Nancy Lynch and Cameron Musco for helpful technical discussion in the static network part of this thesis. I would like to thank my coauthor and my best friend Chi-Ning Chou for countless sleepless nights with me working on the plastic network part of the thesis. I would like to thank Kai-Min Chung for helpful discussions on Oja's rule and algorithms in neuroscience. I want to thank Nancy Lynch again for organizing a brain algorithm reading group at MIT and all the participants for the inspiring conversation.

Contents

Chapter 1

Introduction

The purpose of this thesis is to mathematically understand some principles of brain computation. Brains comprise networks of neurons that individually have relatively simple dynamics. However, the brain can execute complicated tasks, such as playing violin, or learning concepts from the environment, such as language acquisition. Although scientists have obtained many experimental findings and theoretical insights about the brain, most underlying principles behind brain computation have remained elusive. There are two aspects of brain computation: the *static network dynamic*, which describes how the electrical activities of neurons evolve under fixed connections and the *activity-dependent synaptic plasticity*, which describes how the strength of the synapses varies based on the activities of the neurons. For static networks, understanding the underlying computational principles is already difficult because of the possibility of chaotic behaviors in recurrent connections and the highly nonlinear nature of spiking dynamics. Synaptic plasticity further adds an layer of complication. Therefore, a theoretical understanding of brain computation is a challenging task.

There have been many attempts to model brain computationally. At a single-neuron level, theoretical neuroscientists were able to model the dynamics of a single neuron to high accuracy with the Hodgkin-Huxley model [35]. At a static circuit level, to make the analysis tractable, neuroscientists approximated detailed dynamics of neurons with simplified models such as the nonlinear integrate-and-fire model [24] and the spiking response model [42]. On the other hand, since the experimental

finding of synaptic plasticity [15, 41, 22], theoretical neuroscientists have considered numerous learning rules that govern the dynamics of the synapses to model synaptic plasticity such as Oja's rule [59], BCM rule [14], covariance rule [71], spike-timing-dependent plasticity [12], etc. It remains a big mystery how these relatively simple neuronal dynamics and plasticity rules can generate complicated behavioral outcomes such as language acquisition.

In this thesis, we take a mathematical approach to model two particular biological phenomena: processing of temporal information and sensory adaptation, using static and plastic neural circuits respectively.

1.1 Processing of temporal information

One of the most important questions in neuroscience is how humans integrate information over time. Sensory inputs such as visual and auditory stimuli are inherently temporal; yet brains can integrate the temporal information into a single concept, such as recognizing a moving object in a visual scene or forming an entity in a sentence. In the above examples, the temporal information spans over a time scale of 1-10 seconds. However, individual neurons only have transient activities with the time scale of 10-100ms. It is not clear how neurons with transient components can process temporal information over a long time range. In the static network part of this thesis, we are going to present a static network to process temporal information and translate it into spatial information with transient components.

There are two kinds of neuronal codings: *rate coding* and *temporal coding*. Rate coding is a neural coding scheme assuming most of the information is coded in the firing rate of the neurons. It is most commonly seen in muscle when the higher firing rates of motor neurons correspond to higher intensity in muscle contraction [2]. On the other hand, rate coding cannot be the only neural coding brains employ. A fly is known to react to new stimuli and change its direction of flight within 30-40 ms. For a neuron that spikes at around $50Hz$, which is much higher than the average spiking rate, there is only time to produce 1-2 spikes within this window. There is simply not enough time for neurons to decode rate coding accurately [13]. Therefore,

neuroscientists proposed the idea of temporal coding, assuming the information is coded in the temporal firing patterns. One of the popular temporal codings is the first-to-spike coding, in which the information is encoded in the duration between the stimulus onset and the first spike. By plotting the timing of the first spike in retina ganglion cells, one can recover an approximately accurate image on a retina [27].

We propose two toy problems to model how brains extract temporal information from different coding with transient components. *"First consecutive spikes counting" (FCSC)* counts the first consecutive interval of spikes, which is equivalent to counting the distance between the first two spikes, a prevalent temporal coding scheme in the sensory cortex. *"Total spikes counting" (TSC)* counts the number of the spikes over an arbitrary interval, which is an example of rate coding. To model the transient components of neurons, we consider a memoryless synchronous spiking neuron model where the firing of a neuron only depends on the spike events one time step ago.

In this thesis, we design two networks that solve the above two problems by translating temporal information into spatial information in time 1 with $O(\log T)$ neurons. We further show that any network with less than T neurons cannot solve the problems in time 0. It should be noted that Hitron and Parter also considered the TSC problem [34] with the time bound $O(\log T)$. In this context, we improve the time bound on the TSC problem from $O(\log T)$ to 1 by carefully updating all digits in binary representation at once instead of sequentially. We would like to remark that although our problems are biologically inspired, the optimal solutions we propose are not biologically plausible. The networks are not noise-tolerant, whereas the neuronal dynamics are highly noisy and it is hard to conceive that the brain uses binary representation as a neuronal representation. However, the analysis serves as a proof of concept that the brain can process temporal information over a long time range using transient components.

1.2 Oja's rule and sensory adaptation

One of the most influential theoretical ideas in neuroscience is Barlow's efficient coding principle for sensory systems [11]. Barlow hypothesized that the main goal of a

sensory system is to reduce the redundancy in the sensory input and maximize the information transmitted to downstream brain areas. One of its key predictions is that the sensory neurons in the brain adapt to natural stimuli. Indeed, neuroscientists have shown in numerous sensory systems that by maximizing the mutual information transmitted on natural stimuli, one can recover the response filters in the respective sensory system. In the visual system, the structures of both the center-surround receptive field of the retina ganglion cells [6, 7, 29] and Gabor filters of V1 simple cells can be mathematically derived from the efficient coding principle [62]. In the auditory system, the temporal cochlear filters of inner ears can also be derived from optimizing mutual information on natural sounds [48]. However, most works on efficient coding of a sensory system have focused on optimizing the statistics of one natural environment. In reality, the environmental statistics can change drastically and the sensory system needs to continuously adapt to the changing environment in a matter of seconds while having high dimensional sensory inputs. For example, although the retina processes visual inputs from 100 million photoreceptors to 1 million retina ganglion cells, it can change its receptive field to adapt to environments with different illumination [74], contrast [74, 9, 75], spatial frequency [75, 37], orientation and temporal correlation [37] in the time scale of seconds. Therefore, it is important to have a theoretical understanding of how the efficient coding principle can adapt to changing environments in a biologically realistic timescale with a biologically plausible synaptic learning rule. In this thesis, we give the first theoretical demonstration of sensory adaptation under the efficient coding principle in biologically realistic timescale through studying the convergence rate and behaviors of *Oja's rule* [59].

It is known that Oja's rule maximizes the mutual information under Gaussian inputs and linear networks by adapting to the direction that maximizes the variance of the presynaptic inputs through solving Principal Component Analysis (PCA) [50]. Therefore, studying its convergence rate and behaviors can shed light on fast sensory adaptation under the efficient coding principle. Since the dimensionality of the sensory inputs is usually large, for Oja's rule to behave in a biologically realistic time scale, the convergence rate needs to have no dependency or only log dependency on

the input dimension. In addition to its relation to the efficient coding principle, as a biologically plausible synaptic modification rule, Oja's rule serves as a plasticity candidate to investigate sensory adaptation. Oja's rule is one of the earliest local learning rules that incorporate both *Hebbian* and *homeostatic plasticity* [59], two major activity-dependent synaptic modification mechanisms [1]. Both mechanisms work together to form memory and drive learning behaviors in the brain. Hebbian plasticity is a synapse-specific correlation-based plasticity mechanism that strengthens the connection when the input has a high correlation with the weights while weakening the connection when the input has a poor correlation [41, 22, 14]. However, this type of mechanism alone can often make networks unstable since the highly correlated input will keep strengthening synapses unboundedly [1]. Homeostatic plasticity, in contrast, stabilizes the network by keeping the activities of the neurons relatively constant through calcium sensors [80]. Synaptic scaling is a specific kind of homeostatic plasticity where the strength of the incoming synapses is normalized while still encoding the information from Hebbian learning in their relative strength after normalization [79]. It is thus an important problem in computational neuroscience to understand the interplay between Hebbian and homeostatic plasticity [78]. Oja's rule is one example of this. Concretely, Oja's rule can be expressed as the following

$$w_t = w_{t-1} + \eta_t(x_t y_t - y_t^2 w_{t-1})$$

where w_t is the strength of the synapse at time t, x_t, y_t are the firing rates of presynaptic, and postsynaptic neurons respectively, and η_t is the learning rate. One can see that $x_t y_t$ term corresponds to the Hebbian plasticity while $y_t^2 w_{t-1}$ term corresponds to the homeostatic plasticity. One can then show the synaptic scaling property where $\|w_t\| \approx 1$ for all t.

Despite being a subject of extensive theoretical [59, 61, 70, 33, 60, 68, 21, 88, 87, 23, 5] and experimental [16, 43, 31, 40, 18, 73, 77, 51, 5] studies aimed at understanding its performance, the theoretical understanding of the Oja's rule remains incomplete. The state-of-the-art theoretical analysis only provides a guarantee on convergence in the limit [23] through Kushner-Clark methods [44]. However, to the

best of our knowledge, there is no prior work showing the convergence time of Oja's rule. Specifically, if the convergence time of Oja's rule does not depend on the input dimension or depend on only logarithmic factors of the input dimension, Oja's rule can serve as an example of sensory adaptation under the efficient coding principle in a biologically realistic time scale.

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

Theorem 1.2.1 (informal)**.** *Biological Oja's rule efficiently solves streaming PCA with (nearly) optimal convergence rate. Specifically, the convergence rate we obtain matches the information-theoretic 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 dependency on logarithmic factors of the dimension when the initial vector is random. Therefore, biological Oja's rule solves streaming PCA on a biologically realistic time scale.

Also, we show for-all-time convergence with a slowly diminishing learning rate. Most convergence results in the literature show that

$$\Pr(\text{error at time T} > \epsilon) < \delta.$$

However, this is not enough in a biological system. The sensory system cannot afford to only be functional at time T. It needs to be functional constantly. In contrast, the convergence result we can show is

$$\Pr(\exists t \geq T, \text{ error at time t} > \epsilon) < \delta,$$

which guarantees the convergence at all time. Furthermore, in order to achieve this, our learning rate η_t only needs to be scaled as $\eta_t = O(\frac{1}{\log t})$, in particular $\sum_t \eta_t^2 = \infty$. In contrast, Kushner-Clark theorem requires $\sum_t \eta_t^2 < \infty$ where the learning rate is commonly set as $\eta_t = O(\frac{1}{t})$. 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.

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

1.3 Summary

In this thesis, we present mathematical analysis on static and plastic biological neural networks that show principles of temporal processing and sensory adaptation respectively. In addition to biological relevance, these two problems also show interesting mathematical elements. For example, as a result of analyzing Oja's rule, we come up with a powerful framework to analyze general stochastic dynamics. However, we would like to comment that both networks in this thesis do not capture certain aspects of biology. The static network we present is not noise-tolerant, while Oja's rule does not demonstrate BCM-rule-like behaviors, which are experimentally observed. Nonetheless, they both capture some important principles of brain computation. To move forward in theoretical neuroscience, the theory needs to combine with the biology. With too many biological details, the underlying principle becomes intractable to understand while with too few biological details, the theory might not capture the essence of the computation. It is important to have biological intuition to identify the important details that are relevant to the computation and it is also important to have the mathematical theory to facilitate the design of biological experiments. In the future, we would like to create theories that are more closely integrated with biology to understand the principles of brain computation further.

Chapter 2

Static Neural Circuit: Processing of Temporal Information

2.1 Introduction

One of the most important questions in neuroscience is how humans integrate information over time. Sensory inputs such as visual and auditory stimuli are inherently temporal; yet brains can integrate the temporal information into a single concept, such as recognizing a moving object in a visual scene or forming an entity in a sentence. In the above examples, the temporal information spans over a time scale of 1-10 seconds. However, individual neurons only have transient activities with the time scale of 10-100ms. It is not clear how neurons with transient components can process temporal information over a long time range. In this chapter, we are going to present a static network to process temporal information and translate it into spatial information with transient components.

There are two kinds of neuronal codings: *rate coding* and *temporal coding*. Rate coding is a neural coding scheme assuming most of the information is coded in the firing rate of the neurons. It is most commonly seen in muscle when the higher firing rates of motor neurons correspond to higher intensity in muscle contraction [2]. On the other hand, rate coding cannot be the only neural coding brains employ. A fly is known to react to new stimuli and change its direction of flight within 30-40 ms.

For a neuron that spikes at around $50Hz$, which is much higher than the average spiking rate, there is only time to produce 1-2 spikes within this window. There is simply not enough time for neurons to decode rate coding accurately [13]. Therefore, neuroscientists proposed the idea of temporal coding, assuming the information is coded in the temporal firing patterns. One of the popular temporal codings is the first-to-spike coding, in which the information is encoded in the duration between the stimulus onset and the first spike. By plotting the timing of the first spike in retina ganglion cells, one can recover an approximately accurate image on a retina [27].

We propose two toy problems to model how brains extract temporal information from different coding with transient components. *"First consecutive spikes counting" (FCSC)* counts the first consecutive interval of spikes, which is equivalent to counting the distance between the first two spikes, a prevalent temporal coding scheme in the sensory cortex. *"Total spikes counting" (TSC)* counts the number of the spikes over an arbitrary interval, which is an example of rate coding. To model the transient components of neurons, we consider a memoryless synchronous spiking neuron model where the firing of a neuron only depends on the spike events one time step ago.

In this chapter, we design two networks that solve the above two problems by translating temporal information into spatial information in time 1 with $O(\log T)$ neurons. We further show that any network with less than T neurons cannot solve the problems in time 0. It should be noted that Hitron and Parter also considered the TSC problem [34] with the time bound $O(\log T)$. In this context, we improve the time bound on the TSC problem from $O(\log T)$ to 1 by carefully updating all digits in binary representation at once instead of sequentially. We would like to remark that although our problems are biologically inspired, the optimal solutions we propose are not biologically plausible. The networks are not noise-tolerant, whereas the neuronal dynamics are highly noisy and it is hard to conceive that the brain uses binary representation as a neuronal representation. However, the analysis serves as a proof of concept that the brain can process temporal information over a long time range using transient components.

The organization of the rest of the section is as follows. In Section 2.1.1, we

formally define the spiking neuron model we are working in. In Section 2.1.2, we define the two biologically-inspired problems “First Consecutive Spikes Counting" and “Total Spikes Counting" which correspond to temporal coding and rate coding respectively. In Section 2.1.3, we provide our main results, solving the two problems optimally in both time and the number of the neurons and showing that we cannot do better.

2.1.1 Model

In this work, to model the transient aspect of the neurons, we consider a network of memoryless spiking neurons with deterministic synchronous firing at discrete times. Formally, a neuron z consists of the following data with $t \geq 1$

$$z^{(t)} = \Theta(\sum_{y \in P_z} w_{yz} y^{(t-1)} - b_z)$$

where $z^{(t)}$ is the indicator function of neuron z firing at time t. b_z is the threshold (bias) of neuron z. P_z is the set of presynaptic neurons of z, w_{yz} is the strength of connection from neuron y to neuron z and Θ is a nonlinear function. Here we take Θ as the Heaviside function given by $\Theta(x) = 1$ if $x > 0$ and 0 otherwise. At $t = 0$, we let $z^{(0)} = 0$ if z is not one of the input neurons.

For the rest of the chapter, we fix an input neuron x and m output neurons $\{y_i\}_{0 \leq i < m}$ in a network.

2.1.2 Problem Statement

First Consecutive Spikes Counting(T) (FCSC(T))

Given an input neuron x and the max input length T, we consider any input firing sequence such that for all $t \geq T$, $x^{(t)} = 0$. Define L_x in terms of this firing sequence as follows: if $x^{(t)} = 1$ for some t, then there must exist integers $\hat{t}, L$ such that for all $t, t < \hat{t}$ we have $x^{(t)} = 0$, for all $i, 0 \leq i < L$ we have $x^{(\hat{t}+i)} = 1$, and $x^{(\hat{t}+L)} = 0$. Define $L_x = L$. (i.e., L is the length of the first consecutive spikes interval in the sequence.) Otherwise, that is if for all $t \geq 0$, $x^{(t)} = 0$, then define $L_x = 0$.

Then we say a network of neurons solves FCSC(T) in time t' with m' neurons if

there exists an injective function $F : \{0, \cdots, T\} \to \{0,1\}^m$ such that for all x and for all $t, t \geq T + t'$ we have $y^{(t)} = F(L_x)$ and the network has m' total neurons.

Intuitively, FCSC serves as a toy model for encoding distance between spikes, a prevalent spike coding in the sensory cortex. For mathematical convenience, we model the problem as counting the distance between non-spikes which is mathematically equivalent as counting the distance between spikes in our model.

Total Spikes Counting(T) (TSC(T))

Given an input neuron x and the max input length T, we consider any input firing sequence such that for all $t \geq T$, $x^{(t)} = 0$. Define $L_x = |\{t : x^{(t)} = 1, 0 \leq t < T\}|$ as the total number of spikes in the sequence. Then we say a network of neurons solves TSC(T) in time t' with m' neurons if there exists an injective function $F : \{0, \cdots, n\} \to \{0,1\}^m$ such that for all x and for all $t, t \geq T + t'$ we have $y^{(t)} = F(L_x)$ and the network has m' total neurons.

Intuitively, TSC serves as a toy model for rate coding implemented by spiking neural networks because the network can extract the rate information by counting the number of spikes over arbitrary intervals.

Notice that in both problems above, a network solves a task in time t' if, for all $t \geq T + t'$ and for all inputs with max length T, the network outputs the solution of the task at time t. The definition is equivalent to Maass's time complexity for spiking neurons [58]. This definition of the time bound makes natural sense since given a max input length of T, it is unreasonable to count the time before the end of the input.

2.1.3 Main Theorems

Our contributions in this work are to design networks that solve these two problems respectively with matching lower bounds in numbers of neurons.

Theorem 2.1.1. *There exists a network that solves FCSC(T) problem with* $O(\log T)$ *neurons in time* 1.

Theorem 2.1.2. *There exists a network that solves TSC(T) problem with* $O(\log T)$ *neurons in time* 1.

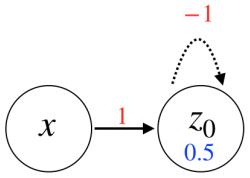


Figure 2-1: mod 2 Base Network

It is easy to see that we also have the corresponding information-theoretical lower bound on the number of neurons all being $\Omega(\log T)$ by the requirements of the tasks.

In terms of time bound, we also show that our networks are optimal for FCSC and TSC problem in the following sense:

Theorem 2.1.3. *There does not exists a network with less than T neurons that solves FCSC(t) problem in time 0 for all $0 \leq t \leq T$.*

Theorem 2.1.4. *There does not exists a network with less than T neurons that solves TSC(t) problem in time 0 for all $0 \leq t \leq T$.*

2.2 First Consecutive Spikes Counting

We present the constructions in two stages. At the first stage, we count consecutive spikes in binary transiently. At the second stage, we transform the transient firing into persistent firing. By composing the two stages, we get our desired network.

2.2.1 First Stage: Counter Network

The network contains neurons $z_0, \cdots, z_n$, $in_1, \cdots, in_n$ and we build the network inductively. To construct mod 2 Base Network which counts mod 2, we have

$$w_{xz_0} = 1, w_{z_0z_0} = -1, b_{z_0} = 0.5.$$

By noticing that for $t \geq 1$, $z_0^{(t)} = 1$ if and only if $x^{(t-1)} = 1$ and $z_0^{(t-1)} = 0$, we have the following lemma

Lemma 2.2.1. *For the mod 2 base network, given $t \geq 0$ if for all $t' such that 0 \leq t' \leq t$ we have $x^{(t')} = 1$, then at time t, $z_0^{(t)} = t \bmod 2$.*

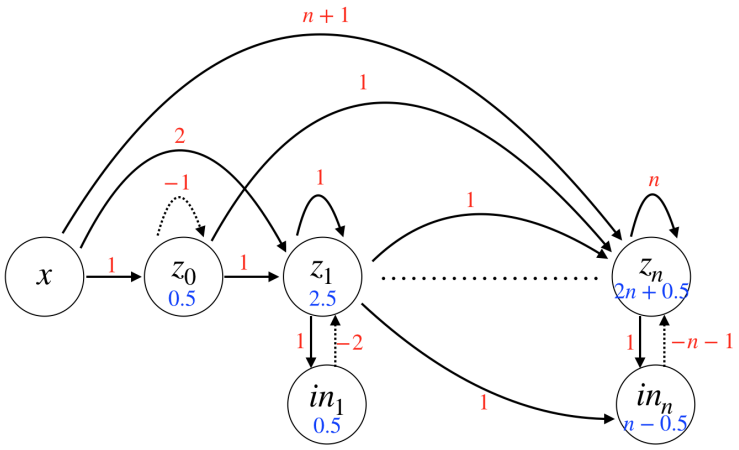


Figure 2-2: First Stage

Now we iteratively build the network where $1 \leq i \leq n$ on top of the mod 2 base network with the following rule:

$$w_{xz_i} = i{+}1,\ w_{z_jz_i} = 1,\ \forall j, 0 \leq j < i, w_{z_kin_i} = 1, \forall k, 0 < k \leq i, w_{in_iz_i} = -i{-}1, w_{z_iz_i} = i$$

$$b_{z_i} = 2i + 0.5,\ b_{in_i} = i - 0.5.$$

This completes the construction. From the construction, we can deduce the following lemma.

