Fall 10-1-2021

On Neuroscience-Inspired Statistical and Computational Problems

Zifan Li
Yale University Graduate School of Arts and Sciences, lizifanterry@gmail.com

Abstract

On Neuroscience-Inspired Statistical and Computational Problems

Zifan Li

2021

Recent years have witnessed a surge of problems lying at the intersection of statistics and neuroscience. In this thesis, we explore various statistical and computational problems that are inspired by neuroscience. This thesis consists of two main parts, each inspired by a different system in the brain.

In the first part, we study problems related to the visual system. In Chapter 2, we investigate the problem of estimating the collision time of a looming object using a theoretical formulation based on statistical hypothesis testing. In Chapter 3, we build computational models for the compound eye of Drosophila, and analyze how the models recover features of actual visual loom-selective neurons.

In the second part, we study problems related to the memory system. In Chapter 4, we consider approaches for accelerating and reducing memory requirements for reinforcement learning algorithms, with provable guarantees on the performance of the algorithms.

On Neuroscience-Inspired Statistical and Computational Problems

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Zifan Li

Dissertation Director: John Lafferty

December 2021

Contents

List of Figures

List of Tables

Dedicated to my parents Xiaobing and Liying

for their endless support and love

Acknowledgements

Having John Lafferty as my advisor is one of the best decisions I have ever made. During my Ph.D. years, I received numerous inspirations and incredible support from him. His constant feedback and guidance have helped me overcome many difficulties during my study. His enthusiasm for research and breadth of knowledge have inspired me to become a better researcher.

I would like to thank Professor Damon Clark, Professor Ilker Yildirim, and Professor Dragomir Radev for guiding me through research projects with their expert knowledge. They have brought me many fresh and interesting perspectives on the application of statistics and data science to neuroscience, psychology and natural language processing.

In addition to the professors, I have worked with many student collaborators during my Ph.D. study, including Baohua Zhou, Xinyi Zhong, Qi Lin, Sunnie Kim, Tao Yu, and many others. I feel priviledged that I had the opportunity to work with so many brilliant researchers, and I benefited a lot by learning from their knowledge about neuroscience, psychology, statistics, machine learning, and natural language processing.

I want to thank the faculty of the Department of Statistics and Data Science at Yale for the outstanding courses they provided, which helped me build a solid foundation for research. I am very grateful to Professor Zhou Fan and Professor Andrew Barron for attending my oral exam and asking insightful questions. I would also like to thank Professor Jay Emerson for being great director of graduate studies, and the department's staff including Joann DelVecchio, Elizavette Torres, and Karen Kavanaugh for their support.

The past four years wouldn't have been so enjoyable without my family and friends. I would like to thank Haitong for being a constant source of joy in my life. I am also very grateful for the endless support provided by my family.

Chapter 1

Introduction

Neuroscience seeks to understand how brains function in terms of principles that transform molecular dynamics to behaviors [Kass et al., 2018]. Abundant interesting statistical and computational problems arise as a result of the vast amount of high-dimensional and dynamic data we obtain from measuraments of the complex brain system. Advanced statistical theory and novel computational tools could be utilized to understand empirical findings in neuroscience. At the same time, knowledge about brain functions and behaviors revealed by experiments serve as inspirations for new statistical models and learning algorithms. In this thesis, we provide rigorous statistial analysis for computationally efficient algorithms inspired from neuroscience. We also build computational models and investigate how their behaviors align with empirical evidence from neuroscience. We focus on problems related to two important systems in the brain, the visual system and the memory system.

One crucial ability of the visual system is to detect looming stimuli, which refer to stimuli of objects directly approaching the observer. Recently, two neuron types in *Drosophila* have been identified as candidates for loom detectors [Ache et al., 2019a, Klapoetke et al., 2017]. However, several important related questions, such as the relationship between these receptive fields, or the nonlinear computations of these neurons

remain unclear. We study problems related to looming stimuli detection from both theoretical and computational perspectives.

From a theoretical perspective, we study the problem of estimating the collision time of a looming object in Chapter 2 using a formulation based on statistical hypothesis testing. The looming stimuli are abstracted as a sequence of two-dimensional images of an object as projected onto the retina of an observer toward which the object is moving. The time-to-contact is estimated as the ratio between the retinal image size and the rate of expansion of the image at a given instant. The problem is parameterized by the number of pixels or sensors in the retinal background image. We consider a testing framework where the null hypothesis is the event that the time-to-contact is less than a fixed threshold. With the number of pixels increasing, we aim to develop efficient and optimal testing procedures for which the false-negative and false-positive probabilities are both asymptotically zero.

From a computational perspective, in Chapter 3 we aim to investigate computational models for loom detection in the compound eye, informed by current understanding of the neurophysiology of the visual system in Drosophila. We follow anatomical data to model a population of loom-sensitive cells that each receive input from overlapping regions of a grid of local motion sensors. Of particular interest is the shared filter used by these neurons when adapted to detect looming stimuli. We aim to study the behavior of the filters learned under different computational models in a data-driven manner by simulating optical flow of looming and non-looming objects.

Given that human brain does not have unlimited memory capacity, a tradeoff between the amount and fidelity of the memory must exist in our brain system. However, behavioral evidence suggests that humans can store massive amount of memory with vivid detail. [Brady et al., 2011]. Drawing inspirations from this phenomenon, we consider approaches for accelerating and reducing memory requirement for learning algorithms, with provable guarantees on the performance of the algorithms. In Chapter 4 we study guarantees of reinforcement learning algorithms under memory constraints. More specifically, we con-

sider how matrix sketching technique that reduce computational and memory complexity can be applied in reinforcement learning problems and its impact on the regret guarantee.

Note that Chapter 3 is based on joint work with Baohua Zhou, Sunnie Kim, Damon Clark, and John Lafferty. My contributions to the joint work include the formulation of the task as a machine learning problem, implementation of the models in Python, running of computational experiments, analysis of simulation data, contributing to refinements of the models, and to the write-up of results for journal submission.

Chapter 2

Estimating Time-to-Contact for Approaching Objects

2.1 Introduction

The ability to detect and avoid approaching objects is critical to the survival of many animal species; it is also a key component of AI systems that govern the sensing and control of autonomous vehicles. In this chapter we study the problem of estimating the collision time of an approaching object, using a formulation based on statistical hypothesis testing. Our starting point is the geometry of a sequence of two-dimensional images of an object as projected onto the retina of an observer toward which the object is moving. The time-to-contact is estimated as the ratio between the retinal image size and the rate of expansion of the image at a given instant. The problem is parameterized by the number of pixels or sensors in the retinal background image. We consider a testing framework where the null hypothesis is the event that the time-to-contact is less than a fixed threshold. With the number of pixels increasing, we consider tests for which the false-negative and false-positive probabilities are both asymptotically zero. In this setting, we establish efficient and optimal testing procedures for one-dimensional intervals, two-dimensional rectangles, and two-dimensional disks, using a penalized multiscale scan statistic for identification

and estimation of area. The signal-to-noise ratios achievable by our testing procedures match recently established statistical lower bounds for signal detection for these same geometries.

Taking inspiration from natural models of visual perception in animals, we abstract the retina as a two-dimensional pixel field with fixed length and width, which we model as $[0,1]^2$, discretized to an n-by-n array of pixels. In our statistical formulation, we let the number of pixels n increase to study the asymptotic behavior. The image of the object at an initial time t_0 is denoted as S_0, which we consider to be a disk with radius r_0 centered at the center of the pixel field. We let S_1 denote the expanded image at time t_1, and assume $t_1 - t_0$ is small. For simplicity, we focus on the case where the expansion is symmetric, corresponding to the event that the object is directly approaching the observer. Thus, S_1 is another disk with radius $r_1 \approx r + dr$. With $|S_0|$ and $|S_1|$ denoting the area of disks S_0 and S_1 respectively, the underlying geometry implies that the time-to-contact is well-approximated as

$$T = \frac{r_0}{dr} = \frac{r_0}{r_1 - r_0} = \frac{1}{\sqrt{\frac{|S_1|}{|S_0|}} - 1}$$

as we explain in the following section. For each pixel $p = (\frac{i}{n}, \frac{j}{n})$ in field of vision, the object is observed as a uniform intensity with additive Gaussian noise,

$$Y_t(p) = \mu_n \mathbb{1}\{p \in S_t\} + Z_t(p),$$

where Z_t is a $n \times n$ matrix of Gaussian white noise with variance one, and $\mu_n > 0$ is the signal amplitude that regulates the signal-to-noise ratio.

We formulate a statistical test of whether T is smaller or larger than a threshold T_{thresh}. Informally, this asks the question "is there enough time to avoid the approaching object"? In this work our objective is to identify the minimal threshold on the signal strength μ_n that permits a sequence of tests $\{\phi_n\}_{n=1}^{\infty}$ for which the false-negative and false-positive proba-

bilities are both asymptotically zero. Our main result is that for two-dimensional disks and rectangles, a signal strength scaling as $\mu_n = \Omega(1/n)$ suffices. Moreover, by computing the test statistics hierarchically, an algorithm can be devised with complexity $\widetilde{\mathcal{O}}(n^2)$ [1], linear in the number of pixels up to a logarithmic factor. We show that asymptotically optimal tests can be derived by estimates of the retinal image area at time t_0 and t_1. To estimate this area, we use penalized scan statistics, which compute the maximum of a Gaussian process parameterized by rectangles with varying scales that scan over the background, with a penalization term depending on its scale. Our analysis leverages previous work that achieves fast algorithms through the use of modified families of dyadic rectangles.