Lemma 2.2.2. *For $i > 0$, neurons z_i, in_i fire according to the following rules:*

1. *$z_i^{(t)} = 1$ if and only if $x^{(t-1)} = 1$, $in_i^{(t-1)} = 0$, and (either for all $j, 0 \leq j < i$ we have $z_j^{(t-1)} = 1$ or $z_i^{(t-1)} = 1$).*

2. *$in_i^{(t)} = 1$ if and only if for all $j, 1 \leq j \leq i$ we have $z_j^{(t-1)} = 1$.*

Proof. **Case (1):** The potential of $z_i^{(t)}$ is

$$w_{xz_i}x^{(t-1)} + \sum_{j=0}^{i-1} w_{z_jz_i}z_j^{(t-1)} + w_{in_iz_i}in_i^{(t-1)} + w_{z_iz_i}z_i^{(t-1)}$$

$$= (i+1)x^{(t-1)} + \sum_{j=0}^{i-1} z_j^{(t-1)} - (i+1)in_i^{(t-1)} + iz_i^{(t-1)}.$$

Only if: Let's show the only if direction for the firing rule of $z_i^{(t)}$ by proving the contrapositive.

If $x^{(t-1)} = 0$, then the potential of $z_i^{(t)}$ is

$$\sum_{j=0}^{i-1} x_j^{(t-1)} - (i+1)in_i^{(t-1)} + iz_i^{(t-1)} \leq 2i < 2i + 0.5 = b_{z_i}.$$

If $in_i^{(t-1)} = 1$, then the potential of $z_i^{(t)}$ is

$$(i+1)x^{(t-1)} + \sum_{j=0}^{i-1} z_j^{(t-1)} - (i+1) + iz_i^{(t-1)} \leq 2i < 2i + 0.5 = b_{z_i}.$$

If there exists $\hat{j}, 0 \leq \hat{j} < i$ such that $z_{\hat{j}}^{(t-1)} = 0$ and $z_i^{(t-1)} = 0$, then the potential of $z_i^{(t)}$ is

$$\sum_{j \neq \hat{j}, 0 \leq j \leq i-1} z_j^{(t-1)} + (i+1)x^{(t-1)} - (i+1)in_i^{(t-1)} \leq 2i < 2i + 0.5 = b_{z_i}.$$

In all three cases, we have $z_i^{(t)} = 0$.

If: For the if direction, if $x^{(t-1)} = 1$, $in_i^{(t-1)} = 0$ and for all $j, 0 \leq j < i$ we have $z_j^{(t-1)} = 1$, then the potential of $z_i^{(t)}$ is

$$(i+1) + \sum_{j=0}^{i-1} 1 + iz_i^{(t-1)} \geq 2i + 1 > 2i + 0.5 = b_{z_i}.$$

If $x^{(t-1)} = 1$, $in_i^{(t-1)} = 0$ and $z_i^{(t-1)} = 1$, then the potential of $z_i^{(t)}$ is

$$(i+1) + \sum_{j=0}^{i-1} z_j^{(t-1)} + i \geq 2i + 1 > 2i + 0.5 = b_{z_i}.$$

In both cases, we have $z_i^{(t)} = 1$.

Case (2): The firing rule of $in_i^{(t)}$ can be analyzed similarly.

The potential of $in_i^{(t)}$ is

$$\sum_{j=1}^{i} w_{z_j in_i} z_j^{(t-1)} = \sum_{j=1}^{i} z_j^{(t-1)}.$$

Only If: For the only if direction, if there exists $\hat{j}, 1 \leq \hat{j} \leq i$ such that $x_{\hat{j}}^{(t-1)} = 0$, then the potential of $in_i^{(t)}$ is

$$\sum_{j \neq \hat{j}, 1 \leq j \leq i} z_j^{(t-1)} \leq i - 1 < i - 0.5 = b_{in_i}.$$

We have $in_i^{(t)} = 0$.

If: For the if direction, if for all $j, 1 \leq j \leq i$ we have $z_j^{(t-1)} = 1$, then the potential of $in_i^{(t)}$ is

$$\sum_{j=1}^{i} 1 = i > i - 0.5 = b_{in_i}.$$

We have $in_i^{(t)} = 1$ as desired. □

Using the above lemma, we can verify that indeed the network at the first stage fires in binary, with z_i encoding the ith digit in the binary representation.

Theorem 2.2.3. *Given $i \geq 1$ and $t \geq 0$, if for all t' such that $0 \leq t' \leq t$ we have $x^{(t')} = 1$, then*

1. *$z_i^{(t)} = a_i$ for $t = \sum_{j=0}^{\infty} a_j 2^j$ where $a_j \in \{0, 1\}$.*

2. *$in_i^{(t)} = 1$ if and only if $t \bmod 2^{i+1} = 2^{i+1} - 1$ or 0.*

Proof. First, let's verify that the claim is true for z_0. Since for all $t', 0 \leq t' \leq t$ we have $x^{(t')} = 1$, $z_0^{(t')} = 1$ if and only if $z_0^{(t'-1)} = 0$. This implies exactly $z_0^{(t)} = t \bmod 2$ as desired (for all the modular arithematic at this work, we choose the smallest nonnegative number from the equivalence class). Now let's do the induction on t and we will verify the induction by checking z_i, in_i fires in according to the induction hypothesis for all $i \geq 1$. When $t = 1$, the induction statement is trivially satisfied for all $i \geq 1$. Fix i, we have the following cases:

1. $0 < t \bmod 2^{i+1} < 2^i, z_i^{(t-1)} = 0$:

   This implies that $0 \leq t - 1 \bmod 2^i < 2^i - 1$. By induction hypothesis, not all $z_j^{(t-1)} = 1$ for $0 \leq j < i$. Now by Lemma 2.2.2, we have $z_i^{(t)} = 0 = a_i, in_i^{(t)} = 0$ as desired.

2. $t \bmod 2^{i+1} = 2^i, z_i^{(t-1)} = 0, in_i^{(t-1)} = 0$:

   This implies that $t - 1 \bmod 2^i = 2^i - 1$. By induction hypothesis, for all $j, 0 \leq j < i$ we have $z_j^{(t-1)} = 1$. Now by Lemma 2.2.2, we have $z_i^{(t)} = 1 = a_i, in_i^{(t)} = 0$ as desired.

3. $2^i < t \bmod 2^{i+1} < 2^{i+1} - 1, z_i^{(t-1)} = 1, in_i^{(t-1)} = 0$:

   This implies that $0 \leq t - 1 \bmod 2^i < 2^i - 2$. By induction hypothesis, not all $j, 1 \leq j < i$ we have $z_j^{(t-1)} = 1$. Now by Lemma 2.2.2, we have $z_i^{(t)} = 1 = a_i, in_i^{(t)} = 0$ as desired.

4. $t \bmod 2^{i+1} = 2^{i+1} - 1, z_i^{(t-1)} = 1, in_i^{(t-1)} = 0$:

   This implies that $t - 1 \bmod 2^i = 2^i - 2$. By induction hypothesis, for all $j, 1 \leq j < i$ we have $z_j^{(t-1)} = 1$. Now by Lemma 2.2.2, we have $z_i^{(t)} = 1 = a_i, in_i^{(t)} = 1$ as desired.

5. $t \bmod 2^{i+1} = 0, z_i^{(t-1)} = 1, in_i^{(t-1)} = 1$:

   This implies that $t - 1 \bmod 2^i = 2^i - 1$. By induction hypothesis, for all $j, 1 \leq j < i$ we have $z_j^{(t-1)} = 1$,. Now by Lemma 2.2.2, we have $z_i^{(t)} = 0 = a_i, in_i^{(t)} = 1$ as desired.

This completes the induction. □

2.2.2 Second Stage: Capture Network

Now the second stage is a simple "capture network" with input neurons x, z_i for all $i, 0 \leq i \leq n$, output neurons y_i for $0 \leq i \leq n$ and an auxilary neuron s. Intuitively, the network persistently captures the state of z_i for all $i, 0 \leq i \leq n$ into y_i for all $i, 0 \leq i \leq n$. We will specify the timing of the states of z_i being captured later. The network is defined as the following:

$$\forall\, 0 \leq i \leq n, w_{xy_i} = -2, w_{y_iy_i} = 4, w_{z_iy_i} = 1, w_{z_is} = w_{y_is} = 1, w_{sy_i} = -1.5,$$

and

$$w_{xs} = -n - 1, w_{ss} = n + 2, b_s = 0.5, \forall\, 0 \leq i \leq n, b_{y_i} = 0.5.$$

Notice that the above weight ensures the following one step firing rule:

Lemma 2.2.4. *For $0 \leq i \leq n$, neurons $y_i^{(t)}, s^{(t)}$ fire according to the following rules:*

1. *$y_i^{(t)} = 1$ if and only if $y_i^{(t-1)} = 1$, or ($y_i^{(t-1)} = 0, x^{(t-1)} = 0, s^{(t-1)} = 0$ and $z_i^{(t-1)} = 1$).*

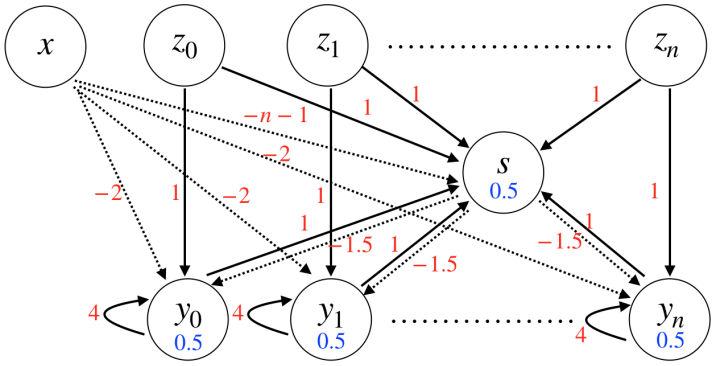


Figure 2-3: Second Stage

2. $s^{(t)} = 1$ *if and only if* $s^{(t-1)} = 1$, *or (there exists* i, i' *such that* $z_i^{(t-1)} = 1$ *or* $y_{i'}^{(t-1)} = 1$, *and* $x^{(t-1)} = 0$*).*

Proof. **Case (1):** The potential of $y_i^{(t)}$ is

$$w_{xy_i}x^{(t-1)} + w_{y_iy_i}y_i^{(t-1)} + w_{z_iy_i}z_i^{(t-1)} + w_{z_iz_i}z_i^{(t-1)} + w_{sy_i}s^{(t-1)}$$
$$= -2x^{(t-1)} + 4y_i^{(t-1)} + z_i^{(t-1)} - 1.5s^{(t-1)}.$$

Only If: Let's show the only if direction for the firing rule of $y_i^{(t)}$ first. If $y_i^{(t-1)} = 0, x^{(t-1)} = 1$, the potential of $y_i^{(t)}$ is

$$-2 + z_i^{(t-1)} - 1.5s^{(t-1)} \leq -1 < 0.1 = b_{y_i}.$$

If $y_i^{(t-1)} = 0, s^{(t-1)} = 1$, the potential of $y_i^{(t)}$ is

$$-2x^{(t-1)} + z_i^{(t-1)} - 1.5 \leq -0.5 < 0.1 = b_{y_i}.$$

If $y_i^{(t-1)} = 0, z_i^{(t-1)} = 0$, the potential of $y_i^{(t)}$ is

$$-2x^{(t-1)} - 1.5s^{(t-1)} \leq 0 < 0.1 = b_{y_i}.$$

In all three cases, we have $y_i^{(t)} = 0$.

If: For the if direction, if $y_i^{(t-1)} = 1$, then the potential of $y_i^{(t)}$ is

$$-2x^{(t-1)} + 4 + z_i^{(t-1)} - 1.5s^{(t-1)} \geq 0.5 > 0.1 = b_{y_i}.$$

If $y_i^{(t-1)} = 0, x^{(t-1)} = 0, s^{(t-1)} = 0, z_i^{(t-1)} = 1$, the potential of $y_i^{(t)}$ is

$$4y_i^{(t-1)} + 1 \geq 1 > 0.1 = b_{y_i}.$$

In both cases, we have $y_i^{(t)} = 1$.

Case (2): The potential of $s^{(t)}$ is

$$\sum_{j=0}^{n} w_{z_j s} z_j^{(t-1)} + \sum_{j=0}^{n} w_{y_j s} y_j^{(t-1)} + w_{xs} x^{(t-1)} + w_{ss} s^{(t-1)}$$
$$= \sum_{j=0}^{n} z_j^{(t-1)} + \sum_{j=0}^{n} y_j^{(t-1)} - (n+1)x^{(t-1)} + (n+2)s^{(t-1)}.$$

Only If: For the only if direction, if $s^{(t-1)} = 0$ and for all $j, 0 \leq j \leq n$ we have $y_j^{(t-1)} = z_j^{(t-1)} = 0$, then the potetntial of $s^{(t)}$ is

$$-(n+1)x^{(t-1)} \leq 0 < 0.5 = b_s.$$

If $s^{(t-1)} = 0, x^{(t-1)} = 1$, the potetntial of $s^{(t)}$ is

$$\sum_{j=0}^{n} z_j^{(t-1)} + \sum_{j=0}^{n} z_j^{(t-1)} - (n+1) \leq 0 < 0.5 = b_s.$$

In both cases, we have $s^{(t)} = 0$.

If: For the if direction, if there exists $i, 0 \leq i \leq n$ such that $y_i^{(t-1)} = 1$ and $x^{(t-1)} = 0$, then the potential of $s^{(t)}$ is

$$\sum_{j=0}^{n} z_j^{(t-1)} + \sum_{j \neq i, 0 \leq j \leq n}^{n} y_j^{(t-1)} + 1 + (n+2)s^{(t-1)} \geq 1 > 0.5 = b_s.$$

If there exists $i, 0 \leq i \leq n$ such that $z_i^{(t-1)} = 1$ and $x^{(t-1)} = 0$, the potential of $s^{(t)}$ is

$$\sum_{j=0}^{n} y_j^{(t-1)} + \sum_{j \neq i, 0 \leq j \leq n}^{n} z_j^{(t-1)} + 1 + (n+2)s^{(t-1)} \geq 1 > 0.5 = b_s.$$

If $s^{(t-1)} = 1$, the potential of $s^{(t)}$ is

$$\sum_{j=0}^{n} z_j^{(t-1)} + \sum_{j=0}^{n} y_j^{(t-1)} - (n+1)x^{(t-1)} + (n+2) \geq 1 > 0.5 = b_s.$$

In all three cases, we have $s^{(t)} = 1$ as desired. □

Now we can describe the behaviors of the capture network in the following theorem. The network persistantly captures the state of z_i for all $i, 0 \leq i \leq n$ at the first time point such that $x = 0$ and there exists some $\hat{i}$ such that $z_{\hat{i}} = 1$ into y_i for all $i, 0 \leq i \leq n$.

Theorem 2.2.5. *For the network at the second stage, let $t' \geq 0$ be such that $x^{(t')} = 0$ and there exists $\hat{j}$ such that $z_{\hat{j}}^{(t')} = 1$, and for all $t, 0 \leq t < t'$, either $x^{(t)} = 1$ or for all $i, 0 \leq i \leq n$ we have $z_i^{(t)} = 0$. Then for all i, t such that $0 \leq i \leq n, t > t'$ we have $y_i^{(t)} = z_i^{(t')}$.*

Proof. First by Lemma 2.2.4, for all $t, 0 < t \leq t'$ and for all $i, 0 \leq i \leq n$ we have $y_i^{(t)} = s^{(t)} = 0$. Now at time $t' + 1$, by Lemma 2.2.4, we see that $y_i^{(t'+1)} = z_i^{(t')}, \forall i, 0 \leq i \leq n$ and $s^{(t'+1)} = 1$. Now by Lemma 2.2.4, we know that for all $t, t > t'$ we have $s^{(t)} = 1$. Now by Lemma 2.2.4 again, if $y_i^{(t'+1)} = 0$, then since for all $t, t > t'$ we have $s^{(t)} = 1$, for all $t > t'$ we have $y_i^{(t)} = 0$; and if $y_i^{(t'+1)} = 1$, then we also have for all $t, t > t'$, $y_i^{(t)} = 1$ as desired. □

2.2.3 Wrap up

Now we are ready to prove the main Theorem 2.1.1 by setting $n = m = \lceil \log T' \rceil$

Proof. We are going to prove the main theorem by composing the networks from stage one and two together. If for all $t, 0 \leq t \leq T$ we have $x^{(t)} = 0$, then the network satisfies the criterion trivially since for all $0 \leq t \leq T$, $y_i^{(t)} = 0$. If not, then there exists $\hat{t} \geq 0, L_x > 0$ such that for all $t, 0 \leq t < \hat{t}$ we have $x^{(t)} = 0$, for all $i, 0 \leq i < L_x$ we have $x^{(\hat{t}+i)} = 1$, and $x^{(\hat{t}+L_x)} = 0$ where L_x is the length of the first consecutive spikes interval. Let $L_x = \sum_{j=0}^{\infty} a_j 2^j$; then by Theorem 2.2.3 and Lemma 2.2.1, for all $i, 0 \leq i \leq n$, we have $z_i^{(\hat{t}+L_x-1)} = a_i$. Now because $L_x > 0$, we know there exists $\hat{j}$ such that $z_{\hat{j}}^{(\hat{t}'+L_x)} = 1$ by Theorem 2.2.3. And by Lemma 2.2.2, we know for all i, t such that $0 \leq t \leq \hat{t}, 0 \leq i \leq n$, we have $z_i^{(t)} = 0$. Now the assumption of Theorem 2.2.5 is satisfied with $t' = \hat{t} + L_x$. By Theorem 2.2.5, we get for all t, i such that $0 \leq i \leq n, t \geq \hat{t} + L_x$ we have $y_i^{(t)} = a_i$ and $T + 1 \geq \hat{t} + L_x$ as desired. This shows that the above network solves FCSC(T) problem in time 1 with $O(\log T)$ neurons. □

Notice that in fact by the proof above, FCSC network enjoys an early convergence property. The network actually converges at time $\hat{t} + L_x$. Therefore we have the following stronger version of Theorem 2.1.1.

Corollary 2.2.6. *For all $t, 0 \leq t \leq T$, FCSC network with $O(\log T)$ neurons solves FCSC(t) problem in time* 1.

2.3 Total Spikes Counting

To count the total number of spikes in an arbitrary interval requires the persistence of neurons without external spikes. Notice that on FCSC network, each neuron toggles itself according to binary representation without delay. However, the persistence of neurons and toggles without delays are conflicting objectives; persistence of neurons stabilizes the network while toggling without delays changes the firing patterns of the network. For example, we use self-inhibition to count mod 2 but if we use self-inhibition to count mod 2, the neuron cannot maintain the count during intervals with no inputs. In this section, we circumvent this difficulty by allowing the network to enter an unstable intermediate state that still stores the information of the count when the spikes arrive; however, the network will converge to a *clean state* that according to binary representation after one step of computation without external signals, and this *clean state* is stable in an arbitrary interval with no input.

In this section, because the self-inhibition used in Section 3 to count mod 2 cannot induce persistence, we build a network of four neurons to count mod 4 to replace the function of z_0, z_1 in Section 3. We then iteratively build the rest of the network that approximately fires in binary on top of the mod 4 counter network.

2.3.1 Mod 4 Counter Network

The construction of the mod 4 counter network is the following:

$$w_{xf_i} = 1, w_{f_if_i} = 2, 0 \leq i \leq 3, w_{f_{j+1}f_j} = -3, 0 \leq j \leq 2,$$
$$w_{f_1f_2} = w_{f_2f_3} = w_{f_3f_0} = 1, w_{f_0f_3} = w_{f_3f_1} = -3$$

and

$$b_{f_1} = 0.5, b_{f_i} = 1.5, i \neq 1.$$

We have the following lemma to specify the firing rules of f_i:

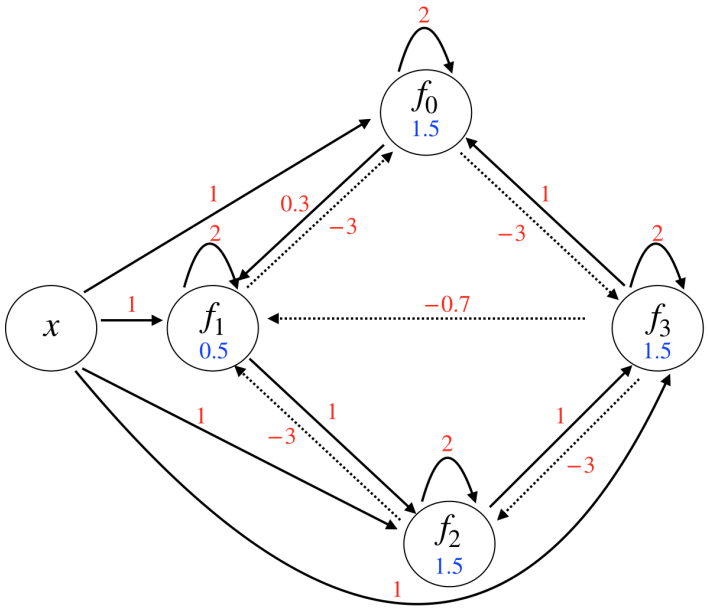


Figure 2-4: mod 4 Counter Network

Lemma 2.3.1. *For all t, i such that $t \geq 1, 0 \leq i < 4$, neurons $f_i^{(t)}$ fire according to the following rules:*

1. *$f_1^{(t)} = 1$ if and only if $f_2^{(t-1)} = 0$, and ($x^{(t-1)} = 1, f_3^{(t-1)} = 0$ or $f_1^{(t-1)} = 1$ or $x^{(t-1)} = 1, f_0^{(t-1)} = 1$).*

2. *For $i \neq 1$ we have $f_i^{(t)} = 1$ if and only if $f_{(i+1) \bmod 4}^{(t-1)} = 0$, and ($x^{(t-1)} = 1, f_{(i-1) \bmod 4}^{(t-1)} = 1$ or $f_i^{(t-1)} = 1$).*

Proof. **Case (1):** The potential of $f_1^{(t)}$ is

$$w_{xf_1}x^{(t-1)} + w_{f_1f_1}f_1^{(t-1)} + w_{f_2f_1}f_2^{(t-1)} + w_{f_3f_1}f_3^{(t-1)} + w_{f_0f_1}f_0^{(t-1)}$$
$$= x^{(t-1)} + 2f_1^{(t-1)} - 3f_2^{(t-1)} - 0.7f_3^{(t-1)} + 0.3f_0^{(t-1)}.$$

Only If: Let's show the only if direction for the firing rule of $f_1^{(t)}$ first. If $f_2^{(t-1)} = 1$, then the potential of $f_1^{(t)}$ is

$$x^{(t-1)} + 2f_1^{(t-1)} - 3 - 0.7f_3^{(t-1)} + 0.3f_0^{(t-1)} \leq 0.3 < 0.5 = b_{f_1}.$$

If $f_1^{(t-1)} = 0, x^{(t-1)} = 0$, then the potential of $f_1^{(t)}$ is

$$-3f_2^{(t-1)} - 0.7f_3^{(t-1)} + 0.3f_0^{(t-1)} \leq 0.3 < 0.5 = b_{f_1}.$$

If $f_1^{(t-1)} = 0, f_3^{(t-1)} = 1, f_0^{(t-1)} = 0$, then the potential of $f_1^{(t)}$ is

$$x^{(t-1)} - 3f_2^{(t-1)} - 0.7 \leq 0.3 < 0.5 = b_{f_1}.$$

In all three cases, we have $f_1^{(t)} = 0$.