Contributions Our main contribution in this chapter is three-fold. First, we present a novel formulation of the visual time-to-contact estimation problem using a statistical hypothesis testing framework, and study its asymptotic property as the number of pixels of the retinal background image goes to infinity. Second, we propose efficient testing procedure based on penalized multiscale scan statistic that is provably optimal when the underlying signal has the form of either a rectangle or a disk. While Arias-Castro et al. [2005] studies multiscale scan statistics for general geometric objects, their scan statistics are not optimal in terms of detection thresholds as they are not penalized. Another related work is Kou [2017] which studies the identification threshold of penalized scan statistic for hyperrectangles. However, their results are not directly applicable to disks and they rule out signals with very large scale (area remains constant as $n \to \infty$), which corresponds to our setting. Last, we perform extensive numerical simulations on the efficient testing procedure under a variety of settings. Simulation results corroborate our theory and demonstrate that the efficient procedure could attain comparable performance as the more accururate but inefficient counterpart in shorter time.

[1] Here $\widetilde{\mathcal{O}}$ hides only constants and poly-logarithmic factors.

Content The rest of the chapter is organized as follows. In Section 2.2 we introduce the geometry of the perception of approaching objects. In Section 2.3 we give a statistical formulation of the hypothesis testing problem and introduce the concept of asymptotically powerful tests. Section 2.4 briefly surveys recent work on signal detection/identification and scan statistics. In Section 2.5 we present an efficient estimation procedure based on a penalized multiscale scan statistic in the one dimensional case as a warm-up, and establish its optimality. In Section 2.6 we present an efficient estimation procedure for two-dimensional rectangular signals. Most results in Section 2.6 are straightforward extensions from results in the one dimensional case. In Section 2.7 we extend the efficient estimation procedure to two-dimensional disk signals. Section 2.8 presents numerical simulations results. A sketch of the proof of the main results in one-dimensional case is provided in Section 2.9, with the detailed proofs of more general cases collected in Section 2.10.

2.2 Perception of Approaching Objects

Gibson [1979] suggests that the symmetrical expansion of the retinal image is a crucial feature for determining whether an object is on a direct collision course with an observer. As a result, sensitivity to optical expansion is critical for selection of an appropriate response in order to avoid a collision—for example, when crossing the road. The time before collision could be derived from knowledge of the distance and speed of the object, but these quantities are not directly available to the observer. Lee [1976] suggests that many animals form a decision by estimating the ratio between retinal image size and the rate of expansion of the image at a given instant.

Consider the visual geometry illustrated in Figure 1. Suppose that the two dimensional projection of the object onto the vision field is a disk of radius R and distance $D(t)$ from the lens of the eye at time t. It subtends an angle of $2\theta(t)$, where $\theta(t)$ is given by the formula $\tan(\theta(t)) = \frac{R}{D(t)}$. Suppose the constant distance between the retina and the focal

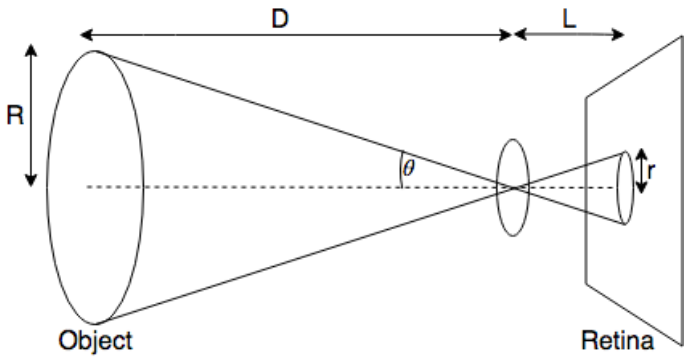

Figure 2.1: Visual geometry of an object approaching an observer.

point is L and the radius of the retinal image is $r(t)$ at time t. Further suppose that the object is moving at a constant velocity v directly towards the observer. In this case, the time-to-contact T is simply

$$T = \frac{D(t)}{v} = -\frac{D(t)}{\dot{D}(t)}.$$

Lee [1976] defines τ as the ratio of optical size $\theta(t)$ and the rate of looming $\dot{\theta}(t)$ at time t. When the image is not too large relative to its distance, i.e., when $\theta(t)$ is small, the small angle approximation suggests that $\theta(t) \approx \frac{R}{D(t)}$ and the time-to-contact is well-approximated as

$$\tau := \frac{\theta(t)}{\dot{\theta}(t)} \approx -\frac{D(t)}{\dot{D}(t)}.$$

Considerable research has been conducted on humans' use of τ in estimating T. [Schiff and Detwiler, 1979, McLeod and Ross, 1983, Regan and Hamstra, 1993, Wann et al., 2011]. Yan et al. [2011] show that time-to-contact specified by the optical variable τ was weighted more by subjects than estimates derived from distance, speed, or object size when making relative time-to-contact judgments.

Going beyond the optical variable τ, if we consider the radius of the retinal image $r(t)$ instead of optical size $\theta(t)$ at time t, then the ratio of retinal image radius $r(t)$ and its expansion rate $\dot{r}(t)$ is exactly equal to T, even without the small angle assumption. This is because $\frac{r(t)}{L} = \tan(\theta(t)) = \frac{R}{D(t)}$ implying that $r(t) = \frac{RL}{D(t)}$, and therefore $\frac{r(t)}{\dot{r}(t)} = -\frac{D(t)}{\dot{D}(t)} =$

T. In the following sections we base our statistical formulation of the problem in terms of this geometrical model.

2.3 Statistical Formulation

Suppose that when an object approaches, the quantity T_{thresh} represents the time needed to act in order to avoid collision. For instance, if one is standing at the beginning of a crosswalk and needs to determine if there is sufficient time to cross when a vehicle approaches, then T_{thresh} would be the time required to cross the crosswalk. We formulate the problem in terms of statistical testing:

$$H_0 : T > T_{\text{thresh}} \quad \text{versus} \quad H_1 : T < T_{\text{thresh}}, \tag{2.1}$$

where the null hypothesis H_0 corresponds to the scenario that the time-to-contact is larger than the time needed to get to safety. The alternative H_1 corresponds to the scenario where there is insufficient time to get to safety. This chapter aims to build statistical models and derive conditions under which the above hypothesis testing problem can be reliably solved.

Our goal is to solve a hypothesis testing problem similar to (2.1) only using information at two consecutive time frames, since the response in animals is usually determined almost instantaneously. The retina can be thought of as a two dimensional pixel field with fixed length and width, which we model as $[0, 1]^2$, discretized to an n-by-n array of pixels. The retinal image of the object at time t_0, denoted as S_0, is a disk with radius r_0 centered at the center of the pixel field. Let S_1 denote the expanded retinal image after one time frame. We assume that the expansion is symmetric and the lapse between time frames is short; so S_1 is another disk with radius $r_1 \approx r + dr$ and also centered at the center of the pixel field.

With $|S_0|$ and $|S_1|$ denoting the area of disks S_0 and S_1 respectively, we have

$$T = \frac{r_0}{dr} = \frac{r_0}{r_1 - r_0} = \frac{1}{r_1/r_0 - 1} = \frac{1}{\sqrt{\frac{|S_1|}{|S_0|}} - 1}.$$

For each pixel $p = (\frac{i}{n}, \frac{j}{n})$ in the pixel field where $0 < i, j \leq n$, we observe data

$$Y_t(p) = \mu_n \mathbb{1}\{p \in S_t\} + Z_t(p), \tag{2.2}$$

where Z_t is a $n \times n$ matrix of Gaussian white noise with variance 1, and $\mu_n > 0$ is the signal amplitude. The $n \times n$ grid can be thought of as a discretization of the retina, and $n \to \infty$ corresponds to a finer and finer discretization. Note that the ratio between the radius of the retinal image and the length of the retina is fixed regardless of the value of n. Hence, the areas of the signal at different times $|S_0|$ and $|S_1|$ are assumed to be fixed as $n \to \infty$.

Our objective is to identify the minimal threshold on the signal strength μ_n such that a sequence of tests $\{\phi_n\}_{n=1}^{\infty}$ exists for which the false-negative and false-positive probabilities are both asymptotically zero.

Definition 2.3.1. *We say that a sequence of tests $\{\phi_n\}_{n=1}^{\infty}$ is **asymptotically powerful** if for all $\epsilon > 0$,*

$$\sup_{T \in H_{0,\epsilon}} \mathbb{P}_T\{\phi_n = 1\} + \sup_{T \in H_{1,\epsilon}} \mathbb{P}_T\{\phi_n = 0\} \longrightarrow 0$$

as $n \to \infty$, where $H_{0,\epsilon}$ and $H_{1,\epsilon}$ are defined as

$$H_{0,\epsilon} : T > T_{thresh} + \epsilon \quad \textit{versus} \quad H_{1,\epsilon} : T < T_{thresh} - \epsilon \tag{2.3}$$

with implicit dependence on n through model (2.2).

Remark. *Note that in contrast to the original hypotheses (2.1), the new hypotheses (2.3) are separated by* $\inf\{T \in H_{n,0}\} - \sup\{T \in H_{n,1}\} = 2\epsilon$. *The* 2ϵ *separation is necessary to make asymptotically powerful testing possible. Indeed, if the two hypotheses are defined as in (2.1), then trivially the sum of supremum of the false-negative and false-positive probabilities is at least 1.*

The idea behind constructing the asymptotically powerful test $\{\phi_n\}_{n=1}^{\infty}$ is that we will first construct a consistent estimator $\hat{T}_n$ of T, i.e., $\hat{T}_n$ converges to T in probability, using multiscale scan statistic. Then, we reject the null hypothesis $H_{n,0}$ if $\hat{T}_n < T$, and accept the null hypothesis if $\hat{T}_n \geq T$. From the expression $T = \frac{1}{\sqrt{\frac{|S_1|}{|S_0|}}-1}$ it is intuitive that a "good" estimator of the area of the signal is sufficient for constructing a consistent estimator of T; this is shown formally in the following lemma.

Lemma 2.3.2. *Assume that under the static setting, i.e., when there is only one time frame, we have an estimator* $\hat{S}$ *of the area of the disk* S *such that* $\frac{\hat{S}}{|S|} \xrightarrow{P} c$ *as* $n \to \infty$ *for some absolute constant* $c > 0$. *Then* $\hat{T}_n := \left(\sqrt{\frac{\hat{S}_1}{\hat{S}_0}} - 1\right)^{-1}$ *is a consistent estimator of* T. *Moreover, the sequence of tests* $\{\phi_n\}_{n=1}^{\infty}$ *that reject the null hypothesis* $H_{n,0}$ *if* $\hat{T}_n < T_{\textit{thresh}}$ *and accept otherwise is asymptotically powerful.*

Throughout the chapter we use the following standard notation. For two sequences of numbers a_n and b_n, we write $a_n = o(b_n)$ if $\frac{a_n}{b_n} \to 0$ as $n \to \infty$, and $a_n = \mathcal{O}(b_n)$ if there exists a finite M such that $\frac{a_n}{b_n} < M$ for all n. For a sequence of random variables X_n and a corresponding sequence of numbers a_n, we write $X_n = o_p(a_n)$ if X_n/a_n converges to zero in probability. Similarly, we write $X_n = \mathcal{O}_p(a_n)$ if for any $\epsilon > 0$, there exists a finite M such that $\mathbb{P}(|X_n/a_n| > M) < \epsilon$ for all n. Unless stated otherwise, all limits are with regard to $n \to \infty$.

2.4 Review of Previous Work

The problem of determining whether the signal is present is called *signal detection*. General signal detection is an important problem that arises in many applications, e.g., epidemiology [Neill, 2009] and copy number variation [Jeng et al., 2010]. On the theoretical side, researchers have considered signal detection for geometric objects [Arias-Castro et al., 2005], cluster detection [Arias-Castro et al., 2011], and density inference [Dümbgen and Walther, 2008]. While Gaussian noise is the most common setting, Bernoulli noise [Walther, 2010], Poisson noise [Rivera and Walther, 2013, Kulldorff et al., 2005], and general exponential family [König et al., 2018] models have also been considered.

In our setting of time-to-contact estimation, the signal is always assumed to be present. We are going to use a penalized multiscale scan statistic for identification and estimation of area, which is based on the maximum of an ensemble of local test statistics, penalized and properly scaled. A great deal of prior work exists that investigates scan-type procedures [Kulldorff, 1997, Siegmund and Yakir, 2000, Glaz et al., 2009]. Recently, Sharpnack and Arias-Castro [2016] give exact asymptotics for the scan statistics where the underlying signal is a $d-$dimension hyperrectangle under Gaussian noise. However, this earlier work considers scan statistics with no penalty term. Moreover, the vast majority of the work in which a scan statistic is applied consider detection problems, i.e., testing the existence of signal instead of identification, which is the focus of this work. In contrast, Brown et al. [2008] and Kou [2017] consider identification problems of varying spatial scales.

Chan and Walther [2013] note that the scan statistic is dominated by signals on small spatial scales; this results in a loss of power for detecting large scale signals, which corresponds to our setting. One solution is to use different levels of critical values to test for signals at different scales [Walther, 2010]. Another common solution is to introduce a size penalty, see e.g. Dümbgen and Spokoiny [2001], Datta and Sen [2018]. While both can be

optimal for detecting signals with arbitrary extent [Chan and Walther, 2013], it is easier to adopt the penalized scan statistic for identification problems, as done by Kou [2017].

2.5 One-Dimensional Signal Identification

In this section, we assume the underlying signal is a fixed-length interval with unknown position. Consider the following one-dimensional model at any given time

$$Y(p) = \mu_n \mathbb{1}\{p \in I_n^*\} + Z(p), \quad p \in \left\{\frac{1}{n}, \ldots, \frac{n}{n}\right\} \tag{2.4}$$

where $Z(p)$ are i.i.d. standard normal random variables and the unknown signal is denoted $I_n^* \in \mathcal{I}_n$ where $\mathcal{I}_n = \{(\frac{j}{n}, \frac{k}{n}], 0 \leq j < k \leq n)\}$ is the set of all possible signals. For $I \in \mathcal{I}_n$ define $\mathbf{Y}_n(I) = \frac{\sum_{p \in I} Y(p)}{\sqrt{n|I|}}$ where $|I|$ is the length of the interval I.

Before giving the estimation procedure, we introduce the approximation scheme used by Arias-Castro et al. [2005] for efficient computation. Let n be a dyadic integer $n = 2^J$ and let $\mathcal{J}_n$ denote the collection of all dyadic subintervals

$$I_{j,k} = \left(\frac{k2^j}{n}, \frac{(k+1)2^j}{n}\right]$$

where $0 \leq j \leq J$ and $0 \leq k < n/2^j$. The dyadic intervals are singled out as a special subset of the collection of all intervals because they have cardinality $2n$ rather than $n^2/2$ and yet constitute an $\epsilon-$net for the space of intervals in some special metric. But using only dyadic intervals is not enough, as they cannot approximate all intervals arbitrarily well. Instead, we use dyadic intervals as the "base" and form compound intervals by attaching additional dyadic intervals at the ends. Formally, the interval J_l is an $l-$*level extension* if it can be constructed as follows.

1. Start from a base J_0 which is either a dyadic interval $I_{j,k}$ or the union of two adjacent

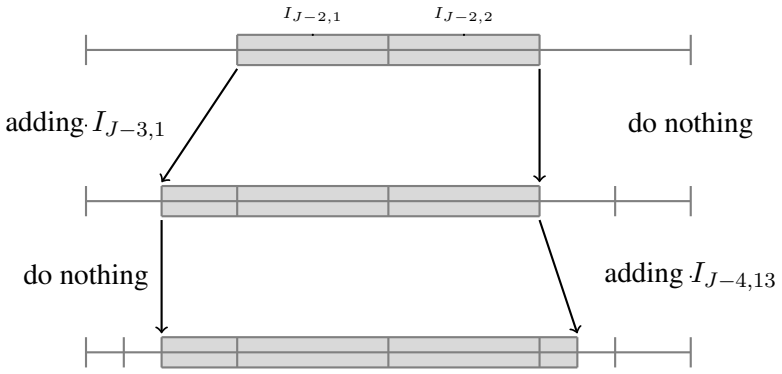

Figure 2.2: Extending dyadic intervals by starting from a dyadic base and appending shorter dyadic intervals at each level.

dyadic intervals $I_{j,k}$ and $I_{j,k+1}$, where k is odd (or else the union of $I_{j,k}$ and $I_{j,k+1}$ is already a dyadic interval).

2. At stage $g = 1, \ldots, l$, extend J_{g-1} to produce J_g by attaching a dyadic interval of length $2^{-g}|I_{j,k}|$ at either or both ends of J_{g-1}, or by doing nothing (so that $J_g = J_{g-1}$).

The result will be a compound interval as depicted in Figure 2.

The collection of all $l-$level extensions of a dyadic interval I will be denoted $\mathcal{J}_l[I]$; the collection of all $l-$level extensions will be denoted $\mathcal{J}_{n,l}$. We have the following facts from Arias-Castro et al. [2005].

Proposition 2.5.1. *For each $I \in \mathcal{I}_n$, there is an interval $J \in \mathcal{J}_{n,l}$ such that*

$$\delta(I, J) = \sqrt{1 - \frac{|I \cap J|}{\sqrt{|I||J|}}} \leq 2^{-l/2}. \tag{2.5}$$

Moreover, $\arg\max_{I \in \mathcal{J}_{n,l}} \mathbf{Y}_n(I)$ *can be computed in* $\mathcal{O}(n4^l)$ *time.*

Here $\delta(I, J)$ is a measure of difference between two intervals I and J that is between 0 and 1. If I and J are exactly the same, $\delta(I, J) = 0$. If $\delta(I, J)$ are disjoint, $\delta(I, J) = 1$. To compute $\arg\max_{I \in \mathcal{J}_{n,l}} \mathbf{Y}_n(I)$, we first evaluate sums for all dyadic intervals I, which

has cardinality $2n$. Superficially, this would take $\mathcal{O}(n^2)$ operations. However, observe that the dyadic sums obey the recursion

$$S[I_{j,k}] = S[I_{j-1,2k}] + S[I_{j-1,2k+1}]$$

for $1 \leq j \leq \log_2(n), 0 \leq k < n/2^j$, where $S[I_{j,k}] = \sum_{p \in I_{j,k}} Y(p)$. Therefore, we start at the finest level by

$$S[I_{0,k}] = Y\left(\frac{k}{n}\right), \ k = 1, \ldots, n$$

and use sums at finer levels to compute sums at coarser levels. Once the $2n$ dyadic sums are computed, we obtain the statistics $\mathbf{Y}_n(I_{j,k}) = 2^{-j/2} S[I_{j,k}]$. Therefore, the overall complexity in this stage is $\mathcal{O}(n)$. After this stage, there are $\mathcal{O}(n)$ dyadic bases and each requires $\mathcal{O}(4^l)$ work; therefore, the overall complexity is $\mathcal{O}(n4^l)$.

The next theorem shows that we can come up with a consistent estimator of $|I_n^*|$ that is also computationally efficient using this approximation scheme.

Theorem 2.5.2. *Assume model* (2.4) *and let the length of the interval signal be denoted* $|I_n^*| \in (0, 1]$. *Suppose the signal size* μ_n *satisfies*

$$\mu_n = \frac{b_n}{\sqrt{n}} \text{ with } b_n \to \infty. \tag{2.6}$$

Let $l = \log_2 \log_2(n)$ *and define*

$$\hat{I}_n = \arg\max_{I \in \mathcal{J}_{n,l}} \left(\mathbf{Y}_n(I) - \sqrt{2 \log \frac{1}{|I|}} \right). \tag{2.7}$$

Then we have $\frac{|\hat{I}_n|}{|I_n^*|} \xrightarrow{P} 1$ *as* $n \to \infty$. *Moreover, there is an algorithm to compute* $|\hat{I}_n|$ *in* $\mathcal{O}(n \log_2^2 n)$ *time.*

Remark. *The condition* (2.6) *matches the detection lower bound for intervals [Chan and*

Walther, 2013]. Since the estimation problem is harder than the detection problem, the condition (2.6) is optimal and cannot be improved.