If: For the if direction, if $f_2^{(t-1)} = 0, f_1^{(t-1)} = 1$, then the potential of $f_1^{(t)}$ is

$$x^{(t-1)} + 2 - 0.7f_3^{(t-1)} + 0.3f_0^{(t-1)} \geq 1.3 > 0.5 = b_{f_1}.$$

If $f_2^{(t-1)} = 0, x^{(t-1)} = 1, f_3^{(t-1)} = 0$, then the potential of $f_1^{(t)}$ is

$$1 + 2f_1^{(t-1)} + 0.3f_0^{(t-1)} \geq 1 > 0.5 = b_{f_1}.$$

If $f_2^{(t-1)} = 0, x^{(t-1)} = 1, f_0^{(t-1)} = 1$, then the potential of $f_1^{(t)}$ is

$$1 + 2f_1^{(t-1)} - 0.7f_3^{(t-1)} + 0.3 \geq 0.6 > 0.5 = b_{f_1}.$$

In all three cases, we have $f_1^{(t)} = 1$.

Case (2): For $i \neq 1$, The potential of $f_i^{(t)}$ is

$$w_{xf_i}x^{(t-1)} + w_{f_if_i}f_i^{(t-1)} + w_{f_{(i-1) \bmod 4}f_i}f_{(i-1) \bmod 4}^{(t-1)} + w_{f_{(i+1) \bmod 4}f_i}f_{(i+1) \bmod 4}^{(t-1)}$$
$$= x^{(t-1)} + 2f_i^{(t-1)} + f_{(i-1) \bmod 4}^{(t-1)} - 3f_{(i+1) \bmod 4}^{(t-1)}.$$

Only If: For the only if direction, if $f_{(i+1) \bmod 4}^{(t-1)} = 1$, then the potential of $f_i^{(t)}$ is

$$x^{(t-1)} + 2f_i^{(t-1)} + f_{(i-1) \bmod 4}^{(t-1)} - 3 \leq 1 < 1.5 = b_i.$$

If $x^{(t-1)} = 0, f_i^{(t-1)} = 0$, then the potential of $f_i^{(t)}$ is

$$f_{(i-1) \bmod 4}^{(t-1)} - 3f_{(i+1) \bmod 4}^{(t-1)} \leq 1 < 1.5 = b_i.$$

If $f_{(i-1) \bmod 4}^{(t-1)} = 0, f_i^{(t-1)} = 0$, then the potential of $f_i^{(t)}$ is

$$x^{(t-1)} - 3f_{(i+1) \bmod 4}^{(t-1)} \leq 1 < 1.5 = b_i.$$

In all three cases, we have $f_i^{(t)} = 0$.

If: For the if direction, if $f_{(i+1) \bmod 4}^{(t-1)} = 0, x^{(t-1)} = 1, f_{(i-1) \bmod 4}^{(t-1)} = 1$, then the potential of $f_i^{(t)}$ is

$$1 + 2f_i^{(t-1)} + 1 \geq 2 > 1.5 = b_i.$$

If $f^{(t-1)}_{(i+1) \bmod 4} = 0, f^{(t-1)}_i = 1$, then the potential of $f^{(t)}_i$ is

$$x^{(t-1)} + 2 + f^{(t-1)}_{(i-1) \bmod 4} \geq 2 > 1.5 = b_i.$$

In both cases, we have $f^{(t)}_i = 1$ as desired. □

For $0 \leq i < 4$, define a *clean state* with value i at time t' of the mod 4 counter network to be a state in which $f^{(t')}_i = 1$ and for all $j, j \neq i$ we have $f^{(t')}_j = 0$. By Lemma 2.3.1, it is trivial to see that if for all $t, t \geq t'$ we have $x^{(t)} = 0$, then for all $t, t \geq t'$ and for all $i, 0 \leq i < 4$ we have $f^{(t)}_i = f^{(t')}_i$. Using Lemma 2.3.1, we have the following lemma describing the behaviors of mod 4 counter network. Intuitively, when a new input arrives, the network enters an intermediate state in which both neurons represent the old count and the new count fire; when there is no input, the neuron that represents the new count will inhibit the neuron that represents the old count to stabilize the network in a *clean state*.

Lemma 2.3.2. *Let the mod* 4 *counter network be at a clean state with value* $\hat{i}$ *at time* t'*. Fix a positive integer* L*. For all* $i, 0 \leq i < L$*, let* $x^{(t'+i)} = 1$ *and* $x^{(t'+L)} = 0$*. Then, at time* $t, t' < t < t' + L + 1$*, we have the state of the network being*

$$f^{(t)}_{(\hat{i}+t-t') \bmod 4} = f^{(t)}_{(\hat{i}+t-t'-1) \bmod 4} = 1, f^{(t)}_{(\hat{i}+t-t'-2) \bmod 4} = f^{(t)}_{(\hat{i}+t-t'-3) \bmod 4} = 0.$$

Furthermore, the network will be at a clean state again at time $t' + L + 1$ *with* $f^{(t'+L+1)}_{(\hat{i}+L) \bmod 4} = 1$.

Proof. First, let's use induction on t to prove at time $t, t' < t < t' + L + 1$, we have the state of the network be

$$f^{(t)}_{(\hat{i}+t-t') \bmod 4} = f^{(t)}_{(\hat{i}+t-t'-1) \bmod 4} = 1, f^{(t)}_{(\hat{i}+t-t'-2) \bmod 4} = f^{(t)}_{(\hat{i}+t-t'-3) \bmod 4} = 0.$$

Base Case: By Lemma 2.3.1, we have

$$f^{(t'+1)}_{(\hat{i}+1) \bmod 4} = f^{(t'+1)}_{(\hat{i}+t-t') \bmod 4} = 1, f^{(t'+1)}_{(\hat{i}-1) \bmod 4} = f^{(t'+1)}_{(\hat{i}-2) \bmod 4} = 0$$

for the base case.

Inductive Step: Now assume the induction hypothesis is true for $t = k$, since we

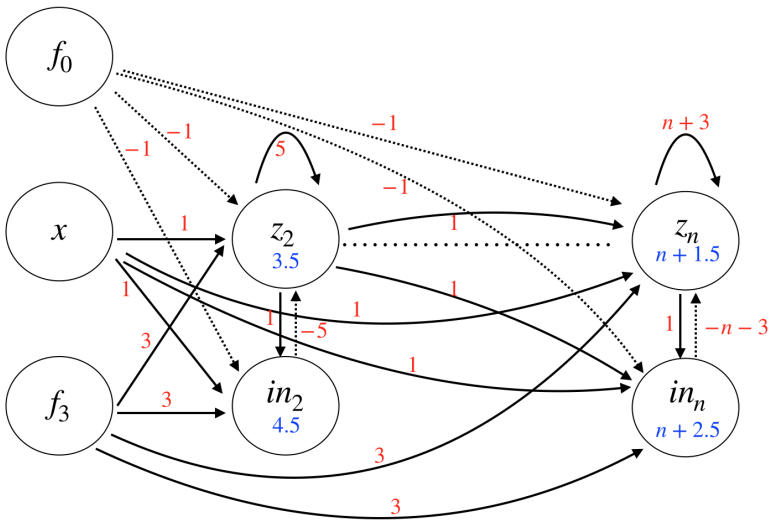


Figure 2-5: Total spikes counting (TSC) Network

have $x^{(k)} = 1$ by Lemma 2.3.1, we indeed have

$$f^{(k+1)}_{(\hat{i}+k+1-t') \bmod 4} = f^{(k+1)}_{(\hat{i}+k+1-t'-1) \bmod 4} = 1, f^{(k+1)}_{(\hat{i}+k+1-t'-2) \bmod 4} = f^{(k+1)}_{(\hat{i}+k+1-t'-3) \bmod 4} = 0.$$

This completes the induction.

Now since $x^{(t'+L)} = 0$, by Lemma 2.3.1 we can derive the state of the network at time $t' + L + 1$

$$f^{(t'+L+1)}_{(\hat{i}+L) \bmod 4} = 1, f^{(t'+L+1)}_{j} = 0, \ \forall j \neq (\hat{i} + L) \bmod 4$$

as desired. □

2.3.2 TSC Network

Now we iteratively build the network with the following rule on top of the mod 4 counter network,

$$w_{f_3 z_i} = w_{f_3 in_i} = 3, w_{f_0 z_i} = w_{f_0 in_i} = -1, w_{x z_i} = w_{x in_i} = 1,$$
$$w_{z_j z_i} = w_{z_j in_i} = 1, \ \forall j, 2 \leq j < i, w_{in_i z_i} = -i - 3, w_{z_i in_i} = 1, w_{z_i z_i} = i + 3$$

and

$$b_{z_i} = i + 1.5, \ b_{in_i} = i + 2.5.$$

In the full construction of the TSC network, intuitively, we replace the function of

z_0, z_1 in Section 3 with a mod 4 counter network. We design the weights coming from f_3, f_0 such that they will induce proper carry in an approximate binary representation at $z_i, i \geq 2$, and we use a similar idea as the mod 4 counter network to make TSC network converge to an exact binary representation in one computation step without input.

The following lemma specifies the firing rules of z_i, in_i for $i \geq 2$:

Lemma 2.3.3. *For $i \geq 2$, neurons $z_i^{(t)}, in_i^{(t)}$ fire according to the following rules:*

1. *$z_i^{(t)} = 1$ if and only if $in_i^{(t-1)} = 0$, and either ($f_3^{(t-1)} = 1, f_0^{(t-1)} = 0, x^{(t-1)} = 1$ and for all $j, 2 \leq j < i$ we have $z_j^{(t-1)} = 1$) or $z_i^{(t-1)} = 1$.*

2. *$in_i^{(t)} = 1$ if and only if $z_i^{(t-1)} = 1, f_3^{(t-1)} = 1, f_0^{(t-1)} = 0, x^{(t-1)} = 1$ and for all $j, 2 \leq j < i$ we have $z_j^{(t-1)} = 1$.*

Proof. **Case (1):** The potential of $z_i^{(t)}$ is

$$w_{f_3 z_i} f_3^{(t-1)} + w_{f_0 z_i} f_0^{(t-1)} + \sum_{j=2}^{i-1} w_{z_j z_i} z_j^{(t-1)} + w_{z_i z_i} z_i^{(t-1)} + w_{in_i z_i} in_i^{(t-1)} + w_{x z_i} x^{(t-1)}$$
$$= 3f_3^{(t-1)} - f_0^{(t-1)} + \sum_{j=2}^{i-1} z_j^{(t-1)} + (i+3) z_i^{(t-1)} - (i+3) in_i^{(t-1)} + x^{(t-1)}.$$

Only If: Let's show the only if direction for the firing rule of $z_i^{(t)}$ first. If $in_i^{(t-1)} = 1$, the potential of $z_i^{(t)}$ is

$$3f_3^{(t-1)} - f_0^{(t-1)} + \sum_{j=2}^{i-1} z_j^{(t-1)} + (i+3) z_i^{(t-1)} - (i+3) + x^{(t-1)} \leq i + 1 < i + 1.5 = b_{z_i}.$$

If $f_3^{(t-1)} = 0, z_i^{(t-1)} = 0$, the potential of $z_i^{(t)}$ is

$$-f_0^{(t-1)} + \sum_{j=2}^{i-1} z_j^{(t-1)} - (i+3) in_i^{(t-1)} + x^{(t-1)} \leq i - 1 < i + 1.5 = b_{z_i}.$$

If $f_0^{(t-1)} = 1, z_i^{(t-1)} = 0$, the potential of $z_i^{(t)}$ is

$$3f_3^{(t-1)} - 1 + \sum_{j=2}^{i-1} z_j^{(t-1)} - (i+3) in_i^{(t-1)} + x^{(t-1)} \leq i + 1 < i + 1.5 = b_{z_i}.$$

If $x^{(t-1)}=0, z_i^{(t-1)}=0$, the potential of $z_i^{(t)}$ is

$$3f_3^{(t-1)}-f_0^{(t-1)}+\sum_{j=2}^{i-1}z_j^{(t-1)}-(i+3)in_i^{(t-1)}\leq i+1<i+1.5=b_{z_i}.$$

If $z_i^{(t-1)}=0$ and there exists $\hat{j}, 2\leq\hat{j}<i$ such that $z_{\hat{j}}^{(t-1)}=0$, the potential of $z_i^{(t)}$ is

$$3f_3^{(t-1)}-f_0^{(t-1)}+\sum_{j\neq\hat{j},2\leq j<i}z_j^{(t-1)}-(i+3)in_i^{(t-1)}+x^{(t-1)}\leq i+1<i+1.5=b_{z_i}.$$

In all cases, we have $z_i^{(t)}=0$.

If: For the if direction, if $in_i^{(t-1)}=0, f_3^{(t-1)}=1, f_0^{(t-1)}=0, x^{(t-1)}=1$ and for all $j, 2\leq j<i$ we have $z_j^{(t-1)}=1$, then the potential of $z_i^{(t)}$ is

$$3+\sum_{j=2}^{i-1}1+(i+3)z_i^{(t-1)}+1\geq i+2>i+1.5=b_{z_i}.$$

If $in_i^{(t-1)}=0, z_i^{(t-1)}=1$, the potential of $z_i^{(t)}$ is

$$3f_3^{(t-1)}-f_0^{(t-1)}+\sum_{j=2}^{i-1}z_j^{(t-1)}+(i+3)+x^{(t-1)}\geq i+2>i+1.5=b_{z_i}.$$

In both cases, we have $z_i^{(t)}=1$.

Case (2): The potential of $in_i^{(t)}$ is

$$w_{f_3in_i}f_3^{(t-1)}+w_{f_0in_i}f_0^{(t-1)}+\sum_{j=2}^{i-1}w_{z_jin_i}z_j^{(t-1)}+w_{z_iin_i}z_i^{(t-1)}+w_{xin_i}x^{(t-1)}$$
$$=3f_3^{(t-1)}-f_0^{(t-1)}+\sum_{j=2}^{i-1}z_j^{(t-1)}+z_i^{(t-1)}+x^{(t-1)}.$$

Only If: For the only if direction, if $z_i^{(t-1)}=0$, then the potential of $in_i^{(t)}$ is

$$3f_3^{(t-1)}-f_0^{(t-1)}+\sum_{j=2}^{i-1}z_j^{(t-1)}+x^{(t-1)}\leq i+2<i+2.5=b_{in_i}.$$

If $f_3^{(t-1)}=0$, the potential of $in_i^{(t)}$ is

$$-f_0^{(t-1)}+\sum_{j=2}^{i-1}z_j^{(t-1)}+z_i^{(t-1)}+x^{(t-1)}\leq i<i+2.5=b_{in_i}.$$

If $f_0^{(t-1)} = 1$, the potential of $in_i^{(t)}$ is

$$3f_3^{(t-1)} - 1 + \sum_{j=2}^{i-1} z_j^{(t-1)} + z_i^{(t-1)} + x^{(t-1)} \leq i + 2 < i + 2.5 = b_{in_i}.$$

If $x^{(t-1)} = 0$, the potential of $in_i^{(t)}$ is

$$3f_3^{(t-1)} - f_0^{(t-1)} + \sum_{j=2}^{i-1} z_j^{(t-1)} + z_i^{(t-1)} \leq i + 2 < i + 2.5 = b_{in_i}.$$

If there exists $\hat{j}, 2 \leq \hat{j} < i$ such that $z_{\hat{j}} = 0$, the potential of $in_i^{(t)}$ is

$$3f_3^{(t-1)} - f_0^{(t-1)} + \sum_{j \neq \hat{j}, 2 \leq j < i} z_j^{(t-1)} + z_i^{(t-1)} + x^{(t-1)} \leq i + 2 < i + 2.5 = b_{in_i}.$$

In all cases, $in_i^{(t)} = 0$.

If: For the if direction, if $z_i^{(t-1)} = 1, f_3^{(t-1)} = 1, f_0^{(t-1)} = 0, x^{(t-1)} = 1$ and for all $j, 2 \leq j < i$ we have $z_j^{(t-1)} = 1$, then the potential of $in_i^{(t)}$ is

$$3 + \sum_{j=2}^{i-1} 1 + 1 + 1 \leq i + 3 > i + 2.5 = b_{in_i}.$$

We have $in_i^{(t)} = 1$ as desired. □

Define a clean state at time t' of TSC network with value X stored be one in which

1. $f_{X \bmod 4}^{(t')} = 1, f_j^{(t')} = 0, \ \forall j \neq X \bmod 4$ (i.e., the mod 4 counter subnetwork is clean with value X mod 4).

2. For $X = \sum_{i=0}^{\infty} a_i 2^i, a_i \in \{0, 1\}$, $z_k^{(t')} = a_k, \forall k \geq 2$.

3. $in_i^{(t')} = 0$ if $X \bmod 2^{i+1} = 2^{i+1} - 1$.

So being at a clean state for TSC network with value X stored implies being at a clean state with value X mod 4 for its mod 4 counter subnetwork with z_i in binary representation for $i \geq 2$. By Lemma 2.3.3, it is trivial to see that if for all $t \geq t'$ we have $x^{(t)} = 0$, then for all $i \geq 2$ and for all $t, t \geq t'$ we have $f_i^{(t)} = f_i^{(t')}$. Using Lemma 2.3.3, we have the following lemma describing the behaviors of the TSC network.

Lemma 2.3.4. *Let TSC network be at a clean state at time t' with value X stored. Fix a positive integer L. For all i such that $0 \leq i < L$, let $x^{(t'+i)} = 1$ and $x^{(t'+L)} = 0$. Then, at $t, t' < t < t' + L + 1$, z_i, in_i fire with the following rules for all $i \geq 2$:*

1. *for $1 = X + t - t' \bmod 2^{i+1} < 2^i$, $z_i^{(t)} = 0$.*
2. *for $1 < X + t - t' \bmod 2^{i+1} < 2^i$, $z_i^{(t)} = in_i^{(t)} = 0$.*
3. *for $X + t - t' \bmod 2^{i+1} \geq 2^i$, we have $z_i^{(t)} = 1, in_i^{(t)} = 0$.*
4. *for $X + t - t' \bmod 2^{i+1} = 0$, we have $z_i^{(t)} = 1, in_i^{(t)} = 1$.*

Furthermore, the network will be at a clean state with value $X + L$ stored at time $t' + L + 1$.

Proof. Just like the mod 4 counter network case, we want to deduce the behaviors of network at $t, t' < t < t' + L + 1$ using induction first.
Base Case: Fix i, for $t = t' + 1$, we have the following cases

1. $0 < X + 1 \bmod 2^{i+1} < 2^i$:
   This implies that $0 \leq X \bmod 2^{i+1} < 2^i - 1$. This shows that not all $j, j < i$ we have $z_j^{(t-1)} = 1$ or $f_3^{(t-1)} = 0$ or $f_0^{(t-1)} = 1$. By Lemma 2.3.3, we have $z_i^{(t)} = in_i^{(t)} = 0$.

2. $X + 1 \bmod 2^{i+1} \geq 2^i$:
   This implies that $2^i - 1 \leq X \bmod 2^{i+1} < 2^{i+1} - 1$. This shows that either for all $j, j < i$ we have $f_3^{(t-1)} = 1, f_0^{(t-1)} = 0, z_j^{(t-1)} = 1$ or $z_i^{(t-1)} = 1$ but not both. By Lemma 2.3.3, we have $z_i^{(t)} = 1, in_i^{(t)} = 0$.

3. $X + 1 \bmod 2^{i+1} = 0$:
   This implies that $X \bmod 2^{i+1} = 2^{i+1} - 1$. This shows that $f_3^{(t-1)} = 1, f_0^{(t-1)} = 0$ and for all $j \leq i$ we have $z_j^{(t-1)} = 1$ and by the definition of a clean state, we have $in_i^{(t-1)} = 0$. Now by Lemma 2.3.3, we have $z_i^{(t)} = 1, in_i^{(t)} = 1$.