Remark. *In (2.7) $\hat{I}_n$ is defined as the argmax over a subset of all possible intervals $\mathcal{J}_{n,l}$. Taking the argmax over all possible intervals $\mathcal{I}_n$ would also work; however, implementing the algorithm takes $\mathcal{O}(n^2)$ time.*

We give the proofs of these results in Section 2.9.

2.6 Two-Dimensional Rectangular Signal Identification

In this section, we assume the underlying signal is a fixed-area rectangle with sides parallel to the axes but position unknown. Consider the following two-dimensional model at any given time

$$Y(p) = \mu \mathbb{1}\{p \in I_n^*\} + Z(p), \;\; p = (p_1, p_2), \;\; p_1, p_2 \in \left\{\frac{1}{n}, \ldots, \frac{n}{n}\right\} \tag{2.8}$$

where $Z(p)$ are i.i.d. standard normal random variables and the unknown rectangular signal is given by

$$I_n^* = \left(\frac{j_{n,1}^*}{n}, \frac{k_{n,1}^*}{n}\right] \times \left(\frac{j_{n,2}^*}{n}, \frac{k_{n,2}^*}{n}\right], \;\; 0 \le j_{n,d}^* < k_{n,d}^* \le n \text{ for } d = 1, 2.$$

We denote the area of the rectangle I_n^* by $|I_n^*| = \frac{(k_{n,1}^* - j_{n,1}^*) \times (k_{n,2}^* - j_{n,2}^*)}{n^2}$. Let $\mathcal{I}_n^{(2)}$ be the set of all rectangles on the pixel field with edges parallel to the axes, i.e., $\mathcal{I}_n^{(2)} = \{(\frac{j_{n1}}{n}, \frac{k_{n1}}{n}] \times (\frac{j_{n2}}{n}, \frac{k_{n2}}{n}] \,|\, 0 \le j_{nd} < k_{nd} \le n\}$ for $d = 1, 2$. For $I \in \mathcal{I}_n^{(2)}$, define $\mathbf{Y}_n(I) = \frac{\sum_{p \in I} Y(p)}{\sqrt{n^2 |I|}}$. Note that this definition of $\mathbf{Y}_n(I)$ differs from that for the one-dimensional model in the denominator. Similar to the one dimensional case, we introduce the approximation set $\mathcal{J}_{n,l}^{(2)} = \mathcal{J}_{n,l} \times \mathcal{J}_{n,l} \in \mathcal{I}_n^{(2)}$. We have the following facts from Arias-Castro et al. [2005].

Proposition 2.6.1. *For each $I \in \mathcal{I}_n^{(2)}$, there is an interval $J \in \mathcal{J}_{n,l}^{(2)}$ such that*

$$\delta(I, J) = \sqrt{1 - \frac{|I \cap J|}{\sqrt{|I||J|}}} \leq \epsilon_{n,l}. \tag{2.9}$$

for some $\epsilon_{n,l} \to 0$ as $n, l \to \infty$. Moreover, $\arg\max_{I \in \mathcal{J}_{n,l}^{(2)}} \mathbf{Y}_n(I)$ can be computed in $\mathcal{O}(n^2 4^{2l})$ time.

The identification threshold on μ_n is established in the following theorem.

Theorem 2.6.2. *Assume model (2.8) and suppose the signal size μ_n satisfies*

$$\mu_n = \frac{b_n}{n} \text{ with } b_n \to \infty. \tag{2.10}$$

Let $l = \log_2 \log_2(n)$ and define

$$\hat{I}_n = \arg\max_{I \in \mathcal{J}_{n,l}^{(2)}} \left(\mathbf{Y}_n(I) - \sqrt{2 \log \frac{1}{|I|}} \right). \tag{2.11}$$

Then $\frac{|\hat{I}_n|}{|I_n^*|} \xrightarrow{P} 1$ as $n \to \infty$. Moreover, there is an algorithm to compute $|\hat{I}_n|$ in $\mathcal{O}(n^2 \log_2^4 n)$ time.

Remark. *The results in this section can be extended to $d-$dimensional hyperrectangles as well, but for concreteness we only study the two dimensional case.*

Remark. *The condition (2.10) matches the detection lower bound for rectangles [Kou, 2017]; thus condition (2.10) gives the optimal signal scaling and cannot be improved.*

Remark. *Kou [2017] establishes optimal identification thresholds for rectangular signals belonging to a large class of spatial scales but does not handle our case where $\liminf_{n\to\infty} |I_n^*| > 0$.*

2.7 Two-Dimensional Disk Signal Identification

In this section the underlying signal is assumed to be a fixed-area disk contained in $[0,1]^2$ with unknown location. Consider the following two-dimensional model at any given time

$$Y(p) = \mu \mathbb{1}\{p \in S_n^*\} + Z(p), \;\; p = (p^1, p^2), \;\; p_1, p_2 \in \left\{\frac{1}{n}, \ldots, \frac{n}{n}\right\} \tag{2.12}$$

where $Z(p)$ are i.i.d standard normal random variables and the unknown disk-like signal

$$S_n^* = \left\{(p^1, p^2) \,|\, p_1, p_2 \in \{\frac{1}{n}, \ldots, \frac{n}{n}\}, (p^1 - c^1)^2 + (p^2 - c^2)^2 \leq r^2\right\}.$$

We denote the area of the disk by $|S_n^*| = \pi r^2 \in (0, \frac{\pi}{4}]$. With a slight abuse of notation let $\mathcal{I}_n^{(2)}$ be the set of all *squares* on the pixel field with edges parallel to the axes, i.e., $\mathcal{I}_n^{(2)} = \{\left(\frac{j_n}{n}, \frac{k_n}{n}\right] \times \left(\frac{j_n}{n}, \frac{k_n}{n}\right] \big| \, 0 \leq j_n < k_n \leq n\}$. For $I \in \mathcal{I}_n^{(2)}$, define $\mathbf{Y}_n(I) = \frac{\sum_{p \in I} Y(p)}{\sqrt{n^2 |I|}}$. We use the approximation set $\mathcal{J}_{n,l}^{(2)} = \{I \times I | I \in \mathcal{J}_{n,l}\} \in \mathcal{I}_n^{(2)}$. In this case, Proposition 2.6.1 still holds, and we have the same identification threshold on μ_n as the case of rectangular signals.

Theorem 2.7.1. *Assume model (2.12) and suppose the signal size μ_n satisfies*

$$\mu_n = \frac{b_n}{n} \text{ with } b_n \to \infty. \tag{2.13}$$

Let $l = \log_2 \log_2(n)$ and define

$$\hat{I}_n = \arg\max_{I \in \mathcal{J}_{n,l}^{(2)}} \left(\mathbf{Y}_n(I) - \sqrt{2 \log \frac{1}{|I|}} \right). \tag{2.14}$$

Then we have $\frac{|\hat{I}_n|}{|S_n^*|} \xrightarrow{P} c$ as $n \to \infty$ for some absolute constant $c > 0$. Moreover, there is an algorithm to compute $|\hat{I}_n|$ in $\mathcal{O}(n^2 \log_2^4 n)$ time.

Theorem 2.7.1 gives a sufficient condition on μ_n that, once satisfied, enables asymptotically powerful tests for the two-dimensional model. The following theorem shows that condition (2.13) is also necessary for constructing asymptotically powerful tests.

Theorem 2.7.2. *Assume that the retinal image of the object is a disk with area* $(0, \frac{\pi}{4}]$ *at time* t_0 *under model* (2.12)*. If condition* (2.13) *does not hold, then for any sequence of tests* $\{\phi_n\}$ *there is some* $\epsilon > 0$ *and* $\beta > 0$ *such that*

$$\sup_{T > T_{thresh} + \epsilon} \mathbb{P}_T\{\phi_n = 1\} + \sup_{T < T_{thresh} - \epsilon} \mathbb{P}_T\{\phi_n = 0\} > \beta.$$

2.8 Simulation Study

Theorem 2.7.1 combined with Lemma 2.3.2 yields efficient testing procedure. For comparison, we also consider the inefficient counterpart where the estimation of area is done by exhaustive search over all possible squares $\mathcal{I}_n^{(2)}$ instead of the approximation set $\mathcal{J}_{n,l}^{(2)}$ in (2.14). The efficacy of both procedure is tested in this section through extensive simulation study.

We pick $T_{\text{thresh}} \in \{3, 5, 10\}$ and $\epsilon \in \{\frac{T}{2}, \frac{T}{5}\}$. We pick $n \in \{8, 16, 32, 64, 128\}$ to run and $n \in \{8, 16, 32, 64, 128, 256, 512\}$ to run the efficient algorithm using the approximation scheme described in Section 2.5. We allow larger n for the efficient algorithm both because the efficient one runs much faster so can afford large n and also because the approximation set introduced by the efficient algorithm requires larger n to be sufficiently accurate. For concreteness, the time it took to run inefficient algorithm on $n = 128$ is more than that to run the efficient algorithm on $n = 512$.

The signal strength is $\mu_n = \frac{b_n}{n}$ where $b_n \in \{16 \log_2(\log_2(n)), 128 \log_2(\log_2(n)), 1024 \log_2(\log_2(n))\}$. The choice of this particular form of $b_n = A \log_2(\log_2(n))$ is to stay below the minimum threshold $\Omega(\log n)$ of the unpenalized scan statistic [Chan and Walther, 2013] while ensuring that $b_n \to \infty$ at the same time. To approximate

$\sup_{T>T_{\text{thresh}}+\epsilon} \mathbb{P}_T\{\phi_n = 1\}$, we generate signals so that $T = T_{\text{thresh}} + \epsilon$, and compute the empirical fraction $\mathbb{P}_{T=T_{\text{thresh}}+\epsilon}(\hat{T}_n < T_{\text{thresh}})$ among 100 repetitions. An approximation of $\sup_{T<T_{\text{thresh}}-\epsilon} \mathbb{P}_T\{\phi_n = 0\}$ is computed analogously using the same number of repetitions. For each T, we generate the signal by first generating r_1, the radius of retinal object at time t_1, uniformly at random in the range $[0.05, 0.5]$. Then, using the identity $r_1 = r_0 + dr = r_0 + \frac{r_0}{T}$, r_0, the radius of retinal object at time t_0, is computed. After r_0 and r_1 are fixed, we generate the $n \times n$ pixel matrix at time t_0 (t_1) point by having all pixels with distance smaller than r_0 (r_1) following Gaussian distribution with mean μ_n and variance 1, while the rest of the pixels following Gaussian distribution with mean 0 and variance 1.

Figures 2.3, 2.4 and 2.5 illustrate the behavior of testing error $\sup_{T>T_{\text{thresh}}+\epsilon} \mathbb{P}_T\{\phi_n = 1\} + \sup_{T<T_{\text{thresh}}-\epsilon} \mathbb{P}_T\{\phi_n = 0\}$ against $\log_2(n)$ with varying $T_{\text{thresh}}, \epsilon$, and μ_n. A few observations can be made. Firstly, the overall trend that the testing error decreases as n increases is clear regardless of our choice of parameter $T_{\text{thresh}}, \epsilon$, and μ_n. Secondly, for a fixed T_{thresh}, smaller ϵ corresponds to a harder problem so leads to larger testing error. Moreover, when keeping ϵ as a fixed fraction of T_{thresh}, testing error still increases as T_{thresh} increases. This might be because under our setting $T_{\text{thresh}} - \epsilon$ and $T_{\text{thresh}} + \epsilon$ are increasing, so the ratio $\frac{r_1}{r_0} = 1 + \frac{1}{T}$ is decreasing and the problem becomes harder in the sense that it requires larger n to make the testing error sufficiently small. In addition, we observe that stronger signal, which implies easier problem, leads to smaller testing error. This effect is particularly noticeable when comparing the first two figures ($b_n = 16\log_2(\log_2(n))$ versus $b_n = 128\log_2(\log_2(n))$), but is less noticeable when comparing the last two figures ($b_n = 128\log_2(\log_2(n))$ versus $b_n = 1024\log_2(\log_2(n))$). Lastly, restricting to the easier scenario where $\epsilon = \frac{T_{\text{thresh}}}{2}$, we see that the efficient procedure could attain comparable performance in shorter time, compared to the more accurate but inefficient counterpart.

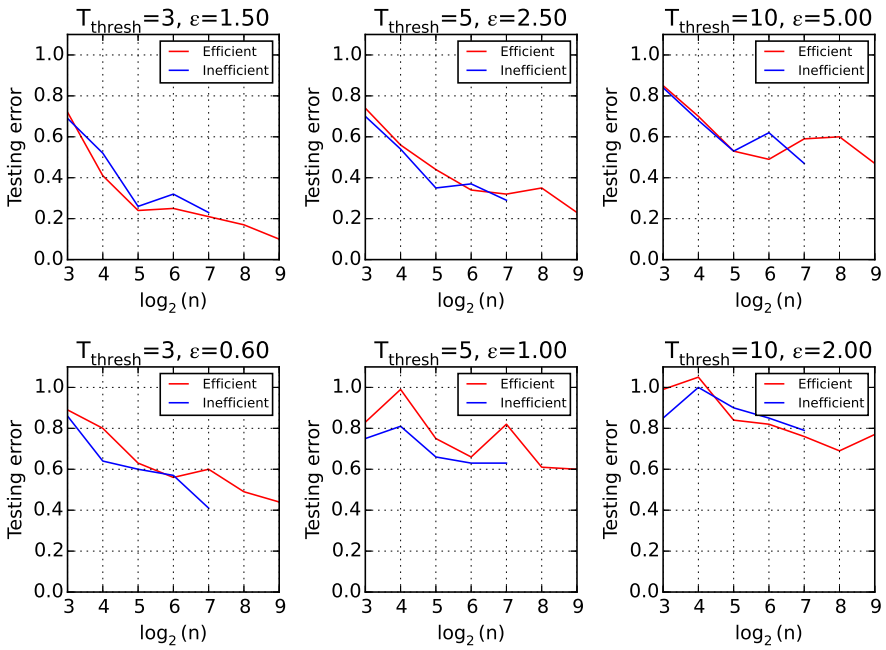

Figure 2.3: Testing error against $log_2(n)$ when $b_n = 16\log_2(\log_2(n))$.

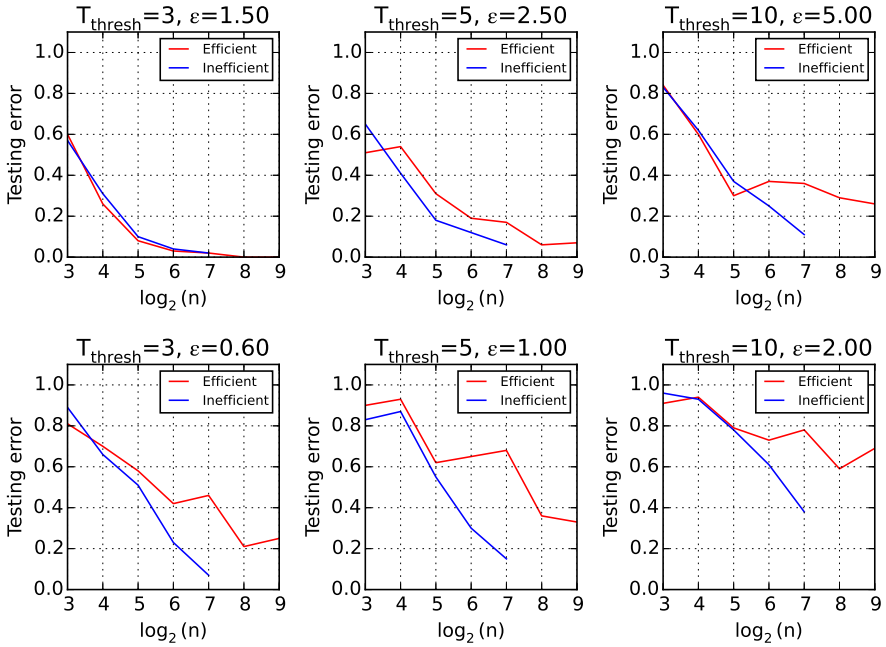

Figure 2.4: Testing error against $log_2(n)$ when $b_n = 128\log_2(\log_2(n))$.

2.9 Proof Outline

In this section we sketch the structure of the proofs, beginning with the one-dimensional case.

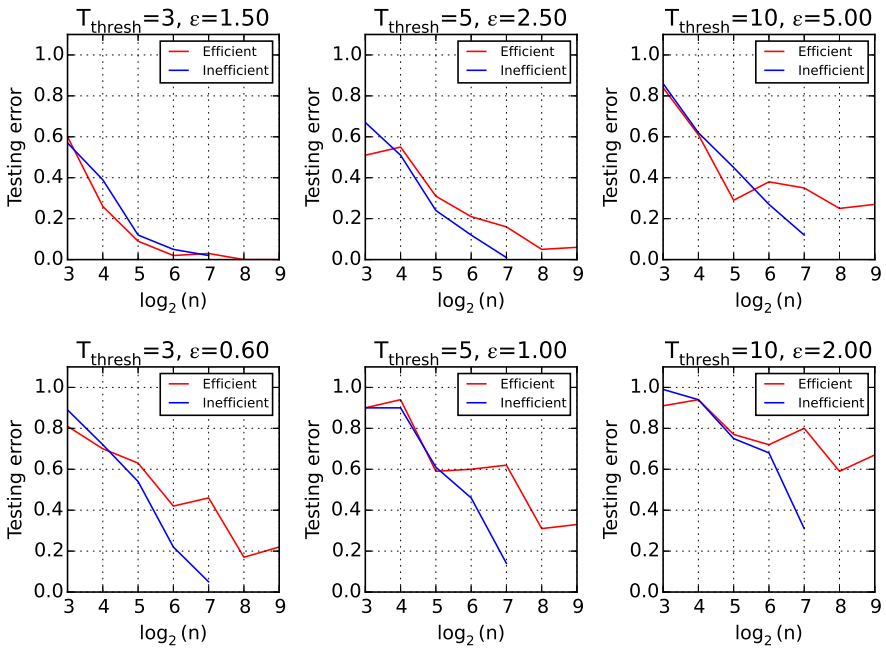

Figure 2.5: Testing error against $log_2(n)$ when $b_n = 1024 \log_2(\log_2(n))$.

Proof of Lemma 2.3.2. By assumption we can construct two estimators $\hat{s}_0$ and $\hat{s}_1$ such that $\frac{\hat{s}_0}{|S_0|} \xrightarrow{P} c$ and $\frac{\hat{s}_1}{|S_1|} \xrightarrow{P} c$. As a result $\frac{\hat{s}_1}{\hat{s}_0} \xrightarrow{P} \frac{|S_1|}{|S_0|}$. Therefore,

$$\hat{T}_n = \frac{1}{\sqrt{\frac{\hat{s}_1}{\hat{s}_0} - 1}} \xrightarrow{P} \frac{1}{\sqrt{\frac{|S_1|}{|S_0|} - 1}} = T,$$

i.e., $\hat{T}_n$ is a consistent estimator of T. This implies that

$$\sup_{T \in H_{n,0}} \mathbb{P}_T\{\phi_n \text{ rejects } H_{n,0}\} = \sup_{T > T_{\text{thresh}} + \epsilon} \mathbb{P}_T\{\hat{T}_n < T\} \tag{2.15}$$

$$\leq \sup_{T > T_{\text{thresh}} + \epsilon} \mathbb{P}\{\hat{T}_n - T < -\epsilon\} \longrightarrow 0$$

and

$$\sup_{T \in H_{n,1}} \mathbb{P}_T\{\phi_n \text{ accepts } H_{n,0}\} = \sup_{T < T_{\text{thresh}} - \epsilon} \mathbb{P}_T\{\hat{T}_n \geq T\} \tag{2.16}$$

$$\leq \sup_{T<T_{\text{thresh}}-\epsilon} \mathbb{P}\{\hat{T}_n - T \geq \epsilon\} \longrightarrow 0.$$

Combing the above two equations finishes the proof. □

Proof of Theorem 2.5.2. Before we prove this theorem, we introduce the following key lemma, the proof of which is very similar to the proof of Lemma 1 in [Chan and Walther, 2013].