Inductive Step: Assume the induction hypothesis is accurate for $t = k$. We have the following cases

1. $1 = X + k + 1 - t' \bmod 2^{i+1} < 2^i$:

   This implies that $X + k - t' \bmod 2^{i+1} = 0$. Now by induction hypothesis and Lemma 2.3.2, we know that $f_3^{(k)} = 1, f_0^{(k)} = 0$ and for all $j, i \geq j \geq 2$ we have $z_j^{(k)} = 1, in_j^{(k)} = 1$. By Lemma 2.3.3, we have $z_i^{(k+1)} = 0, in_i^{(k+1)} = 1$.

2. $1 < X + k + 1 - t' \bmod 2^{i+1} < 2^i$:

   This implies that $1 \leq X + k - t' \bmod 2^{i+1} < 2^i - 1$. By induction hypothesis and Lemma 2.3.2, this shows that not all $j, j < i$ we have $z_j^{(k)} = 1$ or $f_3^{(k)} = 0$ or $f_0^{(k)} = 1$. By Lemma 2.3.3, we have $x_i^{(k+1)} = in_i^{(k+1)} = 0$.

3. $X + k + 1 - t' \bmod 2^{i+1} \geq 2^i$:

   This implies that $2^i - 1 \leq X + k - t' \bmod 2^{i+1} < 2^{i+1} - 1$. By induction hypothesis and Lemma 2.3.2, this shows that either for all $j, j < i$ we have $f_3^{(k)} = 1, f_0^{(k)} = 0, z_j^{(k)} = 1$ or $z_i^{(k)} = 1$ but not both. By Lemma 2.3.3, we have $z_i^{(k+1)} = 1, in_i^{(k+1)} = 0$.

4. $X + k + 1 - t' \bmod 2^{i+1} = 0$:

   This implies that $X + k - t' \bmod 2^{i+1} = 2^{i+1} - 1$. By induction hypothesis and Lemma 2.3.2, this shows that all $f_3^{(k)} = 1, f_0^{(k)} = 0, in_i^{(k)} = 0$ and for all $j, j \leq i$ we have $z_j^{(k)} = 1$. Now by Lemma 2.3.3, we have $z_i^{(t)} = 1, in_i^{(t)} = 1$.

This completes the induction.

Now we just need to show that at time $t' + L + 1$ the network is at a clean state with value $X + L$ stored. We have the following cases:

1. $1 = X + L \bmod 2^{i+1} < 2^i$:

   By above induction, we have for $j, j \leq i$, $z_j^{(t'+L)} = 0$. No matter what the value of $in_i^{(t'+L)}$ is, by Lemma 2.3.3 we have $z_i^{(t'+L+1)} = in_i^{(t'+L+1)} = 0$.

2. $1 < X + L \bmod 2^{i+1} < 2^i$, $z_i^{(t)} = in_i^{(t)} = 0$:

   By above induction, we have $z_i^{(t'+L)} = in_i^{(t'+L)} = 0$. By Lemma 2.3.3, we have $z_i^{(t'+L+1)} = in_i^{(t'+L+1)} = 0$.

3. $X + L \bmod 2^{i+1} \geq 2^i$, we have $z_i^{(t'+L)} = 1, in_i^{(t'+L)} = 0$. By Lemma 2.3.3, we have $z_i^{(t'+L+1)} = in_i^{(t'+L+1)} = 0$.

4. $X + L \bmod 2^{i+1} = 0$, we have $z_i^{(t'+L)} = 1, in_i^{(t'+L)} = 1$. By Lemma 2.3.3, we have $z_i^{(t'+L+1)} = 0, in_i^{(t'+L+1)} = 1$.

which is exactly a clean state with value $X + L$ stored combining with Lemma 2.3.2. □

2.3.3 Wrap up

Now we are ready for the main proof of Theorem 2.1.2 by setting $n = \lceil \log T' \rceil$ and let $f_i, z_j, 0 \leq i \leq 3, 2 \leq j \leq n$ be our output neurons.

Proof. Let $f_i, z_j, 0 \leq i < 4, 2 \leq j \leq n$ be our output neurons. Let there be X spikes in T time steps. Let $[t_0, t_0+X_0-1], \cdots, [t_k, t_k+X_k-1]$ be the disjoint maximal intervals of spikes ordered by time (i.e., $x^{(t)} = 1$ if $t \in [t_i, t_i + X_i - 1]$ for some $0 \leq i \leq k$ and $[t_i, t_i + X_i] \cap [t_j, t_j + X_j] = \emptyset$ for all $i \neq j$ and $t_0 < t_1 < \cdots < t_k$, $\sum_{i=0}^{k} X_k = X$). Now I claim that at time $t_i + X_i + 1$, the network is at a clean state with value $\sum_{j=0}^{i} X_j$ stored. We will prove the claim with induction on i. For $i = 0$, apply Lemma 2.3.4, we get that the network is at a clean state with value X_0 stored. Assume the network is at a clean state with value $\sum_{j=0}^{i} X_j$ stored at time $t_i + X_i + 1$. Then apply Lemma 2.3.4 again, we get at time $t_{i+1} + X_{i+1} + 1$, the network is at a clean state with value $\sum_{j=0}^{i+1} X_j$ stored at time $t_{i+1} + X_{i+1} + 1$. So at time $t_k + X_k + 1 \leq T + 1$, the network is at a clean state with value $\sum_{j=0}^{k} X_j = X$ stored as desired. This shows that the above network solves TSC(T) problem in time 1 with $O(\log T)$ neurons. □

Notice that in fact by the proof above, TSC network enjoys an early convergence property. The network actually converges at time $t_k + X_k + 1$. Therefore we have the following stronger version of Theorem 2.1.2.

Corollary 2.3.5. *For all $t, 0 \leq t \leq T$, TSC network with $O(\log T)$ neurons solves FCSC(t) problem in time* 1.

2.4 Time Lower Bound for FCSC and TSC

In Section 4, we mentioned that there is a conflicting objective between stabilizing the output and toggling without delays. We therefore introduced the idea of carrying

information of the count at an unclean state and then converging to a clean state, which introduces one time step of delay. In this Section, we are going to show that this delay is unavoidable.

Intuitively, the proof of the time lower bound uses the fact that if the network has to solve the problem without delay, the network must stabilize immediately at each time step. Therefore, the neurons that fire at the last round will stay firing. By injectivity of the representation, we can conclude that the network can at most count up to the network size.

The proof of Theorem 2.1.3 is the follows. The proof of Theorem 2.1.4 is identical.

Proof. Consider the following input sequence such that for all $0 \leq t < T$ we have $x^{(t)} = 1$ and for all $t \geq T$ we have $x^{(t)} = 0$. Let X be the collections of all neurons in the network. Assume for all $0 \leq t \leq T$, the network solves FCSC(t) at time 0. For all $0 \leq j \leq T$, let $S_j = \{y_i : y_i^{(j)} = 1, 1 \leq i \leq m\}$. We want to show that $S_T \supsetneq S_{T-1} \supsetneq \cdots \supsetneq S_0$. To prove this by induction on t, we strengthen our induction hypothesis to become $S_t \supsetneq S_{t-1} \supsetneq \cdots \supsetneq S_0$ and for all $y_j \in S_{t-1}$ we have $w_{xy_j} > 0$.
Base Case: When $t = 1$, notice that $S_0 = \emptyset$ by construction. Now by injectivity of the counter representation, we have $S_1 \supsetneq S_0$ and for $y_j \in S_0$, $w_{xy_j} > 0$ is vacuously true.
Induction Step: Now assume $S_t \supsetneq S_{t-1} \supsetneq \cdots \supsetneq S_1$ and $w_{xy_j} > 0$ for $y_j \in S_{t-1}$. At time step $t+1$, since the network solves FCSC(t) at time 0, the neurons in y is stabilized even without the input from x. This means that

$$\sum_{z \in X/\{x\}} w_{zy_j} z^{(t)} - b_{y_j} > 0 \text{ if } y_j \in S_t$$

Now since $w_{xy_j} > 0$, we know that neurons in S_{t-1} will keep firing at time $t+1$. For neurons in S_t/S_{t-1}, since those neurons fire at time t, we have

$$w_{xy_j} + \sum_{z \in X/\{x\}} w_{zy_j} z^{(t-1)} - b_{y_j} > 0 \text{ if } y_j \in S_t/S_{t-1}$$

And since the network solves FCSC(t-1) at time $t-1$, we also have

$$\sum_{z \in X/\{x\}} w_{zy_j} z^{(t-1)} - b_{y_j} \leq 0 \text{ if } y_j \in S_t/S_{t-1}$$

Substract two equations we get

$$w_{xy_j} > 0 \text{ if } y_j \in S_t / S_{t-1}$$

And hence $S_{t+1} \subset S_t$. By injectivity of the count representation, we have $S_{t+1} \supsetneq S_t$ as desired.

Now we have $S_T \supsetneq S_{T'-1} \supsetneq \cdots \supsetneq S_2 \supsetneq S_1$, but we only have less than T neurons. Contradiction. □

2.5 Discussion and Future Directions

In this work, we model how brains process temporal information over a long time range using neurons with transient activities. We propose two tasks that correspond to two common neural coding schemes, temporal coding and rate coding. "First consecutive spikes counting" (FCSC) is equivalent to counting the distance between the first two spikes, a prevalent temporal coding scheme in the sensory cortex while "Total spikes counting"(TSC) counts the number of the spikes over an arbitrary interval, which is an example of a rate coding. We design two networks with memoryless neurons that solve the above two problems in time 1 with $O(\log T)$ neurons and show that the time bound is tight.

A natural extension is to consider general temporal coding. Instead of coding the distance between the first two spikes, we code an arbitrary spike pattern within a max input interval length T. This can be done as an application of the FCSC network. Given an input pattern with K spikes, we can count the spike interval between each pair of spikes using an FCSC network and have a network that processes an arbitrary spike pattern with K spikes in time 1 with $O(K \log T)$ neurons. Since typically the temporal coding in the brain is sparse, we have $K \ll T$ and therefore the network only uses a small number of neurons.

Out of the spiking neural networks literature, Hitron and Parter [34] tackled a similar problem. Their deterministic neural counter problem is our TSC problem. This work differ in three ways. First, our network has time bound 1 while theirs is $O(\log T)$. Second, we provide a time lower bound result and show our time bound

is optimal. Third, they additionally consider an approximate version of the problem while we consider other forms of neural coding.

Our work follows similar approaches to Lynch et al. [54, 55, 53] by treating neurons as static circuits to explore the computational power of neural circuits. There are three noteworthy points about our model. First, instead of a stochastic model, we use a deterministic one. However, it should be noted that all the results in this work would still hold under the randomized model of Lynch et al. [54, 55, 53] with high probability. Second, we use a model that resets the potential at every round. Therefore, to retain temporal information, many self-excitation connections are employed in our networks. At the other extreme, we could have a model in which the potential does not decay from past rounds. In that model, temporal information can be stored in potentials, but it might require different mechanisms to translate the information from potentials to spikes. The two models thus could lead to different possible computational principles in brains. Third, we used a discrete time model instead of a continuous time model, which would be more biologically plausible. However, this might not be a concern since we could use Maass's synchronization module [58] to simulate our discrete time model from a continuous time model.

In addition, our networks are not noise tolerant, whereas the actual neuronal dynamics are highly noisy. It will be interesting to consider a noise tolerant version of the network. One possible formulation is the following: at each time step t, with probability τ which does not depend on the number of neurons, a spiking event becomes a non-spike event. Can the network still count exactly or approximately with high probability? Can we find a noise tolerant network that can do this with $O(\log T)$ neurons?

Another aspect of the temporal input we have not explored is the time-scale invariance of the problem. In biology, many problems are time-scale invariant. A person who says "apple" fast can be understood as well as a person who says "apple" slowly. If we exploit this invariance, we might be able to reduce the networks' complexity further.

Chapter 3

Plastic Neural Circuit: Oja's Rule and Sensory Adaptation

3.1 Introduction

One of the most influential theoretical ideas in neuroscience is Barlow's efficient coding principle for sensory systems [11]. Barlow hypothesized that the main goal of the sensory system is to reduce the redundancy in the sensory input and maximize the information transmitted to downstream brain areas. One of its key predictions is that the sensory neurons in the brain adapt to natural stimuli. Indeed, neuroscientists have shown in numerous sensory systems that by maximizing the mutual information transmitted on natural stimuli, one can recover the response filters in the respective sensory system. In the visual system, the structures of both the center-surround receptive field of the retina ganglion cells [6, 7, 29] and Gabor filters of V1 simple cells can be mathematically derived from the efficient coding principle [62]. In the auditory system, the temporal cochlear filters of inner ears can also be derived from optimizing mutual information on natural sounds [48]. However, most works on efficient coding of a sensory system have focused on optimizing the statistics of one natural environment. In reality, the environmental statistics can change drastically and the sensory system needs to continuously adapt to the changing environment in a matter of seconds while having high dimensional sensory inputs. For example,

although the retina processes visual inputs from 100 million photoreceptors to 1 million retina ganglion cells, it can change its receptive field to adapt to environments with different illumination [74], contrast [74, 9, 75], spatial frequency [75, 37], orientation and temporal correlation [37] in the time scale of seconds. Therefore, it is important to have a theoretical understanding of how the efficient coding principle can adapt to changing environments in a biologically realistic timescale with a biologically plausible synaptic learning rule. In this chapter, we give the first theoretical demonstration of sensory adaptation under the efficient coding principle in biologically realistic timescale through studying the convergence rate and behaviors of *Oja's rule* [59].

It is known that Oja's rule maximizes the mutual information under Gaussian inputs and linear networks by adapting to the direction that maximizes the variance of the presynaptic inputs through solving Principal Component Analysis (PCA) [50]. Therefore, studying its convergence rate and behaviors can shed light on fast sensory adaptation under the efficient coding principle. Since the dimensionality of the sensory inputs is usually large, for Oja's rule to behave in a biologically realistic time scale, the convergence rate needs to have no dependency or only log dependency on the input dimension. In addition to its relation to the efficient coding principle, as a biologically plausible synaptic modification rule, Oja's rule serves as a plasticity candidate to investigate sensory adaptation. Oja's rule is one of the earliest local learning rules that incorporate both *Hebbian* and *homeostatic plasticity* [59], two major activity-dependent synaptic modification mechanisms [1]. Both mechanisms work together to form memory and drive learning behaviors in the brain. Hebbian plasticity is a synapse-specific correlation-based plasticity mechanism that strengthens the connection when the input has a high correlation with the weights while weakening the connection when the input has a poor correlation [41, 22, 14]. However, this type of mechanism alone can often make networks unstable since the highly correlated input will keep strengthening synapses unboundedly [1]. Homeostatic plasticity, in contrast, stabilizes the network by keeping the activities of the neurons relatively constant through calcium sensors [80]. Synaptic scaling is a specific kind of home-

ostatic plasticity where the strength of the incoming synapses is normalized while still encoding the information from Hebbian learning in their relative strength after normalization [79]. It is thus an important problem in computational neuroscience to understand the interplay between Hebbian and homeostatic plasticity [78]. Oja's rule is one example of this. Concretely, Oja's rule can be expressed as the following

$$w_t = w_{t-1} + \eta_t(x_t y_t - y_t^2 w_{t-1})$$

where w_t is the strength of the synapse at time t, x_t, y_t are the firing rates of presynaptic, and postsynaptic neurons respectively, and η_t is the learning rate. One can see that $x_t y_t$ term corresponds to the Hebbian plasticity while $y_t^2 w_{t-1}$ term corresponds to the homeostatic plasticity. One can then show the synaptic scaling property where $\|w_t\| \approx 1$ for all t.

Despite being a subject of extensive theoretical [59, 61, 70, 33, 60, 68, 21, 88, 87, 23, 5] and experimental [16, 43, 31, 40, 18, 73, 77, 51, 5] studies aimed at understanding its performance, the theoretical understanding of the Oja's rule remains incomplete. The state-of-the-art theoretical analysis only provides a guarantee on convergence in the limit [23] through the Kushner-Clark methods [44]. However, to the best of our knowledge, there is no prior work showing the convergence time of Oja's rule. Specifically, if the convergence time of Oja's rule does not depend on the input dimension or depend on only logarithmic factors of the input dimension, Oja's rule can serve as an example of sensory adaptation under the efficient coding principle in a biologically realistic time scale.

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

Theorem 3.1.1 (informal)**.** *Biological Oja's rule efficiently solves streaming PCA with (nearly) optimal convergence rate. Specifically, the convergence rate we obtain matches the information-theoretic 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 dependency on logarithmic factors of the dimension when the initial vector is random. Therefore, biological Oja's

rule solves streaming PCA on a biologically realistic time scale.

Also, we show for-all-time convergence with a slowly diminishing learning rate. Most convergence results in the literature show that

$$\Pr(\text{error at time T} > \epsilon) < \delta.$$

However, this is not enough in a biological system. The sensory system cannot afford to only be functional at time T. It needs to be functional constantly. In contrast, the convergence result we can show is

$$\Pr(\exists t \geq T, \text{ error at time t} > \epsilon) < \delta,$$

which guarantees the convergence at all time. Furthermore, in order to achieve this, our learning rate η_t only needs to be scaled as $\eta_t = O(\frac{1}{\log t})$, in particular $\sum_t \eta_t^2 = \infty$. In contrast, the Kushner-Clark theorem requires $\sum_t \eta_t^2 < \infty$ where the learning rate is commonly set as $\eta_t = O(\frac{1}{t})$. 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.

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

We organize the rest of the introduction as follows. We first formally define biological Oja's rule and streaming PCA in Section 3.1.1 and state the main results

and their biological relevance in Section 3.1.2. In Section 3.1.3, we provide a technical overview of the proof and the analysis framework. Finally, we conclude the introduction with a survey and comparison of related works in Section 3.1.4.

3.1.1 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* [82, 83] and since biological Oja's rule was defined on a rate-based model, this setting will be the focus of this work.

To understand how 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, which has been informally paraphrased as "cells that fire together wire together" [32]. 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. biologi-

cal 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 [59] 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) [65, 38] 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 [19] for a survey on the literature), the input data arrives in a stream and the algorithm/dynamic only has a limited amount of space, *e.g.,* $O(n)$ space. Streaming PCA is important for biological systems 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 [3]). In the following, we formally define the streaming PCA problem. [1]

Problem 3.1.2 (Streaming PCA)**.** *Let $n, T \in \mathbb{N}$ and $\mathcal{D}$ be a distribution over the unit sphere of $\mathbb{R}^n$. Suppose the input data $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T \overset{i.i.d.}{\sim} \mathcal{D}$ 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

[1] In related works, some (*e.g.,* [3]) measure the error using $1 - \langle \mathbf{w}, \mathbf{v}_1\rangle^2$, some (*e.g.,* [72]) use $1 - \mathbf{w}^\top A\mathbf{w}/\|A\|$, and some (*e.g.,* [39]) 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 [72]. 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.

T.

Biological Oja's rule in solving streaming PCA Oja [59] 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 firing rate vector $\mathbf{x}_t$ and the synaptic weight vector $\mathbf{w}_t$ correspond to the $\mathbf{x}_t, \mathbf{w}$ in the streaming PCA problem in Problem 3.1.2.

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 biological Oja's rule, the dynamic of the synaptic weight vector is described by the following equation.

Definition 3.1.3 (Biological Oja's rule)**.** *For any initial vector $\mathbf{w}_0 \in \mathbb{R}^n$ such that $\|\mathbf{w}_0\|_2 = 1$, the dynamic of biological Oja's rule is the following. For any $t \in \mathbb{N}$, define*

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \eta_t y_t \left(\mathbf{x}_t - y_t \mathbf{w}_{t-1}\right) \tag{3.1.4}$$

where $y_t = \mathbf{x}_t^\top \mathbf{w}_{t-1}$ and $\mathbf{x}_t$ is the input at time t. See also in Figure 3-1 for a pictorial definition of biological Oja's rule in solving streaming PCA.

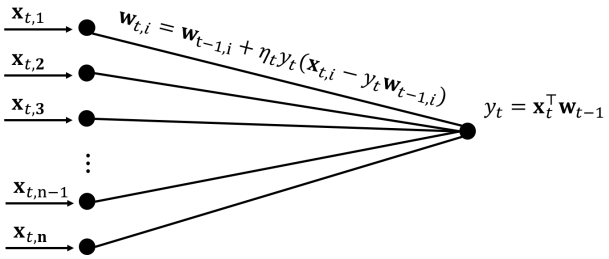


Figure 3-1: A neural network that uses biological Oja's rule to solve streaming PCA. The firing rate vector $\mathbf{x}_t$ is the input and the weight vector $\mathbf{w}_t$ is the output at time t.

Following from the definition, biological Oja's rule is *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 3.5.1), biological Oja's rule achieves the *synaptic scaling guarantee* [1], *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 biological Oja's rule in the retina-optical nerve pathway. See Section 3.1.2 for more discussions.

Oja's derivation for biological Oja's rule Before going into more technical contents, it would be helpful to take a look at the original derivation for biological Oja's rule. Initially, Oja wanted to use the following update rule with normalization[2] to solve the streaming PCA problem.

$$\mathbf{w}_t = \frac{\left(I + \eta_t \mathbf{x}_t \mathbf{x}_t^\top\right) \mathbf{w}_{t-1}}{\| \left(I + \eta_t \mathbf{x}_t \mathbf{x}_t^\top\right) \mathbf{w}_{t-1} \|_2} \,. \tag{3.1.5}$$

However, the normalization term $\| \left(I + \eta_t \mathbf{x}_t^\top \mathbf{x}_t\right) \mathbf{w}_{t-1} \|_2^{-1}$ is *global*[3] 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 biological Oja's rule (*i.e.,* Equation 3.1.4). See Section A.1 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 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.