Lemma 2.9.1. *Define* $\mathbf{Z}_n(I) := \frac{\sum_{p\in I} Z(p)}{\sqrt{n|I|}}$. *Then*

$$\max_{I\in\mathcal{I}_n} \left(|\mathbf{Z}_n(I)| - \sqrt{2\log\frac{1}{|I|}} \right) < \infty$$

almost surely for all n.

Proof of Lemma 2.9.1. Observe that for any particular $I \in \mathcal{I}_n$, $\mathbf{Z}_n(I) \sim N(0,1)$. Writing W for Brownian motion, we have that

$$\begin{aligned}
&\max_{I\in\mathcal{I}_n} \left(|\mathbf{Z}_n(I)| - \sqrt{2\log\frac{1}{|I|}} \right) \\
&\overset{d}{=} \max_{0\leq j/n<k/n\leq 1} \left(\frac{|W(k/n)-W(j/n)|}{\sqrt{k/n-j/n}} - \sqrt{2\log\frac{1}{k/n-j/n}} \right) \\
&\leq \max_{0\leq s<t\leq 1} \left(\frac{|W(t)-W(s)|}{\sqrt{t-s}} - \sqrt{2\log\frac{1}{t-s}} \right) \\
&\leq \max_{0\leq s<t\leq 1} \left(\frac{|W(t)-W(s)|}{\sqrt{t-s}} - \sqrt{2\log\frac{1}{t-s}} \right) / D(t-s) \\
&\overset{d}{=}: L' < \infty \text{ a.s.}
\end{aligned}$$

where $D(r) := \frac{\log\log(e^e/r)}{\sqrt{\log(e/r)}} \leq 1$ for all $r \in (0,1]$. The last line is from Sec 6.1 in Dümbgen and Spokoiny [2001]. □

Equipped with this lemma, we are ready to prove Theorem 2.5.2; the complexity of

the algorithm trivially follows from Proposition 2.5.1. To show that $\frac{|\hat{I}_n|}{|I_n^*|} \xrightarrow{P} 1$ as $n \to \infty$, we need to show that for any $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{|\hat{I}_n|}{|I_n^*|} - 1\right| > \epsilon\right) \longrightarrow 0$$

as $n \to \infty$. Denote $\mathbf{X}_n(I) := \mathbf{Y}_n(I) - \sqrt{2\log\frac{1}{|I|}}$. By (2.5), there is $\tilde{I} \in \mathcal{J}_{n,l}$ such that

$$\delta(\tilde{I}, I_n^*) = \sqrt{1 - \frac{|\tilde{I} \cap I_n^*|}{\sqrt{|\tilde{I}||I_n^*|}}} \leq 2^{-l/2} = \frac{1}{\sqrt{\log_2 n}}.$$

Therefore, there is $\tilde{I} \in \mathcal{J}_{n,l}$ such that

$$\frac{|\tilde{I} \cap I_n^*|}{\sqrt{|\tilde{I}||I_n^*|}} \geq 1 - \frac{1}{\log_2 n}. \tag{2.17}$$

Observe that for any $I \in \mathcal{I}_n$, if $\left|\frac{|I|}{|I_n^*|} - 1\right| > \epsilon$, we have

$$\frac{|I \cap I_n^*|}{\sqrt{|I||I_n^*|}} \leq \max\left\{\frac{1}{\sqrt{1+\epsilon}}, \sqrt{1-\epsilon}\right\} := 1 - d_\epsilon < 1. \tag{2.18}$$

Hence, there is $\tilde{I} \in \mathcal{J}_{n,l}$ such that

$$\left|\frac{|\tilde{I}|}{|I_n^*|} - 1\right| \leq \max\left\{1 - \left(1 - \frac{1}{\log_2 n}\right)^2, \frac{1}{\left(1 - 1/\log_2 n\right)^2} - 1\right\} \longrightarrow 0. \tag{2.19}$$

In particular, assuming that n is sufficiently large, we have $\left|\frac{|\tilde{I}|}{|I_n^*|} - 1\right| < \epsilon$. Define the event

$$\mathcal{K}_{n,l} := \left\{I \in \mathcal{J}_{n,l} : \left|\frac{|I|}{|I_n^*|} - 1\right| > \epsilon\right\}.$$

Then we have the following sequence of probability bounds:

$$
\begin{aligned}
&\mathbb{P}\left(\left|\frac{|\hat{I}_n|}{|I_n^*|}-1\right|>\epsilon\right)\\
&\le \mathbb{P}\left(\max_{I\in\mathcal{K}_{n,l}} \mathbf{X}_n(I)\ge \mathbf{X}_n(\tilde{I})\right)\\
&= \mathbb{P}\Bigg(\max_{I\in\mathcal{K}_{n,l}}\left(\mathbf{Z}_n(I)+\frac{n|I\cap I_n^*|}{\sqrt{n|I|}}\mu_n-\sqrt{2\log\frac{1}{|I|}}\right)\ge\\
&\qquad\qquad \mathbf{Z}_n(\tilde{I})+\frac{n|\tilde{I}\cap I_n^*|}{\sqrt{n|\tilde{I}|}}\mu_n-\sqrt{2\log\frac{1}{|\tilde{I}|}}\Bigg)\\
&= \mathbb{P}\Bigg(\max_{I\in\mathcal{K}_{n,l}}\left(\mathbf{Z}_n(I)+\frac{|I\cap I_n^*|}{\sqrt{|I||I_n^*|}}\sqrt{r}b_n-\sqrt{2\log\frac{1}{|I|}}\right)\ge\\
&\qquad\qquad \mathbf{Z}_n(\tilde{I})+\frac{|\tilde{I}\cap I_n^*|}{\sqrt{|\tilde{I}||I_n^*|}}\sqrt{r}b_n-\sqrt{2\log\frac{1}{|\tilde{I}|}}\Bigg)\\
&\le \mathbb{P}\Bigg(\max_{I\in\mathcal{K}_{n,l}}\left(\mathbf{Z}_n(I)+(1-d_\epsilon)\sqrt{r}b_n-\sqrt{2\log\frac{1}{|I|}}\right)\ge\\
&\qquad\qquad \mathbf{Z}_n(\tilde{I})+\big(1-\frac{1}{\log_2 n}\big)\sqrt{r}b_n-\sqrt{2\log\frac{1}{|\tilde{I}|}}\Bigg)\\
&= \mathbb{P}\left(\max_{I\in\mathcal{K}_{n,l}}\left(\mathbf{Z}_n(I)-\sqrt{2\log\frac{1}{|I|}}\right)\ge \mathbf{Z}_n(\tilde{I})+\big(d_\epsilon-\frac{1}{\log_2 n}\big)\sqrt{r}b_n+\sqrt{2\log\frac{1}{|\tilde{I}|}}\right).
\end{aligned}
$$

The last inequality is from (2.17) and (2.18). By Lemma 2.9.1,
$\max_{I\in\mathcal{K}_{n,l}}\left(\mathbf{Z}_n(I)-\sqrt{2\log\frac{1}{|I|}}\right)$ is $\mathcal{O}_p(1)$. Moreover, $\mathbf{Z}_n(\tilde{I})\sim N(0,1)$,
$\big(d_\epsilon-\frac{1}{\log_2 n}\big)\sqrt{r}b_n\to\infty$, and from (2.19) we know $\sqrt{2\log\frac{1}{|\tilde{I}|}}\to\sqrt{2\log\frac{1}{|I_n^*|}}=\sqrt{2\log\frac{1}{r}}$.
Therefore, the last probability converges to zero, completing the proof of Theorem 2.5.2.

□

The proofs of the two-dimensional cases are given in the next section.

2.10 Additional Proofs

The proof of Theorem 2.6.2 is almost identical to the proof of Theorem 2.5.2 so the bulk of the proof is omitted. However, we need to show a two dimensional version of Lemma 2.9.1.

Lemma 2.10.1. *Define* $\mathbf{Z}_n(I) := \frac{\sum_{p\in I} Z(p)}{\sqrt{n^2|I|}}$. *Then*

$$\max_{I\in\mathcal{I}_n^{(2)}} \left(|\mathbf{Z}_n(I)| - \sqrt{2\log\frac{1}{|I|}} \right) < \infty$$

is uniformly bounded almost surely.

Proof of Lemma 2.10.1. In proving Lemma 2.10.1 we shall make use of the following lemma from Datta and Sen [2018].

Lemma 2.10.2. *[Datta and Sen, 2018] Let* $\psi(x) = \mathbb{1}\{x \in [-1,1]^2\}$. *For a vector* $h = (h_1, h_2)$ *with* $h_1, h_2 > 0$ *let* $A_h = \{t \in \mathbb{R}^2 : h_i \le t_i \le 1 - h_i \text{ for } i = 1, 2\}$. *For* $t \in A_h$, *let*

$$\psi_{t,h}(x) = \psi(\frac{x_1 - t_1}{h_1}, \ldots, \frac{x_d - t_d}{h_d})$$

for $x = (x_1, x_2) \in [0,1]^2$ *and define*

$$\hat{\Psi}(t,h) = \frac{1}{2\sqrt{h_1 h_2}} \int_{[0,1]^2} \psi_{t,h}(x) dW(x)$$

where W *is the 2-dimensional Brownian sheet as defined in Definition 6.1 of Datta and Sen [2018]. Then,*

$$T(W,\psi) = \sup_{h\in(0,1/2]^2} \sup_{t\in A_h} \frac{|\hat{\Psi}(t,h)| - \Gamma(4h_1h_2)}{D(4h_1h_2)} < \infty \text{ a.s.}$$

where $\Gamma(r) := (2\log(1/r))^{1/2}$ *and* $D(r) := \frac{\log\log(e^e/r)}{\sqrt{\log(e/r)}}$.

By Lemma 2.10.2, for any $0 \le r_1 < r_2 \le 1$ and $0 \le t_1 < t_2 \le 1$, if we plug in $h = (\frac{r_2 - r_1}{2}, \frac{t_2 - t_1}{2}) \in (0, 1/2]^2$ and $u = (\frac{r_2 + r_1}{2}, \frac{t_2 + t_1}{2}) \in A_h$, then for $x = (x_1, x_2) \in [0,1]^2$,

$$\psi_{u,h}(x) = \psi(\frac{x_1 - u_1}{h_1}, \frac{x_2 - u_2}{h_2}) = \mathbb{1}\{x \in [r_1, r_2] \times [t_1, t_2]\}.$$

Therefore,

$$|\hat{\Psi}(u,h)| = \frac{\left| \int_{[0,1]^2} \mathbb{1}\{x \in [r_1, r_2] \times [t_1, t_2]\} dW(x) \right|}{\sqrt{(r_2 - r_1)(t_2 - t_1)}}.$$

Hence we have that

$$\begin{aligned}
&\max_{\substack{0 \le r_1 < r_2 \le 1 \\ 0 \le t_1 < t_2 \le 1}} \left(\frac{| \int_{[0,1]^2} \mathbb{1}\{x \in [r_1, r_2] \times [t_1, t_2]\} dW(x)|}{\sqrt{(r_2 - r_1)(t_2 - t_1)}} - \sqrt{2 \log \frac{1}{(r_2 - r_1)(t_2 - t_1)}} \right) \\
&\quad \le \sup_{h \in (0,1/2]^2} \sup_{t \in A_h} \left(|\hat{\Psi}(t,h)| - \Gamma(4h_1 h_2) \right).
\end{aligned}$$

Now, writing j_1, j_2, k_1, k_2 for integer indices and $\stackrel{d}{=}$ for equal in distribution:

$$\begin{aligned}
&\max_{I \in \mathcal{I}_n^{(2)}} \left(|\mathbf{Z}_n(I)| - \sqrt{2 \log \frac{1}{|I|}} \right) \\
&\quad \stackrel{d}{=} \max_{\substack{0 \le j_1/n < j_2/n \le 1 \\ 0 \le k_1/n < k_2/n \le 1}} \left(\frac{\left| \int_{[0,1]^2} \mathbb{1}\{x \in [\frac{j_1}{n}, \frac{j_2}{n}] \times [\frac{k_1}{n}, \frac{k_2}{n}]\} dW(x) \right|}{\sqrt{(j_2 - j_1)(k_2 - k_1)/n^2}} - \right. \\
&\qquad\qquad \left. \sqrt{2 \log \frac{1}{(j_2 - j_1)(k_2 - k_1)/n^2}} \right) \\
&\quad \le \max_{\substack{0 \le r_1 < r_2 \le 1 \\ 0 \le t_1 < t_2 \le 1}} \left(\frac{\left| \int_{[0,1]^2} \mathbb{1}\{x \in [r_1, r_2] \times [t_1, t_2]\} dW(x) \right|}{\sqrt{(r_2 - r_1)(t_2 - t_1)}} - \sqrt{2 \log \frac{1}{(r_2 - r_1)(t_2 - t_1)}} \right) \\
&\quad \le \sup_{h \in (0,1/2]^2} \sup_{t \in A_h} \left(|\hat{\Psi}(t,h)| - \Gamma(4h_1 h_2) \right) \\
&\quad \le \sup_{h \in (0,1/2]^2} \sup_{t \in A_h} \frac{|\hat{\Psi}(t,h)| - \Gamma(4h_1 h_2)}{D(4h_1 h_2)} < \infty \text{ a.s.}
\end{aligned}$$

The first equality is from the property of Brownian sheet W that if $g \in L_2([0,1]^2)$ then $\int_{[0,1]^2} g(t)dW(t) \sim N(0, \|g\|^2)$ where $\|g\|^2 = \int_{[0,1]^2} g^2(x)dx$. □

Proof of Theorem 2.7.1. We first prove the following proposition that is instrumental in proving the theorem.

Proposition 2.10.3. *Let S^* be a disk on $\mathbb{R}^2$. Denote the set of all squares on $\mathbb{R}^2$ by $\mathcal{I}^2$. The function $\frac{|I \cap S^*|}{\sqrt{|I||S^*|}}$ has a unique maximizer among all $I \in \mathcal{I}^2$, denoted as I^*. Moreover, define $K_\epsilon = \{I \in \mathcal{I}^2 : \left|\frac{|I|}{|I^*|} - 1\right| > \epsilon\}$. For any $I \in K_\epsilon$, we have*

$$\frac{|I^* \cap S^*|}{\sqrt{|I^*||S^*|}} - \frac{|I \cap S^*|}{\sqrt{|I||S^*|}} \geq d_\epsilon$$

for some $d_\epsilon > 0$, where $|\cdot|$ denotes the Lebesgue measure of a set.

Proof of Proposition 2.10.3. Since $\frac{|I \cap S^*|}{\sqrt{|I||S^*|}}$ is invariant to simultaneous scaling or translation of I and S^*, we can without loss of generality assume that $S^* = \{(x, y) : x^2 + y^2 \leq 2\}$. For any square I with fixed area, observe that $\frac{|I \cap S^*|}{\sqrt{|I||S^*|}}$ is maximized when I is centered at the center of the disk S^*, which is the origin. Therefore, define

$$I_\delta = \{(x, y) : -\delta \leq x, y \leq \delta\}$$

and

$$f(\delta) = \frac{|I_\delta \cap S^*|}{\sqrt{|I_\delta||S^*|}}.$$

Given that $f(\delta)$ is continuous, to show Proposition 2.10.3 it suffices to show that $f(\delta)$ is monotone increasing when $0 < \delta < \delta^*$, and monotone decreasing when $\delta > \delta^*$ for some $\delta^* \in (1, \sqrt{2})$. When $\delta \leq 1$, $f(\delta) = \sqrt{\frac{|I_\delta|}{|S^*|}} = \frac{2\delta}{\sqrt{2\pi}}$ is monotone increasing in δ. When $\delta \geq \sqrt{2}$, $f(\delta) = \frac{|S^*|}{\sqrt{|I_\delta||S^*|}} = \frac{\sqrt{2\pi}}{2\delta}$ is monotone decreasing in δ. Therefore, it further suffices to show that $f(\delta)$ is monotone increasing when $\delta \in (1, \delta^*)$ and monotone increasing when $\delta \in (\delta^*, \sqrt{2})$.

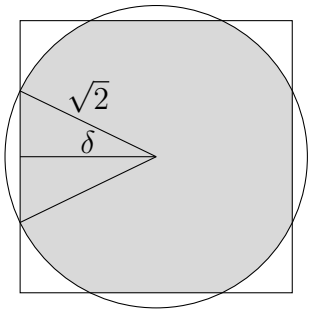

Figure 3. Intersection of S^* and I_δ for some $\delta \in (1, \sqrt{2})$.

From Figure 3 we can derive that

$$f(\delta) = \frac{|I_\delta \cap S^*|}{\sqrt{|I_\delta||S^*|}} = \frac{2\pi - 4(\arccos(\frac{\delta}{\sqrt{2}}) - \delta\sqrt{2-\delta^2})}{2\delta\sqrt{2\pi}}.$$

Hence, it suffices to show that

$$g(\delta) = \sqrt{2\pi} f(\delta) = \frac{\pi}{\delta} - \frac{4\arccos(\frac{\delta}{\sqrt{2}})}{\delta} + 2\sqrt{2-\delta^2}$$

is monotone increasing when $\delta \in (1, \delta^*)$ and monotone decreasing when $\delta \in (\delta^*, \sqrt{2})$ for some $\delta^* \in (1, \sqrt{2})$. We have

$$g'(\delta) = \frac{4}{\delta\sqrt{2-\delta^2}} - \frac{\pi}{\delta^2} + \frac{4\arccos(\frac{\delta}{\sqrt{2}})}{\delta^2} - \frac{2\delta}{\sqrt{2-\delta^2}}.$$

Observe that $g'(1) = 2$ and $g'(1.4) \approx -1 < 0$. It further suffices to show that $g''(\delta) < 0$ when $\delta \in (1, \sqrt{2})$. We then have

$$\begin{aligned} g''(\delta) &= \frac{2(\delta^2-1)}{\delta^2(2-\delta^2)^{3/2}} + \frac{2\pi}{\delta^3} - \frac{4\left(\delta + 2\sqrt{2-\delta^2}\arccos(\frac{\delta}{\sqrt{2}})\right)}{\delta^3\sqrt{2-\delta^2}} - \frac{4}{(2-\delta^2)^{3/2}} \\ &= \frac{2\delta(\delta^2-1) + 2\pi(2-\delta^2)^{3/2}}{\delta^3(2-\delta^2)^{3/2}} - \frac{4(2-\delta^2)\left(\delta + 2\sqrt{2-\delta^2}\arccos(\frac{\delta}{\sqrt{2}})\right) - 4\delta^3}{\delta^3(2-\delta^2)^{3/2}} \\ &= 2 \times \frac{\delta^3 - 5\delta + \pi(2-\delta^2)^{3/2} - 8\sqrt{2-\delta^2}\arccos(\frac{\delta}{\sqrt{2}})}{\delta^3(2-\delta^2)^{3/2}} + \frac{4\delta^2\sqrt{2-\delta^2}\arccos(\frac{\delta}{\sqrt{2}})}{\delta^3(2-\delta^2)^{3/2}} \end{aligned}$$

$$
\begin{aligned}
&\leq 2 \times \frac{\delta^3 - 5\delta + \pi}{\delta^3(2-\delta^2)^{3/2}} \\
&\leq 2 \times \frac{\pi - 4}{\delta^3(2-\delta^2)^{3/2}} \\
&< 0.
\end{aligned}
$$