[2]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 Section 3.1.4 for more discussion.

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

Previous works: Results about convergence in the limit There were many previous works on analyzing the convergence of biological Oja's rule in solving streaming PCA [59, 61, 70, 33, 60, 68, 21, 88, 87, 23]. However, these works only proved guarantees on convergence in the limit. For example, Duflo [23] showed that $\mathbf{w}_t$ converges to the top eigenvector of A in the limit under some constraints on the learning rates.

Theorem 3.1.6 ([23], 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 a convergence rate guarantee because they only provide convergence in the limit.

To the best of our knowledge, prior to this work, there had been no efficiency guarantee for biological Oja's rule. The main technical barrier is due to the nonlinear terms in the update rule which introduces correlations in the traditional step-by-step analysis and thus naive analysis would not work. We explain the difficulty further in Section 3.1.3 and Section A.3. Given this situation, natural questions on the frontier would then be:

> **Question**: What is the convergence rate of biological Oja's rule in solving streaming PCA? Is the convergence rate fast enough to be an example of fast sensory adaptation under the efficient coding principle with high dimensional sensory inputs?

3.1.2 Our results

In this chapter, we answer the above questions by giving the first convergence rate guarantee for 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 [3]. In terms of the techniques, we develop an ODE-inspired framework to analyze stochastic dynamics. We believe this general framework of using tools and insights from ODE and SDE in analyzing stochastic dynamics is elegant

and powerful. We provide more details and intuitions on the ODE-inspired framework in the section on the technical overview (see Section 3.1.3). 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 3.1.7 (Global and local convergence)**.** *With the setting in Problem 3.1.2 and dynamic in Definition 3.1.3, 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*

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

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*

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

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* $\tilde{\Theta}$ *hides the poly-logrithmic factors with respect to* $\epsilon^{-1}, \delta^{-1}, \mathsf{gap}^{-1}, n$*.*

Biological perspectives Our results use biological Oja's rule as an example to provide the first theoretical demonstration of how a sensory system can adapt to changing environments in a biologically realistic time scale under the efficient coding principle. In a linear network with Gaussian input, biological Oja's rule maximizes the mutual

information of the presynaptic inputs by selecting the top principal component as the output. Specifically, we show that *biological Oja's rule is a local synaptic modification mechanism that not only incorporates both Hebbian learning and homeostatic plasticity but also can solve streaming PCA 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 instead of $O(1/\delta)$. The local convergence is particularly important for sensory adaptation because different environments in nature are usually still correlated and therefore the sensory system does not need to start at a random initialization to adapt to the new environment. Since the local convergence does not depend on the dimension of the inputs, this demonstrates that it is possible for a sensory system with high dimensional sensory inputs to adapt to changing environments in seconds.

Furthermore, we prove the *for-all-time* guarantee of biological Oja's rule as a 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. In particular, the dynamic would not leave even temporarily 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 3.1.8 (For-all-time guarantee with slowly diminishing rate)**.** *With the setting in Problem 3.1.2 and dynamic in Definition 3.1.3, 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 3.1.8). This suggests the capability of *continual adaptation*, which is crucial in the biological scenario. For example, if a person walks into a new environment, the retina cells need to quickly adapt to the new environment and this cannot be achieved if the learning rate already diminished too fast in the previous environment. Since our learning rate decreases like $\Omega(1/\log t)$, when the environment changes, the learning rate is still large enough to do efficient adaptation without resetting the learning rate.

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 $\eta_t = o(1/\sqrt{t})$ [23] and the convergence rate analysis for non-biologically-plausible variants of Oja's rule requires $\eta_t = \tilde{O}(1/t)$ [39, 3, 49] or $\eta_t = O(1/\sqrt{t})$ [72]. 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 $\eta_t = \Omega(1/\log t)$ (hence $\sum_t \eta_t^2 = \infty$) for biological Oja's rule.

3.1.3 Technical overview

In this work, we give the first efficiency guarantee for 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 clear understanding of how biological Oja's rule solves streaming PCA. In short, our high-level strategy is to first consider the *continuous* version of 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., $A = \mathsf{diag}(\lambda_1 \ldots, \lambda_n)$ 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 Section 3.5.1.

Insight 1: Inspiration from the continuous dynamics The first insight is to analyze biological Oja's rule in a way inspired by its continuous analog. The advantage of considering the continuous dynamics is that not only does it capture 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 biological Oja's rule degenerates into a simple deterministic ODE almost surely (see Section 3.3 for a derivation). Specifically, for any $t \geq 0$, we have

$$\frac{d\mathbf{w}_{t,1}}{dt} \geq (\lambda_1 - \lambda_2)\mathbf{w}_{t,1}(1 - \mathbf{w}_{t,1}^2) \quad \text{and} \quad \|\mathbf{w}_t\|_2 = 1 \tag{3.1.9}$$

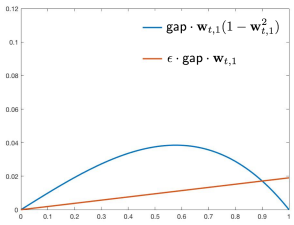
almost surely. Furthermore, observe that the continuous Oja's rule is non-decreasing and has three fixed points 0 and ± 1 for $\mathbf{w}_{t,1}$ while the first is unstable and the later two are stable. Namely, in the continuous dynamic, $\mathbf{w}_t$ will eventually converge to $\pm\mathbf{e}_1$, *i.e.,*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

Note that in a discrete stochastic dynamic, there are two sources of noise: (i) the intrinsic stochasticity from its continuous analog and (ii) the noise due to discretization. Thus, Equation 3.1.9 suggests that the noise in biological Oja's rule comes only from discretization since the continuous Oja's rule is deterministic.

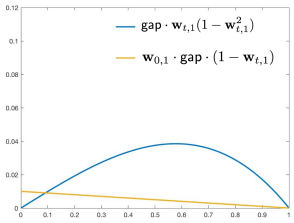
In addition to the limiting behavior, one can also read out finer structures of the continuous dynamic from Equation 3.1.9 by solving the differential equation using standard tools from dynamical system. The right hand side (RHS) of the inequality in Equation 3.1.9 is non-linear which usually does not have a clean solution. A natural idea from dynamical system would then be *linearizing* the differential equation around fixed points and applying the *exact* solution for a linear ordinary differential equation. Moreover, as there are three fixed points in Equation 3.1.9, one can linearize the

differential equation with the 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.

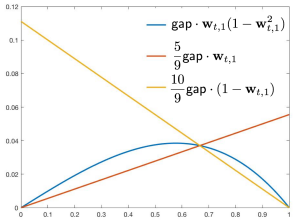
For example, we can linearize at 0 by lower bounding the RHS of Equation 3.1.9 by $\epsilon(\lambda_1 - \lambda_2)\mathbf{w}_{t,1}$ for any $\mathbf{w}_{t,1} \in [0, \sqrt{1-\epsilon}]$ (see Figure 3-2a). Similarly, we can linearize at 1 by using $\mathbf{w}_{0,1}(\lambda_1 - \lambda_2)(1 - \mathbf{w}_{t,1})$ for any $\mathbf{w}_{t,1} \in [\mathbf{w}_{0,1}, 1]$ (see Figure 3-2b). 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 3-2c).



(a) Linearization only at 0.



(b) Linearization only at 1.



(c) Linearization at both 0 and 1.

Figure 3-2: In (a), we only linearize at 0 and use $\epsilon \cdot \mathsf{gap} \cdot \mathbf{w}_{t,1}$ to lower bound Equation 3.1.9 for $\mathbf{w}_{t,1} \in [0, \sqrt{1-\epsilon}]$. In (b), we only linearize at 1 and use $(\mathbf{w}_{0,1} \cdot \mathsf{gap} \cdot (1 - \mathbf{w}_{t,1}))$ for $\mathbf{w}_{t,1} \in [\mathbf{w}_{0,1}, 1]$. On the other hand, in (c), we linearize at both 0 and 1. For $\mathbf{w}_{t,1} \in [0, \frac{2}{3}]$, we use $\frac{5}{9}\mathsf{gap} \cdot \mathbf{w}_{t,1}$ while for $\mathbf{w}_{t,1} \in [\frac{2}{3}, 1]$ we use $\frac{10}{9}\mathsf{gap} \cdot (1 - \mathbf{w}_{t,1})$. One can see that the lower bounds in (c) are much tighter than that in (a) and (b) in the sense that the *slopes* are of order $\Omega(\mathsf{gap})$ instead of $O(\epsilon \cdot \mathsf{gap})$ or $O(\mathbf{w}_{0,1} \cdot \mathsf{gap})$.

The main difference between linearizing only at a single fixed point and linearizing at two fixed points is the *slope* in the linearization. Note that the slopes of the linearizations in Figure 3-2a and Figure 3-2b 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 3-2c. 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 3-2 for a pictorial explanation.

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

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

for some constant a and function $b(t)$. To put into the context, one can think of a as the drifting term and $b(t)$ as the noise term in the continuous Oja's rule due to the linearization[4]. By the standard tool for solving linear ODE, the solution of $y(t)$ at $t = T$ is

$$y(T) = e^{aT} \cdot \left(y(0) + \int_0^T e^{-at} b(t) dt \right) . \tag{3.1.10}$$

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

To sum up, the continuous dynamic informs us how to linearize 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 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.,* [3]).

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

Traditional step-by-step analysis To see the difference, let us illustrate how the step-by-step analysis on biological Oja's rule would work. 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 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 an 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 [3] encountered in their analysis for a variant of biological Oja's rule. They resolved the first difficulty by decomposing the non-linear term in the dynamic into a *multi-dimensional chain* and carefully bounding the chain with strong assumptions on learning rates to enable martingale analysis. They used extremely delicate and complicated techniques tailored to the dynamic to achieve optimal convergence rate. biological Oja's rule, in addition to having the first difficulty, also has the second difficulty (see Section A.3 for more discussions). Therefore, applying the traditional step-by-step analysis of 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 a 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 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[5], we have

$$\mathbf{w}_{T,1} = H^T \cdot \left(\mathbf{w}_{0,1} + \sum_{t=1}^{T} \frac{N_t}{H^t} \right) \tag{3.1.11}$$

where $H > 1$ is the multiplier term and $\{N_t\}$ is the noise term which forms a martingale on the natural filtration. See Corollary 3.7.21 and Corollary 3.6.4 for a precise formulation of H and $\{N_t\}$ in our analysis. Intuitively, one should think of $H^T \mathbf{w}_{0,1}$ as the *drifting term* and the other part as the *noise term*. The goal of the ODE trick in the discrete dynamic is to show that the drifting term dominates the noise term.

To show that the noise in Equation 3.1.11 is small, Azuma's inequality would be a natural tool to start with (see Lemma 3.2.3). 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[6] by $1 + o(1)$. In the meantime, both $\mathbf{w}_{t,1}^2$ and the noise are expected to be very small in the early stage of the dynamic with high probability.

To circumvent this obstacle, we consider the *stopped process* of the original martingale in which the bounded difference is under control. For example, consider the above situation where the noise term $\{N_t\}$ is a martingale and a stopping time τ for the event $\{\mathbf{w}_{\tau,1}^2 \geq 0.1\}$. The stopped process, denoted by $\{N_{t \wedge \tau}\}$ where $t \wedge \tau = \min\{t, \tau\}$,

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

[6] This is because we are able to upper bound $\mathbf{w}_{t-1,1}$ by $1 + o(1)$ almost surely. See Section 3.5.2. Note that there are ways to get better bounded difference conditions in the worst case but this is still not sufficient.

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 3.1.11: we have to show that the concentration for the stopped process $\{N_{t\wedge\tau}\}$ can be extended to the original process $\{N_t\}$. We achieve this task by developing a *pull-out lemma* which is able to utilize the structure of the martingale and pull out the stopping time from a concentration inequality.

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

Let $M_t = \sum_{t'=1}^{t} H^{-t'} N_{t'}$ be the noise term in the ODE trick (*i.e.,* Equation 3.1.11) and τ be a stopping time that ensures good bounded difference condition. Note that as $\{N_t\}$ is a martingale, we know that $\{M_{t\wedge\tau}\}$ is also a martingale. There are two key ingredients to pull out the stopping time from $\{M_{t\wedge\tau}\}$, *i.e.,* the stopped process of the noise term.

First, we use the *maximal* concentration inequality (*e.g.,* Lemma 3.2.4) which gives the following stronger guarantee than the traditional Azuma's inequality.

$$\Pr\left[\sup_{1\leq t\leq T} |M_{t\wedge\tau} - M_0| \geq a\right] < \delta \tag{3.1.12}$$

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

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

$$\Pr\left[\tau \geq t+1 \;\middle|\; \sup_{1\leq t'\leq t} |M_{t'} - M_0| < a\right] = 1\,. \tag{3.1.13}$$

Namely, if the bad event has not happened, then the martingale would not stop immediately. Intuitively, Equation 3.1.13 holds because $\{\sup_{1\le t'\le t}|M_{t'}-M_0|<a\}$ implies the noise term to be small and thus the drifting term dominates in the ODE trick. 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 $\tau \ge t+1$.

Combining the above two ingredients (*i.e.,* Equation 3.1.12 and Equation 3.1.13), we are able to show in the pull-out lemma that

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

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 $\{|M_{\tau'}-M_0|\ge a\}$ 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 3.1.13 to show that the probability of error event is 0 based on a *diagonal argument*. Thus, we have

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

See Section 3.6.2 and Figure 3-3 for more details on the chain condition for biological Oja's rule and how to partition the probability space of the error event.

3.1.4 Related works

Related theory work on biological Oja's variants Computational neuroscientists have proposed several variants of biological Oja's rule to solve streaming

PCA [59, 60, 70, 26, 46, 69, 43, 67]. In a single neuron case, Oja used stochastic approximation theory [44] to prove the global convergence in the limit [59]. In a multi-neurons case, Hornik and Kuan demonstrated the connection between the discrete dynamics and the associated ODE [36] from the Kushner-Clark theory [44]. However, most existing analyses on the multi-neurons dynamics show only local convergence [70, 26, 46, 69, 43, 67]. Even for the ODE dynamic, the global convergence for most networks in a multi-neurons case is difficult to show. Yan et al. provide the only global analysis on Oja's multi-neurons subspace network [60, 86, 85]. Previously there is no work showing the convergence rate on the discrete dynamics. This thesis shows the first convergence rate bound on biological Oja's rule.

Oja's rule in machine learning Unlike the situation in biological Oja's rule, a line of recent exciting results [30, 20, 10, 72, 39, 3] 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 biological Oja's rule.

To see the difference between biological Oja's rule and ML Oja's rules, let us take the update rule from [72, 39, 3] as an example. Note that the other variants of ML Oja's rules also have a similar fundamental difference from 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 3.1.5, *i.e.,* applying Taylor's expansion on the ML Oja's rule and truncating the higher-order terms would result in biological Oja's rule.

A natural idea would be trying to *couple* biological Oja's rule with the ML Oja's rule by showing that for every $t \in \mathbb{N}$, 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 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 biological Oja's rule requires two linear ODEs in different regimes for tight analysis. See Section 3.3 and Section A.3 for more details.

To sum up, biological Oja's rule and the ML Oja's rule are similar but the analysis of the latter 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 biological Oja's rule, in this work we develop a completely different framework (as briefly discussed in Section 3.1.3). This framework not only gives the first and nearly optimal convergence rate guarantee for 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 future work.

Comparing with other streaming PCA algorithms Streaming PCA is a well-studied and challenging computational problem. Many works [20, 72, 49, 39, 3] provided theoretical guarantees for streaming PCA algorithms. Interestingly, all of the streaming PCA algorithms in these works are some variants of 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 [3]. Prior to this work, the state-of-the-art for both global and local convergences are achieved by [3] using ML Oja's rule (see the second to last row of Table 3.1). In this work, as a byproduct, the convergence rate we get for biological Oja's rule outperforms [3] by a logarithmic factor in both settings. See Table 3.1 for a summary.

*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 [20] and [72] cannot handle arbitrary failure probability so we ignore their δ dependency on the table.

‡In [20, 72, 49], their convergence rates are far from the information-theoretic lower bound. So we do not trace down their logarithmic dependencies.

§In [3], they only stated $\Omega(\frac{\lambda_1}{\mathsf{gap}} \cdot \frac{1}{\epsilon})$ lower bound. We observe that their lower bound can be

Algorithm	Reference	Any Input	Global Convergence		Local Convergence	
			Convergence Rate	Degree in Log Terms*	Convergence Rate	Degree in Log Terms*
Biological Oja's Rule	This Work	Y	$\tilde{O}\left(\frac{\lambda_1}{\mathsf{gap}^2}\cdot\frac{1}{\epsilon\wedge\delta^2}\right)$	3	$\tilde{O}\left(\frac{\lambda_1}{\mathsf{gap}^2}\cdot\frac{1}{\epsilon}\right)$	2
ML Oja's Rule	[20]	N	$\tilde{O}\left(\frac{n}{\mathsf{gap}^2}\cdot\frac{1}{\epsilon}\right)$ †	- ‡	$\tilde{O}\left(\frac{n}{\mathsf{gap}^2}\cdot\frac{1}{\epsilon}\right)$ †	- ‡
	[72]	Y	$\tilde{O}\left(\frac{n}{\mathsf{gap}^2}\cdot\frac{1}{\epsilon}\right)$ †	- ‡	$\tilde{O}\left(\frac{n}{\mathsf{gap}^2}\cdot\frac{1}{\epsilon}\right)$ †	- ‡
	[49]	N	$\tilde{O}\left(\frac{\lambda_1 n}{\mathsf{gap}^2}\cdot\frac{1}{\epsilon\delta^6}\right)$	- ‡	$\tilde{O}\left(\frac{\lambda_1 n}{\mathsf{gap}^2}\cdot\frac{1}{\epsilon\delta^4}\right)$	- ‡
	[39]	Y	$\tilde{O}\left(\frac{\lambda_1}{\mathsf{gap}^2}\cdot\frac{1}{\epsilon\delta^3}\right)$	2	$\tilde{O}\left(\frac{\lambda_1}{\mathsf{gap}^2}\cdot\frac{1}{\epsilon\delta^3}\right)$	2
	[3]	Y	$\tilde{O}\left(\frac{\lambda_1}{\mathsf{gap}^2}\cdot\frac{1}{\epsilon\wedge\delta^2}\right)$	≥ 4	$\tilde{O}\left(\frac{\lambda_1}{\mathsf{gap}^2}\cdot\frac{1}{\epsilon}\right)$	≥ 3
Any Algorithm	[3]	$\Omega\left(\frac{\lambda_1}{\mathsf{gap}^2}\cdot\frac{\log\frac{1}{\delta}}{\epsilon}\right)$ §				

Table 3.1: Convergence rate for biological Oja's rule and ML Oja's rule in solving streaming PCA. The "Any Input" column indicates that whether the analysis has higher moment conditions on the unknown distribution $\mathcal{D}$. Note that having higher moment conditions would drastically simplify the problem because the non-linear terms in the update rule can then be non-trivially replaced with the first order term.

Algorithms inspired by biological neural networks In recent years, the study of the algorithmic aspect of mathematical models for biological neural networks is an emerging field in theoretical CS. For example, the efficiency of spiking neural networks in solving the *winner-take-all (WTA)* problem [56, 54, 55, 53, 76], the efficiency of spiking neural networks in storing temporal information [57, 34], assemblies [47, 64], spiking neural networks in solving optimization problems [17, 66] and learning hierarchically structured concepts [52]. Under this context, this work provides an algorithmic insight in a biologically-plausible learning rule that solves streaming PCA.

3.2 Preliminaries

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

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.

3.2.1 Notations

We use $\mathbb{N} = \{1, 2, \dots\}$ and $\mathbb{N}_{\geq 0} = \{0, 1, \dots\}$. For each $n \in \mathbb{N}$, denote $[n] = \{1, 2, \dots, n\}$ and $[n]_{\geq 0} = \{0, 1, \dots, n\}$. For a vector indexed by time t, *e.g.,* $\mathbf{w}_t$, its i^{th} coordinate is denoted by $\mathbf{w}_{t,i}$. The notation $\tilde{O}$ (similarly, $\tilde{\Omega}$ and $\tilde{\Theta}$) is the same as the big-O notation by ignoring extra poly-logarithmic term. $\mathbf{1}_E$ stands for the indicator function for a probability event E. We sometimes abuse the big O notation by $y = O(x)$ meaning $|y| = O(x)$ and this will be clear in the context. Throughout the chapter, 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.

3.2.2 Probability toolbox

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

Definition 3.2.1 (Adapted random process)**.** *Let $\{X_t\}_{t\in\mathbb{N}_{\geq 0}}$ be a sequence of random variables and $\{\mathcal{F}_t\}_{t\in\mathbb{N}_{\geq 0}}$ be a filtration. We say $\{X_t\}_{t\in\mathbb{N}_{\geq 0}}$ is an adapted random process with respect to $\{\mathcal{F}_t\}_{t\in\mathbb{N}_{\geq 0}}$ if for each $t \in \mathbb{N}_{\geq 0}$, the σ-algebra generated by $X_0, X_1, \dots, X_t$ is contained in $\mathcal{F}_t$.*

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

Definition 3.2.2 (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 3.2.3 (Azuma's inequality [8])**.** *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 3.2.4 (Maximal Azuma's inequality [28, 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 3.2.5 (Freedman's inequality [25])**.** *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 3.2.6. *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 3.2.7 (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 chapter, 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 [81, Theorem 10.9] for a proof for this classic result.

Brownian motion In Section 3.3, 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 3.2.8 (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 3.2.9 (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 3.2.9 here and refer the interested readers to standard textbook such as [45] for more details on Brownian motion.

3.2.3 ODE toolbox

Lemma 3.2.10 (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*

$$X_t = H_t X_{t-1} + A_t \tag{3.2.11}$$

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 3.2.10. For each $t_0 < i \leq t$, dividing Equation 3.2.11 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 3.2.12 (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*

$$X_t = H_t X_{t-1} + A_t \tag{3.2.13}$$

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 3.2.12. The proof is a direct induction. □

3.2.4 Approximation toolbox

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

Lemma 3.2.14. *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 3.2.15. *For any $x \in (0, 0.5)$ and $t \in \mathbb{N}$,*

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

Lemma 3.2.16. *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. □

3.3 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. Specifically, in Section 3.3.1, we show a surprising fact, the randomness in the input disappears at the continuous limit. Therefore, the continuous limit of Oja's rule is deterministic almost surely. In Section 3.3.2, we formalize the key insights 1 in Section 3.1.3 by proving that one needs to linearize the continuous dynamic at 0 first and then at 1 to obtain a tight analysis. This provides three insights for analyzing the discrete dynamic. First, it suggests that one should linearize at 0 at 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 an exact characterization of the dynamic would give a tight analysis. Finally, the continuous dynamic is deterministic and will stay around the optimal region for all time after a certain point. This suggests that the *for-all-time* guarantee could potentially happen in the original discrete setting.

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 can 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.

3.3.1 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 Section 3.5.1. In this section, we will show that the continuous dynamic of Oja's rule is actually deterministic almost surely. Specifically, we will derive the

following ODE as the continuous dynamic of Oja's rule

$$d\mathbf{w}_t = \left[\mathsf{diag}(\lambda)\mathbf{w}_t - \mathbf{w}_t^\top \mathsf{diag}(\lambda)\mathbf{w}_t\mathbf{w}_t\right] dt\,. \tag{3.3.1}$$

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,

$$\frac{1}{\Delta}\mathbb{E}\left[\left(B_{t+\Delta} - B_t\right)\left(B_{t+\Delta} - B_t\right)^\top\right] = \mathsf{diag}(\lambda)\,. \tag{3.3.2}$$

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[7] with covariance matrix being A. To put this into the context of biological Oja's rule, one should think of $\eta = \Delta$, $\mathbf{x}_j = \frac{1}{\sqrt{\Delta}}\Delta B_j$, and $y_j = \mathbf{x}_j^\top \mathbf{w}_{j-1}$ for each $j \in \mathbb{N}$ where $\Delta B_j = \left(B_{j\cdot\Delta} - B_{(j-1)\cdot\Delta}\right)$ [8]. 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 $B_{t+\Delta} - B_t$ with dB_t and index the weight vector by $t \geq 0$, *i.e.,* $(\mathbf{w}_t)_{\geq 0}$. We then obtain the following SDE as the continuous Oja's rule dynamic.

$$d\mathbf{w}_t = dB_t^\top \mathbf{w}_t dB_t - \left(dB_t^\top \mathbf{w}_t\right)^2 \mathbf{w}_t\,. \tag{3.3.3}$$

It might look absurd at first glance (for those who have not seen stochastic calculus

[7]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 biological Oja's rule.

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

before) that there is a quadratic term of dB_t in Equation 3.3.3. Nevertheless, it is in fact mathematically well-defined and we recommend a standard resource such as [45] 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 3.2.9). 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 randomness from the input disappears, and the continuous Oja's rule defined in Equation 3.3.3 can be rewritten as Equation 3.3.1, a *deterministic* process, almost surely.

$$d\mathbf{w}_t = \left[\mathsf{diag}(\lambda)\mathbf{w}_t - \mathbf{w}_t^\top \mathsf{diag}(\lambda)\mathbf{w}_t\mathbf{w}_t\right] dt\,. \tag{3.3.1}$$

With the continuous Oja's rule being deterministic as in Equation 3.3.1, it is then not difficult to have a tight analysis on its convergence using tools from ODE as explained in the next section.

3.3.2 One-sided versus two-sided linearization

In this subsection, we analyze Equation 3.3.1 by linearizing the dynamic at 0 and 1 respectively and get two incomparable convergence rates (Theorem 3.3.4 and Theorem 3.3.5).

Theorem 3.3.4 (Linearization at 0)**.** *Suppose* $\mathbf{w}_{0,1} > 0$*. For any* $\epsilon \in (0,1)$*, when* $t \geq \Omega\left(\frac{\log(1/\mathbf{w}_{0,1}^2)}{\epsilon(\lambda_1-\lambda_2)}\right)$*, we have* $\mathbf{w}_{t,1}^2 > 1-\epsilon$*.*

Theorem 3.3.5 (Linearization at 1)**.** *Suppose* $\mathbf{w}_{0,1} > 0$*. For any* $\epsilon \in (0,1)$*, when* $t \geq \Omega\left(\frac{\log(1/\epsilon)}{\mathbf{w}_{0,1}(\lambda_1-\lambda_2)}\right)$*, we have* $\mathbf{w}_{t,1}^2 > 1-\epsilon$*.*

The proofs for Theorem 3.3.4 and Theorem 3.3.5 are based on applying Taylor's expansion on Equation 3.3.1 with a center either being 0 or 1. Then, we approximate the dynamics with linear differential equations and use tools from ODE to get a tight

analysis. See Section A.2 for the analysis on the linearizations of continuous Oja's rule.

When starting with a random vector, *i.e.,* $\mathbf{w}_{0,1} = \Omega(1/\sqrt{n})$ with high probability, the above convergence rates become $O\left(\frac{\log n}{\epsilon(\lambda_1-\lambda_2)}\right)$ and $O\left(\frac{\sqrt{n}\log(1/\epsilon)}{\lambda_1-\lambda_2}\right)$ respectively. This indicates that linearizing only on one side (either at 0 or at 1) would not give tight analysis. Nevertheless, if we invoke Theorem 3.3.4 with the error parameter being 0.5, then for some $t_1 = O\left(\frac{\log n}{\lambda_1-\lambda_2}\right)$, we have $\mathbf{w}_{t_1,1} > 0.5$. Next, we invoke Theorem 3.3.5 starting from $\mathbf{w}_{t_1}$ and with the error parameter being ϵ, then for some $t_2 = O\left(\frac{\log(1/\epsilon)}{\lambda_1-\lambda_2}\right)$, we have $\mathbf{w}_{t_1+t_2,1} > 1-\epsilon$. Putting these together, we have the following theorem combining the linearizations on both sides.

Theorem 3.3.6 (Linearization at both 0 and 1)**.** *Suppose* $\mathbf{w}_{0,1} > 0$*. For any* $\epsilon \in (0,1)$*, when*

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

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

The above theorem for the convergence rate of the continuous Oja's rule gives three key insights. First, it suggests that one should linearize at 0 at 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 an exact characterization of the dynamic would give a tight analysis. Finally, the continuous dynamic is deterministic and will stay around the optimal region for all time after a certain point. This suggests that the *for-all-time* guarantee could potentially happen in the original discrete setting.

3.4 Main Results

In this section, we state the main technical results of this chapter. In the following, all of the theorems and lemmas are stated with respect to the setting of streaming PCA defined as Problem 3.1.2 and discrete biological Oja's rule defined as Definition 3.1.3. Thus, for simplicity, we would not repeat the setup in their statements. Now, let us state the formal version of the main theorem for biological Oja's rule.

In Theorem 3.4.1, 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 3.4.2 we show that once $\mathbf{w}_t$ becomes ϵ-close to the top eigenvector $\mathbf{v}_1$, it will stay in the neighborhood of $\mathbf{v}_1$ for a long time 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 3.4.1 (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 3.4.1 To prove Theorem 3.4.1, we first reduce the general setting where the covariance matrix A is PSD to the special case where $A = \mathsf{diag}(\lambda)$ in Section 3.5. 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 Section 3.6. 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 Section 3.7.

Theorem 3.4.2 (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

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

Proof structure of Theorem 3.4.2 We first reduce the general setting to the special case where $A = \mathsf{diag}(\lambda)$ in Section 3.5. 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 Section 3.6.4.

3.5 Preprocessing

Before the main analysis of biological Oja's rule, we provide two useful observations on the dynamic in this section. Specifically, we show in Section 3.5.1 that considering

the covariance matrix being *diagonal* is sufficient for the analysis and in Section 3.5.2 that $\|\mathbf{w}_t\|_2^2 = 1 \pm O(\eta)$ almost surely for all $t \in \mathbb{N}$.

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

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

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

3.5.2 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 3.5.1. *For any $\eta \in (0, 0.1)$, if for all $t \in \mathbb{N}$, $\eta_t \leq \eta$, then for all $t \in \mathbb{N}_{\geq 0}$, $1 - 10\eta \leq \|\mathbf{w}_t\|_2^2 \leq 1 + 10\eta$ almost surely.*

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

3.6 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 $\mathbf{w}_{0,1}^2 \ge \frac{2}{3}$. The goal of this section is to show that $1 - \mathbf{w}_{t,1}^2 \le \epsilon$ for some $t = O\left(\frac{\lambda_1 \log(1/\epsilon)(\log\log\log(1/\epsilon) + \log(1/\delta))}{\epsilon(\lambda_1 - \lambda_2)^2}\right)$ for any small ϵ. Let us first state the main theorem of this section as follows.

Theorem 3.6.1. *Suppose* $\mathbf{w}_{0,1}^2 \ge 2/3$. *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* $1 - \delta$.

By applying the diagonal reduction argument in Section 3.5.1, we prove the local convergence part of Theorem 3.4.1 as a corollary. The proof structure of Theorem 3.6.1 is as follows. First, in Section 3.6.1 we derive a linearization of the dynamic using a center at 1 instead of 0 based on intuition from the continuous dynamic in Section 3.3. Furthermore, we use the ODE trick to write down the dynamic in closed form with respect to the linearization.

Next, in Section 3.6.2, 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.,* $\tilde{\mathbf{w}}_{t,1} < -2/3$) and thus the bounded difference might become too large to bound the noise effectively with Freedman's inequality. To deal with this issue, we consider a stopping time where $\tilde{\mathbf{w}}_{t,1} < -a$ to give good control on the bounded difference and subsequently bound the stopped process in Lemma 3.6.8. After we show that the stopped process

is small, we want to pull out the stopping time from the stopped process to show the concentration on the original process. In general, pulling out the stopping time is impossible without introducing extra failure probability; however, by exploiting the structure of the dynamics, we are able to pull out the stopping time without additional cost in Lemma 3.6.9.

Finally in Section 3.6.3, by combining the small noise and the ODE trick, we are able to prove Theorem 3.6.1 with an interval analysis. As a corollary of Lemma 3.6.7 in the local convergence, we show that biological Oja's rule has the continual learning capacity in Section 3.6.4. 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 3.6.12 guarantees Oja's rule can maintain the convergence for any finite time length efficiently while Theorem 3.6.13 guarantees Oja's rule can function for all time without sacrificing too much efficiency to adapt to a new environment.

3.6.1 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 3.6.2 and the closed form solution of Oja' rule in Corollary 3.6.4. In addition, we show that the bounded differences and moments of the noise can be controlled in Lemma 3.6.5.

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 3.3. To ease the notation, we define $\tilde{\mathbf{w}}_{t,1} = \mathbf{w}_{t,1} - 1$ and the goal becomes to show that $\tilde{\mathbf{w}}_{t_0+t_2,1} > -\epsilon$ with probability at least $1-\delta$. The following lemma states the linearization for $\tilde{\mathbf{w}}_{t,1}$.

Lemma 3.6.2 (Linearization at 1)**.** *Let* $\tilde{\mathbf{w}}_t = \mathbf{w}_{t,1}^2 - 1$ *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*

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

almost surely, where

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

Proof of Lemma 3.6.2. By expanding $\mathbf{w}_{t,1}^2$ with the Oja's rule (Equation 3.1.4), 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

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

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

$$\geq \mathbf{w}_{t-1,1}^2 + 2\eta(\lambda_1(\mathbf{w}_{t-1,1}^2 - \mathbf{w}_{t-1,1}^4) - \lambda_2(\mathbf{w}_{t-1,1}^2 - \mathbf{w}_{t-1,1}^4)) + A_t$$
$$= \mathbf{w}_{t-1,1}^2 + 2\eta(\lambda_1 - \lambda_2)\mathbf{w}_{t-1,1}^2(1 - \mathbf{w}_{t-1,1}^2) + A_t. \tag{3.6.3}$$

Based on the intuition from the continuous dynamic in Section 3.3, since we want to converge from constant error to ϵ error, we want to linearize at 1. Hence we rewrite Equation 3.6.3 in terms of $\tilde{\mathbf{w}}_{t,1} = \mathbf{w}_{t,1}^2 - 1$ and get

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

as desired. □

We apply the ODE trick (see Lemma 3.2.10) on Lemma 3.6.2 and get the following corollary.

Corollary 3.6.4 (ODE trick)**.** *For any $t_0 \in \mathbb{N}_{\geq 0}$, $t \in \mathbb{N}$, and $\eta \in (0, 1)$, we have*

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

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

Lemma 3.6.5. *Let A_t, B_t be defined as in Lemma 3.6.2. For any $t \in \mathbb{N}$, we have A_t, B_t satisfy the following properties:*

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

Proof. First by Lemma 3.5.1, 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

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

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

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

By Lemma 3.5.1 and the definition of $\tilde{\mathbf{w}}_{t-1}$, we have

$$\begin{aligned}
&\leq |y_t| \cdot \left(|\tilde{\mathbf{w}}_{t-1}| + \left| \sqrt{-\tilde{\mathbf{w}}_{t-1} + 10\eta} \right| \right) \\
&\leq |y_t| \cdot \left(|\tilde{\mathbf{w}}_{t-1}| + \sqrt{|\tilde{\mathbf{w}}_{t-1}|} + \sqrt{10\eta} \right). \qquad (3.6.6)
\end{aligned}$$

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

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

Combining above, Lemma 3.5.1 and the fact that $\mathbf{z}_{t,1} = O(1)$, we have

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

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

For conditional expectation, notice that $\mathbb{E}[y_t^2|\mathcal{F}_{t-1}] = \mathbf{w}_{t-1}^T \mathsf{diag}(\lambda)\mathbf{w}_{t-1} = O(\lambda_1)$. 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

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

By Equation 3.6.6, we have

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

as desired. □

3.6.2 Concentration of noise and pulling out the stopping time

In this subsection, we want to show that the noise term in Corollary 3.6.4 is small. Specifically, we prove the following lemma.

Lemma 3.6.7. *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 $\tilde{\mathbf{w}}_{t_0} \geq v_0$ 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 $\tilde{\mathbf{w}}_t$ might *go back* to the small region (*e.g.,* $\tilde{\mathbf{w}}_t < -2/3$) and thus the bounded difference might become too large to bound the noise effectively with Freedman's inequality. Nevertheless, the continuous dynamic (see Section 3.3) 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 $\tilde{\mathbf{w}}_t$ 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 3.6.8 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 prove Lemma 3.6.7. This can be done by Lemma 3.6.9 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 3-3.

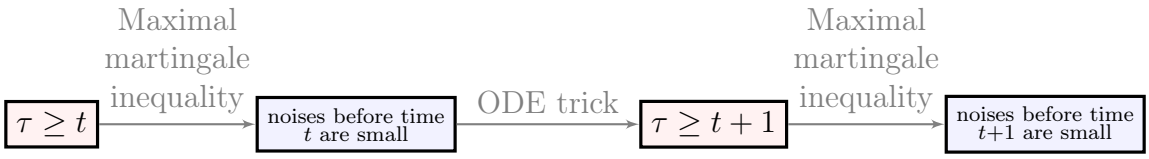


Figure 3-3: Intuition on why it is possible to pull out stopping time in Phase 2.

With the above *chain* structure in the noise terms, we are then able to pull out the stopping time in Lemma 3.6.8 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 3.6.7.

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

Lemma 3.6.8 (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 $\tilde{\mathbf{w}}_{t_0} \geq v_0$ and $v_0 = -\Theta(\epsilon^{1-a})$. Let τ_{v_0} to be the stopping time $\{\tilde{\mathbf{w}}_t < 2v_0\}$ 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. We are going to apply Freedman's inequality Corollary 3.2.6 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 $\bar{A}_t$. Now by Lemma 3.6.5 and geometric series, *i.e.,* $\sum_{i=1}^{T} H^{-i} \leq O\left(\frac{H^{-T}}{\eta(\lambda_1-\lambda_2)}\right)$, we have

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

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

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

By applying the above bounds to Lemma 3.2.5, we have

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

because the deviation term is $O\left(\sqrt{\frac{\log\frac{1}{\delta}\eta\lambda_1\epsilon^{1-2a}}{\lambda_1-\lambda_2}}\right) = O\left(\epsilon^{1-a}\right) \leq \frac{|v_0|}{4}$ and the summation of the conditional expectation terms is $O\left(\frac{\eta\lambda_1\epsilon^{-\frac{a}{2}}}{\lambda_1-\lambda_2}\right) = O\left(\epsilon^{1-a}\right) \leq \frac{|v_0|}{4}$. By stopping time and Lemma 3.6.5, we have

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

By combining both inequalities, we get

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

□

We are going to pull out the stopping time τ_{t_0} in Lemma 3.6.8. 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 3.6.9. *Let $\{M_t\}_{t\in\mathbb{N}_{\geq 0}}$ be an adapted stochastic process and τ be a stopping time. Let $\{M_t^*\}_{t\in\mathbb{N}_{\geq 0}}$ be the maximal process of $\{M_t\}_{t\in\mathbb{N}_{\geq 0}}$ where $M_t^* = \max_{1\leq t'\leq t} M_t$. For any $t \in \mathbb{N}$, $a \in \mathbb{R}$, and $\delta \in (0,1)$, suppose*

1. $\Pr[M_{t\wedge\tau}^* \geq a] < \delta$ *and*

2. 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 3.6.9. The key idea is to introduce another stopping time which helps us partition the probability space. Let τ' be the stopping time for the event $\{M^*_{t\wedge\tau} \ge a\}$. The following claim shows that if τ stopped before time t, then τ' should stop earlier than τ.

Claim 3.6.10. *Let* τ *and* τ' *be stopping times as defined above. Suppose the conditions in Lemma 3.6.9 hold. Then we have*

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

Proof of Claim 3.6.10. The claim can be proved by contradiction as follows. Suppose both $\tau < t$ and $\tau' > \tau$. By the definition of τ', we know that $M^*_\tau < a$ since $\tau < \tau'$. However, by the second condition of the lemma, we then have

$$\Pr[\tau \ge \tau + 1 \mid M^*_\tau < a] = 1,$$

which is a contradiction. □

Next, we will show that $\Pr[M^*_t \ge a] \le \Pr[M^*_{t\wedge\tau} \ge a]$. The idea is partitioning the probability space as follows. We have

$$\begin{aligned}\Pr[M^*_t \ge a] &= \Pr[M^*_t \ge a, \tau \ge t] + \Pr[M^*_t \ge a, \tau < t, \tau' \le \tau] \\ &\quad + \Pr[M^*_t \ge a, \tau < t, \tau' > \tau].\end{aligned}$$

By Claim 3.6.10, we have $\Pr[M^*_t \ge a, \tau < t, \tau' > \tau] = 0$. We have

$$= \Pr[M^*_t \ge a, \tau \ge t] + \Pr[M^*_t \ge a, \tau < t, \tau' \le \tau].$$

For the first term, when $\tau \ge t$, we have $t = t \wedge \tau$ and thus $M^*_t = M^*_{t\wedge\tau}$. As for the second term, when $\tau' \le \tau < t$, we have both $M^*_t, M^*_{t\wedge\tau} \ge a$. Thus, the equation becomes

$$\begin{aligned}&= \Pr[M^*_{t\wedge\tau} \ge a,\ \tau \ge t] + \Pr[M^*_{t\wedge\tau} \ge a,\ \tau < t,\ \tau' \le \tau] \\ &\le \Pr[M^*_{t\wedge\tau} \ge a]\,.\end{aligned}$$

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 3.6.9 on Lemma 3.6.8, we can pull out the stopping time and show concentration on the original process in Lemma 3.6.7.

Lemma 3.6.7. *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 $\tilde{\mathbf{w}}_{t_0} \geq v_0$ 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. Let τ_{v_0} be the stopping time $\{\tilde{\mathbf{w}}_t < 2v_0\}$ such that $t > t_0$. We want to apply Lemma 3.6.9 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 3.6.8. 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 3.6.4

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

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

3.6.3 Interval Analysis

Given $\epsilon \in (0,1)$, let $\tilde{\epsilon} = \frac{\epsilon}{8}$. The goal of this section is to prove the local convergence of Oja's rule (Theorem 3.6.1) with the following interval scheme that shows the improvement of $\tilde{\mathbf{w}}_t$

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

Proof of Theorem 3.6.1. Let $\tilde{\epsilon} = \frac{\epsilon}{8}$ and $v_0 = -\frac{1}{3}, l = \log\log\frac{1}{\tilde{\epsilon}}$. For $1 \leq i \leq l$, choose $T_i \in \mathbb{N}$ such that $\frac{1}{2}\tilde{\epsilon}^{\frac{1}{2^i}} \geq H^{T_i} \geq \frac{1}{4}\tilde{\epsilon}^{\frac{1}{2^i}}$ and $v_i = -\tilde{\epsilon}^{1-\frac{1}{2^i}}$. Let $S_j = \sum_{i=1}^{j} T_i$ and let $T = S_l$. Notice that by Lemma 3.2.16, we have $v_l = -\frac{\epsilon}{4}$.

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

$$\Pr\left[\tilde{\mathbf{w}}_{S_j} \leq v_j \middle| \tilde{\mathbf{w}}_{S_{j-1}} \geq v_{j-1}\right] < \frac{\delta}{l}. \tag{3.6.11}$$

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

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

as desired. What remains to be shown is Equation 3.6.11. Now by Lemma 3.6.7, 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 $1 \leq j \leq l$

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

Now by Corollary 3.6.4, the following is true with probability $1 - \delta$

$$\begin{aligned}
\tilde{\mathbf{w}}_{S_j} &\geq H^{T_j} \cdot \left(\tilde{\mathbf{w}}_{S_{j-1}} + \sum_{i=S_{j-1}+1}^{S_{j-1}+t} \frac{A_i + B_i}{H^{i-S_{j-1}}}\right) \\
&\geq \frac{1}{2}\tilde{\epsilon}^{\frac{1}{2^j}} \cdot 2v_{j-1} \\
&\geq v_j.
\end{aligned}$$

This shows that

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

as desired. □

3.6.4 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 3.6.12, we show that Oja's rule can maintain its convergence for any finite time while in Theorem 3.6.13, 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 Section 3.5.1, we prove the finite continual learning part of Theorem 3.4.2 as a corollary.

Theorem 3.6.12 (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. Given any $1 \leq j \leq l$, by Lemma 3.6.7, we have

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

Notice conditioned on $\tilde{\mathbf{w}}_{(j-1)t'} \geq -\frac{\epsilon}{2}$ 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 3.6.4

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

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

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

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 Section 3.5.1, we prove the for-all-time continual learning part of Theorem 3.4.2 as a corollary.

Theorem 3.6.13 (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. The proof proceeds by recursively choosing η_t in intervals and apply Theorem 3.6.12 repetitively. Let $\delta_i = \frac{\delta}{2i^2}$. Then notice that $\sum_{i=1}^{\infty} \delta_i < \delta$. Now apply Theorem 3.6.12 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 3.6.12, 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 3.6.12 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 $\eta_t \geq \Theta\left(\frac{\epsilon(\lambda_1 - \lambda_2)}{\lambda_1 \log \frac{t}{\delta}}\right)$. And by union bound, we get

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

□

3.7 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 random initialization for the diagonal case. By applying the diagonal reduction argument in Section 3.5.1, we prove the global convergence part of Theorem 3.4.1 as a corollary. The following theorem is the main theorem of this section.

Theorem 3.7.1. *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*

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

Then

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

The main difficulty in the global convergence is that at the beginning the bounded differences of the noise in Lemma 3.7.22 cannot be controlled directly. To be precise, the $|y_t|$ term 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 Section 3.7.1, 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 Section 3.7.2, we construct $n-1$ auxiliary stopping times and use an induction argument to show that the stopping time $\xi_{p,\delta}$ is large with high probability in Theorem 3.7.4.

In Section 3.7.3 we derive a linearization using a center at 0 instead of 1 based on the intuition from the continuous dynamic in Section 3.3. Furthermore, we use the ODE trick to write down the dynamic in closed form with respect to the linearization.

Similar to the local convergence, in Section 3.7.4, we show 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 Section 3.7.5, we prove that $\mathbf{w}_{t,1}^2$ is greater than $2/3$ efficiently with high probability in an interval analysis in Theorem 3.7.29. Finally, in Section 3.7.6, by combining Theorem 3.7.29, Theorem 3.6.1 and Theorem 3.6.12, we prove the efficient global convergence in Theorem 3.7.1.

3.7.1 Initialization and the main stopping time

In this section, we begin with Definition 3.7.2, which introduces the key stochastic processes that we study in Section 3.7.2. Then we give an initialization lemma, Lemma 3.7.3,

which guarantees that the processes perform well with good probability at the first time step.

Definition 3.7.2. *For each $2 \leq j \leq n$, $t \in [T]$, 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 [3, Lemma 5.1] with some straightforward modification below.

Lemma 3.7.3 (Initialization lemma in [3, 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

$$\Pr_{\mathbf{x}_1,\ldots,\mathbf{x}_T \sim \mathcal{D}}[\exists j \in [n],\ t \in [T],\ |f_{t,j}(\mathbf{w}_0)| > \Lambda_{p,\delta}] < \delta$$

and

$$\mathbf{w}_{0,1}^2 \geq \frac{1}{\Lambda'_p}$$

with probability at least $1 - p - \delta$ where the randomness is over $\mathbf{w}_0$. Notice that we denote the above event as $\mathcal{C}_{init}^{p,\delta}$ and we denote the event inside the probability as $\mathcal{C}_0^{p,\delta}$. In particular, the probability inequality reads $\Pr[\mathcal{C}_0^{p,\delta} \mid \mathcal{C}_{init}^{p,\delta}] < \delta$.

Given $p, \delta \in (0,1)$ in Lemma 3.7.3, 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 $\psi, \xi, \Lambda, \Lambda'$.

3.7.2 Bounding the stopping time $\xi_{p,\delta}$

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 $f_{t,n}(\mathbf{w}_{t-1})$ small with high probability. Therefore, the main goal of this section is to show that $\xi_{p,\delta}$ is large with high probability in Theorem 3.7.4. In order to prove Theorem 3.7.4, we consider

a vector linearization and the ODE trick of the auxiliary processes of Definition 3.7.2 in Corollary 3.7.8 and Corollary 3.7.9. Similar to the local convergence, we consider a stopped version of the stochastic processes, but because the randomness in ξ is shifted by 1 we define a special stopped process in Definition 3.7.5. We then obtain the concentration on the stopped processes in Lemma 3.7.15. Finally, to prove the main theorem by induction, we prove the induction step in Lemma 3.7.16 by carefully pull out the stopping time to finish the proof.

The following is the main theorem of this section.

Theorem 3.7.4. *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 $T = \Omega(\frac{1}{\eta\lambda_1})$ and $p = O(\delta)$, 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

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

First notice that the usual notion of the stopped process is not enough to use this stopping time. To give an intuition, we have

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

However, $\mathbf{1}_{\xi\geq t}$ 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. Therefore, we need to consider a different notion of the stopped process. In particular, consider the following notion of a shifted stopped process.

Definition 3.7.5. *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

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

Given a stopping time τ and an adapted stochastic process M_t, the difference of a normal stopped process can be described as

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

Similarly, we want to understand the difference of this shifted stopped process.

Lemma 3.7.6. *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. □

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.

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

Lemma 3.7.7 (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, $2 \leq j \leq n$,*

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

where

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

Proof. 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

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

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

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

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 $i = 2, 3, \ldots, j$, 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 the 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 3.7.7 by rewriting everything into a vector form.

Corollary 3.7.8 (Linearization in 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 3.2.12), we immediately have the following corollary for a closed form solution to $\mathbf{f}_t(\mathbf{w}_s)$.

Corollary 3.7.9 (ODE trick)**.** *For any $t \in [T]$, $s \in [t-1]$, we have*

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

Concentration of the noise terms We want to control the noise term in Corollary 3.7.9. 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 $\{\mathcal{F}_s^{(t)}\}$.

Now we define τ_t to be the stopping time for the first s such that $\{\|\mathbf{f}_t(\mathbf{w}_{s\wedge\psi\star\xi})\|_\infty > 2\Lambda_{p,\delta}\}$. Before we bound the bounded differences and the moments for $\mathbf{1}_{\tau_t,\psi\geq s,\xi>s}\mathbf{A}_{t,s}$, observe that we have the following helper lemma on the conditional expectation.

Lemma 3.7.10. *Let $T \in \mathbb{N}$, ξ be the stopping time specified before and $t \in [T]$. For $s < t$, given*

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

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

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

Proof. By laws of total expectation and rearrange the terms we have

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

So we have

$$\begin{aligned}
&\left|\mathbb{E}[\mathbf{x}_{s,i}\mathbf{x}_{s,j}|\mathcal{F}_{s-1}^{(t)}, \xi > s] - \mathbb{E}[\mathbf{x}_{s,i}\mathbf{x}_{s,j}|\mathcal{F}_{s-1}^{(t)}, \xi \geq s]\right| \\
&= \left|\frac{\mathbb{E}[\mathbf{x}_{s,i}\mathbf{x}_{s,j}|\mathcal{F}_{s-1}^{(t)}, \xi \geq s]\Pr[\xi = s|\xi \geq s] - \mathbb{E}[\mathbf{x}_{s,i}\mathbf{x}_{s,j}|\mathcal{F}_{s-1}^{(t)}, \xi = s]\Pr[\xi = s|\xi \geq s]}{1 - \Pr[\xi = s|\xi \geq s]}\right| \\
&\leq \frac{2\delta'}{1-\delta'}.
\end{aligned}$$

Similarly, we get

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

□

Lemma 3.7.11. *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, τ_t be the stopping times as chosen before. If $\eta = O\left(\frac{1}{\Lambda}\right), T = \Omega(\frac{1}{\eta\lambda_1}), p = O(\delta)$ and the following condition holds*

$$\forall 1 \leq t' \leq t-1,\ \Pr[\xi = t'|\xi \geq t', \mathcal{C}_{init}^{p,\delta}] \leq \frac{1}{n^2 T} \tag{3.7.12}$$

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*

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

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

Proof. The proof is basically direct verification using the definition of stopping time and Lemma 3.7.10. We postpone the proof to Section A.4. □

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

Lemma 3.7.13 (Structure of the stopped processes)**.** *Let $T \in \mathbb{N}, \eta, \delta \in (0,1), t \in [T], \bar{s} \in [t-1]$. Let Λ be the parameter specified before and ξ, τ_t be the stopping times as chosen before. For any $s \in [\bar{s}]$ and $j \in [n-1]$, let $M_{t,s,j}$ be the j^{th} entry of $\sum_{s'=1}^{s\wedge\psi\star\xi\wedge\tau_t} H^{\bar{s}-s'}\mathbf{A}_{t,s'}$. If $\eta = O\left(\frac{1}{\Lambda}\right), T = \Omega\left(\frac{1}{\eta\lambda_1}\right), p = O(\delta)$ and the following condition is true*

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

then the following holds.

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

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

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

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

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

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

Proof of Lemma 3.7.13. For notational convenience given a matrix A, we will denote its jth row as $A_{(j)}$ for the rest of the proof. Notice that $H = VDV^{-1}$ is invertible where

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

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

$$\| \left(VD'V^{-1}\right)_{(i)} \|_1 = d_i + \sum_{j=1}^{i-1} d_j - d_{j+1} = d_1. \tag{3.7.14}$$

Fixed $j \in [n]$. First we have for all $s \in [\bar{s}]$,

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

by Equation 3.7.14. For conditional expectation, we similarly have

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

By Equation 3.7.14 and Lemma 3.7.11, we have

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

So by geometric series, we have

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

For conditional variance, we similarly have

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

By Equation 3.7.14 and Lemma 3.7.11, 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 3.7.11, we are able to prove the following concentration for the stopped processes of the noise terms.

Lemma 3.7.15 (Concentration for the stopped process of the noise vectors)**.** *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, 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 $T = \Omega(\frac{1}{\eta\lambda_1}), p = O(\delta')$ and the following condition is true*

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

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

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

Proof of Lemma 3.7.15. The proof is based on applying the corollary of Freedman's inequality (see Corollary 3.2.6) on each coordinate using Lemma 3.7.13. 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 $O\left(\sqrt{\frac{\eta\lambda_1\Lambda^4_{p,\delta'}\log\frac{1}{\delta}}{\lambda_1-\lambda_2}}\right) < \frac{\Lambda_{p,\delta'}}{2}$ and the sum of conditional expectation term is $O\left(\frac{\eta\lambda_1\Lambda^3_{p,\delta'}}{\lambda_1-\lambda_2}\right) < \frac{\Lambda_{p,\delta'}}{2}$. Now we obtain the desired inequality by union bounding over $i \in [n-1]$. □

Wrap up First fix δ', δ in the Lemma 3.7.15 as $\frac{\delta}{4n^2T}$, $\frac{\delta}{4n^3T^2}$ respectively. The following lemma proves the inductive step toward the main theorem.

Lemma 3.7.16. *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 $T = \Omega(\frac{1}{\eta\lambda_1})$ and*

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

then

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

Proof. We have

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

So it suffices to bound $\Pr[\tau_t < t | \mathcal{C}_{init}^{p,\frac{\delta}{4n^2T}}]$. Notice that we have $\forall 1 \leq t' \leq t-1$,

$$\Pr[\xi = t' | \xi \geq t', \mathcal{C}_{init}^{p,\frac{\delta}{4n^2T}}] \leq \frac{\Pr[\xi = t' | \mathcal{C}_{init}^{p,\frac{\delta}{2nT}}]}{1 - \Pr[\xi < t' | \mathcal{C}_{init}^{p,\frac{\delta}{4n^2T}}]} < \frac{\frac{\delta}{4n^2T}}{1 - \frac{(t-1)\delta}{2nT}} \leq \frac{1}{nT}. \tag{3.7.17}$$

So we satisfy the condition of Lemma 3.7.15.

Let $\mathcal{A}_{s_0} = \{\exists i \in [n], \sum_{s=1}^{s_0\wedge\psi\star\xi} (H^{s_0-s}\mathbf{A}_{t,s})_i \geq \Lambda\}$ and $\mathcal{A}_{s_0}^{\tau_t}$ to be its stopped version $\{\exists i \in [n], \sum_{s=1}^{s_0\wedge\psi\star\xi\wedge\tau_t} (H^{s_0-s}\mathbf{A}_{t,s})_i \geq \Lambda\}$. Recall from Lemma 3.7.3 that $\overline{\mathcal{C}}_0^{p,\frac{\delta}{4n^2T}}$ is the event

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

We claim that

$$\Pr\left[\tau_t \geq s_0 + 1 \,\middle|\, \overline{\mathcal{C}}_0^{p,\frac{\delta}{4n^2T}}, \overline{\mathcal{A}}_{s_0}\right] = 1. \tag{3.7.18}$$

This Equation 3.7.18 is a direct consequence from the ODE trick Corollary 3.7.9. We have for any $t \in [T]$,

$$|\mathbf{f}_t(\mathbf{w}_{s_0 \wedge \psi \star \xi})| = |H^{s_0 \wedge \psi \star \xi} \mathbf{f}_t(\mathbf{w}_0) + \sum_{s'=1}^{s_0 \wedge \psi \star \xi} H^{s_0 - s'} \mathbf{A}_{t,s'}| \leq 2\Lambda$$

by the definition of $\overline{\mathcal{C}}_0^{p, \frac{\delta}{4n^2T}}, \overline{\mathcal{A}}_{s_0}$ and Equation 3.7.14. Now we have by union bound

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

By Equation 3.7.18, the first term is 0, we have

$$\leq 0 + \Pr\left[\mathcal{A}_{t-1}, \overline{\mathcal{C}}_0^{p, \frac{\delta}{4n^2T}} \;\middle|\; \mathcal{C}_{init}^{p, \frac{\delta}{4n^2T}}\right] + \Pr\left[\mathcal{C}_0^{p, \frac{\delta}{4n^2T}} \;\middle|\; \mathcal{C}_{init}^{p, \frac{\delta}{4n^2T}}\right].$$

Then by union bound and Lemma 3.7.3, we have

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

By Equation 3.7.18 again, we can rewrite the terms as

$$\begin{aligned}&= \sum_{s=1}^{t-1} \Pr\left[\mathcal{A}_s, \overline{\mathcal{A}}_{s-1}, \overline{\mathcal{C}}_0^{p, \frac{\delta}{4n^2T}}, \tau_t \geq s \;\middle|\; \mathcal{C}_{init}^{p, \frac{\delta}{4n^2T}}\right] + \frac{\delta}{4n^2T} \\ &= \sum_{s=1}^{t-1} \Pr\left[\mathcal{A}_s^{\tau_t}, \overline{\mathcal{A}}_{s-1}, \overline{\mathcal{C}}_0^{p, \frac{\delta}{4n^2T}}, \tau_t \geq s \;\middle|\; \mathcal{C}_{init}^{p, \frac{\delta}{4n^2T}}\right] + \frac{\delta}{4n^2T} \\ &\leq \sum_{s=1}^{t-1} \Pr\left[\mathcal{A}_s^{\tau_t} \;\middle|\; \mathcal{C}_{init}^{p, \frac{\delta}{4n^2T}}\right] + \frac{\delta}{4n^2T}.\end{aligned}$$

By Lemma 3.7.15, we can bound the first term by $(t-1)n\delta/4n^3T^2$

$$\leq (t-1)n\frac{\delta}{4n^3T^2} + \frac{\delta}{4n^2T} \leq \frac{\delta}{2n^2T}.$$

□

Now the main theorem can be derived as a corollary.

Theorem 3.7.4. *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 $T = \Omega(\frac{1}{\eta\lambda_1})$ and $p = O(\delta)$, 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

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

and

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

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

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

The induction step is exactly Lemma 3.7.16. The second conclusion is exactly Equation 3.7.17 and the last conclusion can be obtained from union bounding over T. □

3.7.3 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 dynamics as explained in Section 3.3. 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 Section 3.7.2 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 3.7.3 and Theorem 3.7.4, we have $\forall t \in [T]$,

$$\Pr[\mathcal{C}_{init}^T] \geq 1 - \frac{\delta}{2},\ \Pr[\xi_T < T \mid \mathcal{C}_{init}^T] < \frac{\delta}{4n^2},\ \Pr[\xi = t | \xi \geq t, \mathcal{C}_{init}^T] \leq \frac{\delta}{2n^2 T}. \quad (3.7.19)$$

We abbreviate the corresponding $\Lambda_{\delta/4,\delta/8n^2T}, \Lambda'_{\delta/4}$ as Λ, Λ' for the rest of the section.

We first derive the linearization with a center at 0.

Lemma 3.7.20 (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

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

Proof of Lemma 3.7.20. By Equation 3.6.3, 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 3.2.10) on Lemma 3.6.2 and get the following corollary.

Corollary 3.7.21 (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 3.7.22. *Let A_t, B_t be defined as in Lemma 3.6.2. 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. 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 3.5.1, 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 3.7.10, Theorem 3.7.4 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 $\mathbb{E}[\mathbf{1}_{\xi_T > t, \psi \geq t} v | \mathcal{F}_{t-1}, \mathcal{C}_{init}^T] - \mathbb{E}[v | \mathcal{F}_{t-1}]$ as $\bar{v}$. Now we also have

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

By applying Lemma 3.7.10 with Equation 3.7.19 and Cauchy-Scharwz, 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 3.7.10, 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. □

3.7.4 Concentration of noise

In this subsection, we want to show that the noise term in Corollary 3.7.21 is small. As in the local analysis, we are going to use a stopping time to control good bounded differences. Specifically,

Lemma 3.7.23 (Concentration of stopped noise 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*

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

Proof. For notational convenience, we denote $\mathbf{1}_{\tau_a, \psi \geq t, \xi > t} A_t$ as $\bar{A}_t$. Now by Lemma 3.7.22 and geometric series, *i.e.,* $\sum_{i=1}^{T} H^{-i} \leq O\left(\frac{1}{\eta(\lambda_1 - \lambda_2)}\right)$, we have

$$\forall t_0 + 1 \leq t \leq t_0 + t', \; \left|\frac{\bar{A}_t}{H^{t-t_0}}\right| \leq O\left(\eta \Lambda a\right),$$

$$\left|\sum_{i=t_0+1}^{t_0+t'} \mathbb{E}\left[\frac{\bar{A}_i}{H^{i-t_0}} \;\middle|\; \mathcal{F}_{i-1}, \mathcal{C}^T_{init}\right]\right| \leq O\left(\frac{\eta \lambda_1 \Lambda^2 a}{\lambda_1 - \lambda_2}\right), \text{ and}$$

$$\left|\sum_{i=t_0+1}^{t_0+t'} \mathsf{Var}\left[\frac{\bar{A}_i}{H^{i-t_0}} \;\middle|\; \mathcal{F}_{i-1}, \mathcal{C}^T_{init}\right]\right| \leq O\left(\frac{\eta \lambda_1 \Lambda^2 a^2}{\lambda_1 - \lambda_2}\right).$$

Apply the above bounds to Lemma 3.2.5, 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 $\sqrt{\frac{\log \frac{1}{\delta'} \eta \lambda_1 \Lambda^2 a^2}{\lambda_1 - \lambda_2}} = O\left(a\right) \leq \frac{a}{4}$ and the summation of conditional expectation term is $\frac{\lambda_1 \eta \Lambda^2 a}{\lambda_1 - \lambda_2} = O\left(a\right) \leq \frac{a}{4}$. By stopping time τ_a and Lemma 3.7.22, we have

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

Combining both inequality we get

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

□

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 3.7.24 (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*

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

Proof. Let τ_a be the stopping time $\{\mathbf{w}_{t\wedge\psi\star\xi_T,1}^2 \geq a\}$. Notice that now we only have controls on $\mathbf{w}_{t\wedge\psi\star\xi_T\wedge\tau_a,1}^2$. To conclude a statement about τ, we need to pull out ψ, τ_a, ξ_T. ξ_T will be pull out by paying union bounds in conditioning. τ_a and ψ will be pulled out similar to Lemma 3.6.9. Denote the event

$$\min_{1\leq t\leq t'} \sum_{i=t_0+1}^{(t_0+t)\wedge\psi\star\xi\wedge\tau_a} \frac{A_i + B_i}{H^{i-t_0}} \leq -\frac{a}{2}$$

as $\mathcal{A}$. By Lemma 3.7.23 we have $\Pr[\mathcal{A}|\mathcal{C}_{init}^T] < \frac{\delta'}{4}$. First let's deal with τ_a first. I claim that we have

$$\Pr\left[\tau_a > t_0 + t', \psi > t_0 + t' | \xi_T > T, \mathcal{C}_{init}^T\right] < \frac{\delta'}{2} \tag{3.7.25}$$

Notice if $\tau_a > t_0 + t', \xi_T > T, \psi > t_0 + t'$ and $\overline{\mathcal{A}}$ are true, by Corollary 3.7.21 we have

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

Contradict with $\tau_a > t_0 + t'$. This implies that

$$\Pr[\tau_a > t_0 + t', \xi_T > T, \psi > t_0 + t', \overline{\mathcal{A}}] = 0. \tag{3.7.26}$$

Now we have

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

The second term vanishes because of Equation 3.7.26. We have

$$\begin{aligned}
&\leq \Pr\left[\mathcal{A} \;\middle|\; \xi_T > T, \mathcal{C}_{init}^T\right] \\
&= \frac{\Pr\left[\mathcal{A} \;\middle|\; \mathcal{C}_{init}^T\right]}{\Pr\left[\xi_T > T \mid \mathcal{C}_{init}^T\right]}
\end{aligned}$$

Since $\Pr\left[\mathcal{A} \mid \mathcal{C}_{init}^T\right] < \frac{\delta'}{4}$ and $\Pr\left[\xi_T > T \mid \mathcal{C}_{init}^T\right] < \frac{1}{2}$ from Equation 3.7.19, we have

$$\leq \frac{\delta'}{2}.$$

Now let's deal with ψ. I claim that we have

$$\Pr\left[\tau_a > t_0 + t', \psi \leq t_0 + t' | \xi_T > T, \mathcal{C}_{init}^T\right] < \frac{\delta'}{2}. \tag{3.7.27}$$

Notice that from $\Pr\left[\mathcal{A} \mid \mathcal{C}_{init}^T\right] < \frac{\delta'}{4}$, we have

$$\Pr\left[\min_{1\leq t\leq t'} \sum_{i=t_0+1}^{(t_0+t)\wedge\psi} \frac{A_i + B_i}{H^{i-t_0}} \leq -\frac{a}{2}, \tau_a > t_0 + t', \xi_T > T \;\middle|\; \mathcal{C}_{init}^T\right] < \frac{\delta'}{4}.$$

We will apply Lemma 3.6.9 to pull out ψ in the above inequality. Denote the event

$$\min_{1\leq t\leq t'} \sum_{i=t_0+1}^{t_0+t} \frac{A_i + B_i}{H^{i-t_0}} \leq -\frac{a}{2}$$

as $\mathcal{B}$. It suffices to check that if $\bar{\mathcal{B}}$ is true, then we have $\psi > t_0 + t$. By Corollary 3.7.21 we have for all $t \in [t']$

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

as desired. By Lemma 3.6.9, this shows that

$$\Pr\left[\mathcal{B}, \tau_a > t_0 + t', \xi_T > T \mid \mathcal{C}_{init}^T\right] < \frac{\delta'}{4}.$$

Then notice that if $\tau_a > t_0 + t', \xi_T > T, \tau \leq t_0 + t'$ and $\overline{\mathcal{B}}$ is true, by second condition of Lemma 3.6.9 we just checked, we have $\tau > t_0 + t'$. Contradict with $\tau \leq t_0 + t'$. This implies that

$$\Pr[\tau_a > t_0 + t', \xi_T > T, \psi \leq t_0 + t', \overline{\mathcal{B}}] = 0.$$

Now we have

$$\begin{aligned}&\Pr\left[\tau_a > t_0 + t', \psi \leq t_0 + t' \mid \xi_T > T, \mathcal{C}_{init}^T\right] \\ &= \Pr\left[\tau_a > t_0 + t', \psi \leq t_0 + t', \mathcal{B} \middle| \xi_T > T, \mathcal{C}_{init}^T\right] \\ &+ \Pr\left[\tau_a > t_0 + t', \psi \leq t_0 + t', \overline{\mathcal{B}} \middle| \xi_T > T, \mathcal{C}_{init}^T\right].\end{aligned}$$

The second term vanishes because of Equation 3.7.4. We have

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

The numerator can be bounded by $\frac{\delta'}{4}$ and the denominator can be bounded by $\frac{1}{2}$ from Equation 3.7.19, so we have

$$\leq \frac{\delta'}{2}.$$

Now we have

$$\Pr[\tau > t_0 + t' | \xi > T, \mathcal{C}_{init}^T]$$
$$\leq \Pr[\tau, \psi > t_0 + t' | \xi_T > T, \mathcal{C}_{init}^T] + \Pr[\tau, \psi \leq t_0 + t' | \xi_T > T, \mathcal{C}_{init}^T].$$

Because $\tau > t_0 + t'$ implies $\tau_a > t_0 + t'$, we have

$$\leq \Pr[\tau_a, \psi > t_0 + t' | \xi_T > T, \mathcal{C}_{init}^T] + \Pr[\tau_a, \psi \leq t_0 + t' | \xi_T > T, \mathcal{C}_{init}^T].$$

By Equation 3.7.25 and Equation 3.7.27, we have

$$< \frac{\delta'}{2} + \frac{\delta'}{2} = \delta'$$

as desired.

□

3.7.5 Interval Analysis

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

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

We first show in Lemma 3.7.28 on how to choose the learning rate without dependency on T and then show that $\mathbf{w}_{t,1}^2$ is going to reach $2/3$ efficiently.

Lemma 3.7.28. *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. 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 3.7.29. *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*

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

Then we have

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

Proof. Choose t' such that $8 \geq H^{t'} \geq 4$ and let $m = \lfloor \log \frac{2}{3\Lambda'} \rfloor + 1$, $v_i = \frac{1}{\Lambda'} 2^i$ 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 3.7.24 with $\delta' = \frac{\delta}{4m}$. Notice that since $\log \Lambda' = O(nT)$, we can choose δ' this way. Now we have

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

By Equation 3.7.19 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 3.7.24, 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. □

3.7.6 Combining Theorem 3.7.29 with the local analysis

In this section, since we have shown that $\mathbf{w}_{t,1}^2$ efficiently reaches $2/3$ in Theorem 3.7.29, by combining Theorem 3.7.29, the local convergence (Theorem 3.6.1) and the finite continual learning (Theorem 3.6.12), we derive Theorem 3.7.1.

Proof of Theorem 3.7.1. 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 3.7.29, Theorem 3.6.1 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 3.6.12 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

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

□

3.8 Discussion and Future Directions

In this work, our contributions are three-fold. In terms of biology, we show that Oja's rule can solve streaming PCA in a biologically realistic time scale as an example of fast sensory adaptation under the efficient coding principle. Moreover, we demonstrate the capacity of Oja's rule for continual learning. With only slowly diminishing learning rate that decreases like $\Omega(1/\log t)$, we show that

$$\Pr[\exists t \ge T, \text{ error at time } t > \epsilon] < \delta.$$

This shows that Oja's rule not only can function indefinitely but also can continuously adapt to different environments without sacrificing much efficiency or resetting the

learning rate.

In terms of algorithms, we give the first convergence rate analysis for biological Oja's rule in solving streaming PCA. As a byproduct, the convergence rate we get for biological Oja's rule outperforms the state-of-the-art upper bound for streaming PCA (using ML Oja's rules) and matches the information-theoretic lower bound up to logarithmic factors.

In terms of mathematics, we develop a novel one-shot framework to analyze a stochastic process using inspiration from the continuous dynamic as a guide. Instead of using the traditional step-by-step analysis, this framework writes down the closed form solution of the dynamic and uses stopping times to obtain precise control of the dynamics. This framework provides a more elegant and more general analysis compared with the previous step-by-step approaches. And we hope it can inspire future works on analyzing stochastic processes.

At the rest of the section, we discuss some future directions in both the biological aspects and the algorithmic aspects.

3.8.1 Biological aspects

Spiking Oja's Rule In this thesis, 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 [1], 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 Section 3.1.4, there are plenty of Hebbian-type learning rules that had been proposed to solve some computational problems [71, 14, 70, 84, 63, 4, 67]. 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* [3]. As mentioned in Section 3.1.4, computational neuroscientists have proposed several variants of biological Oja's rule to solve streaming k-PCA [60, 70, 26, 46, 69, 43, 67]. Some networks use feedforward connections only but the learning rules are not local [60, 70] while some use Hebbian learning on the feedforward connection and use anti-Hebbian learning on the recurrent connection to decorrelate the outputs [26, 46, 69, 43, 67]. 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.

3.8.2 Algorithmic aspects

Improving the guarantees for biological Oja's rule In this thesis, 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.g. 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 [3], 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.

3.9 Contribution Statement

The work in this chapter is done as joint work with Chi-Ning Chou.

- Analysis framework conception: Mien Brabeeba Wang
- Mathematical analysis: Chi-Ning Chou and Mien Brabeeba Wang
- Biological motivation: Mien Brabeeba Wang
- Writing: Chi-Ning Chou and Mien Brabeeba Wang
- Editing: Chi-Ning Chou and Mien Brabeeba Wang

Appendix A

Appendix

A.1 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.

$$\mathbf{w}_t = \frac{\left(I + \eta_t \mathbf{x}_t \mathbf{x}_t^\top\right) \mathbf{w}_{t-1}}{\| \left(I + \eta_t \mathbf{x}_t \mathbf{x}_t^\top\right) \mathbf{w}_{t-1} \|_2} \,. \tag{A.1.1}$$

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

$$= \left(1 + 2\eta_t y_t^2 + O(\eta_t^2)\right)^{-1/2} = 1 - \eta_t y_t^2 + O(\eta_t^2) \,. \tag{A.1.2}$$

Replace the denominator of Equation A.1.1 with Equation A.1.2 and truncate the $O(\eta_t^2)$ term, one recovers biological Oja's rule (Equation 3.1.4).

A.2 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 A.2.1 (Properties of continuous Oja's rule)**.** *Let* $\mathbf{w}_0 \in \mathbb{R}^n$ *such that* $\|\mathbf{w}_0\|_2 = 1$ *and* $\mathbf{w}_{0,1} > 0$*. For any* $t \geq 0$*, we have*

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

almost surely.

Proof of Lemma A.2.1. In the following, everything holds almost surely so we would not mention this condition every time. First, consider

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

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 3.3.4 as follows.

Theorem 3.3.4 (Linearization at 0)**.** *Suppose* $\mathbf{w}_{0,1} > 0$*. For any* $\epsilon \in (0,1)$*, when* $t \geq \Omega\left(\frac{\log(1/\mathbf{w}_{0,1}^2)}{\epsilon(\lambda_1-\lambda_2)}\right)$*, we have* $\mathbf{w}_{t,1}^2 > 1-\epsilon$*.*

Proof of Theorem 3.3.4. Observe that for any $t \geq 0$ such that $\mathbf{w}_{t,1}^2 \leq 1-\epsilon$, by the second item of Lemma A.2.1, we have

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

Let $\tau = \frac{10\log(1/\mathbf{w}_{0,1}^2)}{\epsilon(\lambda_1-\lambda_2)}$ and assume $\mathbf{w}_{\tau,1}^2 \leq 1-\epsilon$ for the sake of contradiction. From the above linearization and $\mathbf{w}_{t,1}$ being non-decreasing (the third item of Lemma A.2.1), we have

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

which is a contradiction to the first item of Lemma A.2.1. Thus, we conclude that for any $t = \Omega\left(\frac{\log(1/\mathbf{w}_{0,1}^2)}{\epsilon(\lambda_1-\lambda_2)}\right)$, $\mathbf{w}_{t,1}^2 > 1-\epsilon$. □

Now, we restate and prove Theorem 3.3.5 as follows.

Theorem 3.3.5 (Linearization at 1)**.** *Suppose* $\mathbf{w}_{0,1} > 0$*. For any* $\epsilon \in (0,1)$*, when* $t \geq \Omega\left(\frac{\log(1/\epsilon)}{\mathbf{w}_{0,1}(\lambda_1-\lambda_2)}\right)$*, we have* $\mathbf{w}_{t,1}^2 > 1-\epsilon$*.*

Proof of Theorem 3.3.5. Observe that for $t \geq 0$, by the second item of Lemma A.2.1, we have

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

As $\mathbf{w}_{t,1}$ is non-decreasing (the third item of Lemma A.2.1) 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 $\mathbf{w}_{t,1}^2 > 1-\epsilon$. □

A.3 Why the Analysis of ML Oja's Rule Cannot be Applied to Biological Oja's Rule

In this section, we discuss what makes biological Oja's rule much harder to analyze compared to the previous approaches for ML Oja's rule [3]. 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 must use at least two objective functions with different linearizations to get a 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 how 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 if this is the case, previous techniques cannot show this.

A.4 Proof of Lemma 3.7.11

Proof of Lemma 3.7.11. The proof is basically direct verification using the definition of ξ, τ_t, ψ and Lemma 3.7.10. 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

$$\|\nabla f_{t,j}(\mathbf{w}_{s-1})\|_2 = O(\frac{\Lambda}{\mathbf{w}_{s-1,1}^2}) = O(\sqrt{\Lambda'}\Lambda). \tag{A.4.1}$$

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

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

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

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

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

where

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

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

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

In particular, $\bar{\mathbf{w}}_{s-1,i} = O(\mathbf{w}_{s-1,i} + \mathbf{w}_{s-1,1}\mathbf{x}_{s,i})$. 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

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

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

$$\leq O\left(\frac{\Lambda}{\mathbf{w}_{s-1,1}^2}\|v\|_2^2\right) \tag{A.4.2}$$

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 $\bar{\mathbf{z}}_s$ and separate $\mathbf{A}_{s,j}^{(t)}$ into two terms where $\mathbf{A}_{s,j}^{(t,1)} = \eta\nabla f_{t,j}(\mathbf{w}_{s-1})^T\bar{\mathbf{z}}_s$ and $\mathbf{A}_{s,j}^{(t,2)} = \eta^2\mathbf{z}_s^T\nabla f_{t,j}^2(\overline{\mathbf{w}}_{s-1})\cdot\mathbf{z}_s$. By Cauchy-Schwarz and Equation A.4.1, We have

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

We also have

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

By Equation A.4.2, 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

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

Notice that we have $\mathbb{E}[\mathbf{z}_s|\mathcal{F}_{s-1}^{(t)},\mathcal{C}_{init}^{p,\delta}] = \mathbb{E}[\mathbf{x}_s\mathbf{x}_s^T\mathbf{w}_{s-1} - \mathbf{w}_{s-1}^T\mathbf{x}_s\mathbf{x}_s^T\mathbf{w}_{s-1}\mathbf{w}_{s-1}|\mathcal{F}_{s-1}^{(t)},\mathcal{C}_{init}^{p,\delta}]$. By Lemma 3.7.10 applying on $\mathbf{x}_s\mathbf{x}_s^T$ and Cauchy-Schawrtz, 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 A.4.1 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

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

Notice we have $\|\mathbb{E}[\mathbf{1}_{\psi,\tau_t\geq s,\xi>s}\mathbf{z}_s\mathbf{z}_s^T|\mathcal{F}_{s-1}^{(t)}]\|_2 = O(y_s^2\lambda_1)$ by Lemma 3.7.10. Again by Equation A.4.2, 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

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

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

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

For the first term, we have

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

Notice we have $\|\mathbb{E}[\mathbf{1}_{\psi,\tau_t\geq s,\xi>s}\bar{\mathbf{z}}_s\bar{\mathbf{z}}_s^T|\mathcal{F}_{s-1}^{(t)},\mathcal{C}_{init}^{p,\delta}]\|_2 = O(y_s^2\lambda_1)$ by Lemma 3.7.10. 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

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

By taking the maximum of the $\mathbf{A}_{s,j}^{(t,1)}$ and combining with Equation A.4.2, 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

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

Combining all the terms, we get

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

□

Bibliography

[1] Larry F. Abbott and Sacha B. Nelson. Synaptic plasticity: taming the beast. *Nature Neuroscience*, 3:1178–1183, 2000.

[2] Edgar D. Adrian and Yngve Zotterman. The impulses produced by sensory nerve-endings: Part ii. the response of a single end-organ. *Journal of Physiology*, 61(2):151–171, 1926.

[3] 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.

[4] 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.

[5] 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.

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

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

[8] Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

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

[10] 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.

[11] Horace B. Barlow. Possible principles underlying the transformations of sensory messages. pages 217–234. The MIT Press, 1961.

[12] Guo-qiang Bi and Mu-ming Poo. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of Neuroscience*, 18:10464–10472, 1998.

[13] William Bialek, Rob de Ruyter van Steveninck, Fred Rieke, and David Warland. *Spikes - exploring the neural code*. MIT Press, Cambridge, MA., 1996.

[14] 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.

[15] Tim V. Bliss and Terje Lømo. Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *Journal of Physiology*, 232(2):331–356, 1973.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] Konstantinos I. Diamantaras and Sun Yuan Kung. *Principal component neural networks: theory and applications*. John Wiley & Sons, Inc., 1996.

[22] Serena M. Dudek and Mark F. Bear. Homosynaptic long-term depression in area ca1 of hippocampus and effects of n-methyl-d-aspartate receptor blockade. *Proceedings of the National Academy of Sciences of the United States of America*, 89:4363–4367, 1992.

[23] Marie Duflo. *Random iterative models*, volume 34. Springer Science & Business Media, 2013.

[24] Nicolas Fourcaud-Trocmé, David Hansel, Carl van Vreeswijk, and Nicolas Brunel. How spike generation mechanisms determine the neuronal response to fluctuating input. *Journal of Neuroscience*, 23:11628–11640, 2003.

[25] David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, pages 100–118, 1975.

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

[27] Tim Gollisch and Markus Meister. Rapid neural coding in the retina with relative spike latencies. *Science*, 319:1108–1111, 2008.

[28] Michel Habib, Colin McDiarmid, Jorge Ramirez-Alfonsin, and Bruce Reed. *Probabilistic methods for algorithmic discrete mathematics*, volume 16. Springer Science & Business Media, 2013.

[29] 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.

[30] 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.

[31] George F. Harpur and Richard W. Prager. *Experiments with simple Hebbian-based learning rules in pattern classification tasks*. Citeseer, 1994.

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

[33] 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.

[34] Yael Hitron and Merav Parter. Counting to ten with two fingers: Compressed counting with spiking neurons. *27th Annual European Symposium on Algorithms*, 2019.

[35] Alan L. Hodgkin and Andrew F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, 117(4):500–544, 1952.

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

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

[38] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[39] 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.

[40] 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.

[41] Stephan R. Kelso, Alan H. Ganong, and Thomas H. Brown. Hebbian synapses in hippocampus. *Proceedings of the National Academy of Sciences of the United States of America*, 83:5326–5330, 1986.

[42] Werner M. Kistler, Wulfram Gerstner, and J. Leo van Hemmen. Reduction of hodgkin-huxley equations to a threshold model. *Neural Computation*, 9:1069–1100, 1997.

[43] 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.

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

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

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

[47] 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.

[48] Michael S. Lewicki. Efficient coding of natural sounds. *Nature Neuroscience*, 5(4):356–363, 2002.

[49] 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.

[50] Ralph Linsker. Towards an organizing principle for a layered perceptual network. In D. Z. Anderson, editor, *Neural Information Processing Systems*, pages 485–494. American Institute of Physics, 1988.

[51] 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.

[52] Nancy A. Lynch and Frederik Mallmann-Trenn. Learning hierarchically structured concepts. *arXiv preprint arXiv:1909.04559*, 2019.

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

[54] 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.

[55] 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.

[56] 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.

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

[58] Wolfgang Maass. Lower bounds for the computational power of networks of spiking neurons. *Neural Computation*, 8:1–40, 1996.

[59] Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15(3):267–273, 1982.

[60] Erkki Oja. Principal components, minor components, and linear neural networks. *Neural networks*, 5(6):927–935, 1992.

[61] 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.

[62] Bruno A. Olshausen and David J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37(23):3311–3325, 1997.

[63] 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.

[64] 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.

[65] Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572, 1901.

[66] 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.

[67] 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.

[68] Mark D. Plumbley. Lyapunov functions for convergence of principal component algorithms. *Neural Networks*, 8(1):11–23, 1995.

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

[70] Terence D. Sanger. Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural networks*, 2(6):459–473, 1989.

[71] Terrence J. Sejnowski. Storing covariance with nonlinearly interacting neurons. *Journal of mathematical biology*, 4:303—-321, 1977.

[72] Ohad Shamir. Convergence of stochastic gradient descent for pca. In *International Conference on Machine Learning*, pages 257–265, 2016.

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

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

[75] 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.

[76] 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.

[77] Christian D. Swinehart and Larry F. Abbott. Dimensional reduction for reward-based learning. *Network: Computation in Neural Systems*, 17(3):235–252, 2006.

[78] Taro Toyoizumi, Megumi Kaneko, Michael P. Stryker, and Kenneth D. Miller. Modeling the dynamic interaction of hebbian and homeostatic plasticity. *Neuron*, 84(1):497–510, 2014.

[79] Gina G. Turrigiano. The self-tuning neuron: Synaptic scaling of excitatory synapses. *Cell*, 135(3):422–435, 2008.

[80] Gina G. Turrigiano. Homeostatic synaptic plasticity: Local and global mechanisms for stabilizing neuronal function. *Cold Spring Harb Perspective in Biology*, 4(1):1–17, 2012.

[81] David Williams. *Probability with martingales*. Cambridge university press, 1991.

[82] Hugh R. Wilson and Jack D. Cowan. Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical journal*, 12(1):1–24, 1972.

[83] 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.

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

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

[86] 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.

[87] 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.

[88] 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.