This concludes the proof of Proposition 2.10.3. □

Now we turn to the proof of Theorem 2.7.1. Let I^* denote the unique maximizer of the function $\frac{|I \cap S_n^*|}{\sqrt{|I||S_n^*|}}$ over all squares on $[0,1]^2$. Let I_n^* be the best approximation of I^* in $\mathcal{I}_n^{(2)}$. Trivially we have

$$
|I_n^*| = |I^*| + o(1) \text{ and } \frac{|I_n^* \cap S_n^*|}{\sqrt{|I_n^*||S_n^*|}} = \frac{|I^* \cap S_n^*|}{\sqrt{|I^*||S_n^*|}} + o(1) \tag{2.20}
$$

Define $c := \frac{|I^*|}{|S_n^*|}$. To show that $\frac{|\hat{I}_n|}{|S_n^*|} \xrightarrow{P} c$, it suffices to show that $\frac{|\hat{I}_n|}{|I^*|} \xrightarrow{P} 1$. So we need to show that for any $\epsilon > 0$,

$$
\mathbb{P}\left(\left|\frac{|\hat{I}_n|}{|I^*|} - 1\right| > \epsilon\right) \to 0 \text{ as } n \to \infty.
$$

By (2.9), there is $\tilde{I} \in \mathcal{J}_{n,l}^{(2)}$ such that

$$
\delta(\tilde{I}, I_n^*) = \sqrt{1 - \frac{|\tilde{I} \cap I_n^*|}{\sqrt{|\tilde{I}||I_n^*|}}} \leq \epsilon_{n,l} \to 0.
$$

Therefore, there is $\tilde{I} \in \mathcal{J}_{n,l}^{(2)}$ such that

$$
\frac{|\tilde{I} \cap I_n^*|}{\sqrt{|\tilde{I}||I_n^*|}} \geq 1 - \epsilon_{n,l}^2 \to 1. \tag{2.21}
$$

Hence, there is $\tilde{I} \in \mathcal{J}_{n,l}^{(2)}$ such that

$$|\tilde{I}| = |I_n^*| + o(1) \text{ and } |\tilde{I} \cap I_n^*| = |I_n^*| + o(1).$$

This implies that

$$\frac{|\tilde{I} \cap S_n^*|}{\sqrt{|\tilde{I}||S_n^*|}} = \frac{|I_n^* \cap S_n^*|}{\sqrt{|I_n^*||S_n^*|}} + o(1). \tag{2.22}$$

Moreover, assuming that n is sufficiently large, we have $\left|\frac{|\tilde{I}|}{|I^*|} - 1\right| < \epsilon$. Define

$$\mathcal{K}_{n,l} = \{I \in \mathcal{J}_{n,l}^{(2)} : \left|\frac{|I|}{|I^*|} - 1\right| > \epsilon\}.$$

Denote $\mathbf{X}_n(I) := \mathbf{Y}_n(I) - \sqrt{2\log\frac{1}{|I|}}$ and let $N(S)$ denote the number of $n \times n$ grid points inside a set S. Define the shorthand $m(I) = \sqrt{2\log\frac{1}{|I|}}$. We then have

$$\begin{aligned}
&\mathbb{P}\Big(\Big|\frac{|\hat{I}_n|}{|I^*|} - 1\Big| > \epsilon\Big) \le \mathbb{P}\Big(\max_{I \in \mathcal{K}_{n,l}} \mathbf{X}_n(I) \ge \mathbf{X}_n(\tilde{I})\Big) \\
&= \mathbb{P}\left(\max_{I \in \mathcal{K}_{n,l}} \left(\mathbf{Z}_n(I) + \frac{N(I \cap S^*)}{n\sqrt{|I|}}\mu_n - \sqrt{2\log\frac{1}{|I|}}\right)\right. \\
&\qquad\qquad \left. \ge \mathbf{Z}_n(\tilde{I}) + \frac{N(\tilde{I} \cap S^*)}{n\sqrt{|\tilde{I}|}}\mu_n - \sqrt{2\log\frac{1}{|\tilde{I}|}}\right) \\
&\le \mathbb{P}\left(\max_{I \in \mathcal{K}_{n,l}} \left(\mathbf{Z}_n(I) + \frac{n^2|I \cap S^*| + \mathcal{O}\left(n^{2/3}\right)}{n\sqrt{|I|}}\mu_n - m(I)\right)\right. \\
&\qquad\qquad \left. \ge \mathbf{Z}_n(\tilde{I}) + \frac{n^2|\tilde{I} \cap S^*| - \mathcal{O}\left(n^{2/3}\right)}{n\sqrt{|\tilde{I}|}}\mu_n - m(\tilde{I})\right) \\
&= \mathbb{P}\left(\max_{I \in \mathcal{K}_{n,l}} \left(\mathbf{Z}_n(I) + \frac{|I \cap S^*|}{\sqrt{|I|}}b_n + o(1)b_n - m(I)\right)\right. \\
&\qquad\qquad \left. \ge \mathbf{Z}_n(\tilde{I}) + \frac{|\tilde{I} \cap S^*|}{\sqrt{|\tilde{I}|}}b_n - o(1)b_n - m(\tilde{I})\right)
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{P}\left(\max_{I\in\mathcal{K}_{n,l}}\left(\mathbf{Z}_n(I) + \frac{|I\cap S^*|}{\sqrt{|I||S^*|}}\sqrt{\pi r^2}b_n + o(1)b_n - m(I)\right)\right. \\
&\qquad\qquad \left. \geq \mathbf{Z}_n(\tilde{I}) + \frac{|\tilde{I}\cap S^*|}{\sqrt{|\tilde{I}||S^*|}}\sqrt{\pi r^2}b_n - o(1)b_n - m(\tilde{I})\right) \\
&\leq \mathbb{P}\left(\max_{I\in\mathcal{K}_{n,l}}\left(\mathbf{Z}_n(I) + \frac{|I\cap S^*|}{\sqrt{|I||S^*|}}\sqrt{\pi r^2}b_n - m(I)\right)\right. \\
&\qquad\qquad \left. \geq \mathbf{Z}_n(I^*) + \left(\frac{|I^*\cap S^*|}{\sqrt{|I^*||S^*|}} - o(1)\right)\sqrt{\pi r^2}b_n - m(\tilde{I})\right) \\
&\leq \mathbb{P}\left(\max_{I\in\mathcal{K}_{n,l}}\left(\mathbf{Z}_n(I) - \sqrt{2\log\frac{1}{|I|}}\right) \geq \mathbf{Z}_n(I^*) + \left(d_\epsilon\sqrt{\pi r^2} - o(1)\right)b_n - \sqrt{2\log\frac{1}{|\tilde{I}|}}\right).
\end{aligned}
$$

The second inequality can be found in Huxley [2003]. The third inequality follows from (2.20) and (2.22). The last step uses Proposition 2.10.3. By Lemma 2.10.1, $\max_{I\in\mathcal{K}_{n,l}}\left(\mathbf{Z}_n(I) - \sqrt{2\log\frac{1}{|I|}}\right) = \mathcal{O}_p(1)$. Moreover, $\mathbf{Z}_n(\tilde{I}) \sim N(0,1)$, $\left(d_\epsilon\sqrt{\pi r^2} - o(1)\right)b_n \to \infty$, and from (2.20) and (2.21) we know $\sqrt{2\log\frac{1}{|\tilde{I}|}} \to \sqrt{2\log\frac{1}{|I^*|}} = \sqrt{2\log\frac{1}{c\pi r^2}}$. Therefore, the last event probability in the chain of equations above converges to zero, finishing the proof of Theorem 2.7.1. □

Proof of Theorem 2.7.2. It suffices to show that if condition (2.13) is not met, then for sufficiently small T', there exists some positive $\beta > 0$ such that

$$
\mathbb{P}_{T=\infty}\{\phi_n \text{ rejects } H_{0,n}\} + \mathbb{P}_{T=T'}\{\phi_n \text{ accepts } H_{0,n}\} > \beta.
$$

Note that when H_0 holds, $T = \infty$ so $S_1 = S_0$. At time t_1 the pixel field will have a disk of area πr^2 with elevated mean μ_n. On the other hand, under H_1, $T = T'$ is sufficiently small such that after one time frame all random pixels on the retina will have elevated mean μ_n. Therefore, distinguishing between these two scenarios is equivalent to testing the existence of a disk of fixed area where inside the disk all random variables follow $N(\mu_n, 1)$ and outside the disk all random variables follows $N(0,1)$. Note that even if

we know the exact position of the disk, and thus focus on the $\mathcal{O}(n^2)$ Gaussian random variables of interest, to test between $\mu = 0$ and $\mu = \mu_n$, we must have

$$\mu_n = \frac{b_n}{n} \text{ for some } b_n \to \infty.$$

□

Chapter 3

Shallow Neural Networks Trained to Detect Collisions Recover Features of Visual Loom-selective neurons

3.1 Introduction

For animals living in dynamic visual environments, it is important to detect the approach of predators or other dangerous objects. Many species, from insects to humans, rely on a range of visual cues to identify approaching, or looming, objects [Regan and Beverley, 1978, Sun and Frost, 1998, Gabbiani et al., 1999, Card and Dickinson, 2008, Münch et al., 2009, Temizer et al., 2015]. Among other cues, looming objects create characteristic visual flow fields. When an object is on a ballistic collision course with an animal, its edges will appear to the observer to expand radially outward, gradually occupying a larger and larger portion of the visual field (Figure 3.3). An object heading towards the animal, but which will not collide with it, also expands to occupy an increasing portion of the visual field, but its edges do not expand radially outwards with respect to the observer. Instead, they expand with respect to the object's center so that opposite edges are perceived to be moving in the same direction (Figure 3.3). A collision detector must distinguish between these two cases, while also avoiding predicting collisions in response to a myriad of other

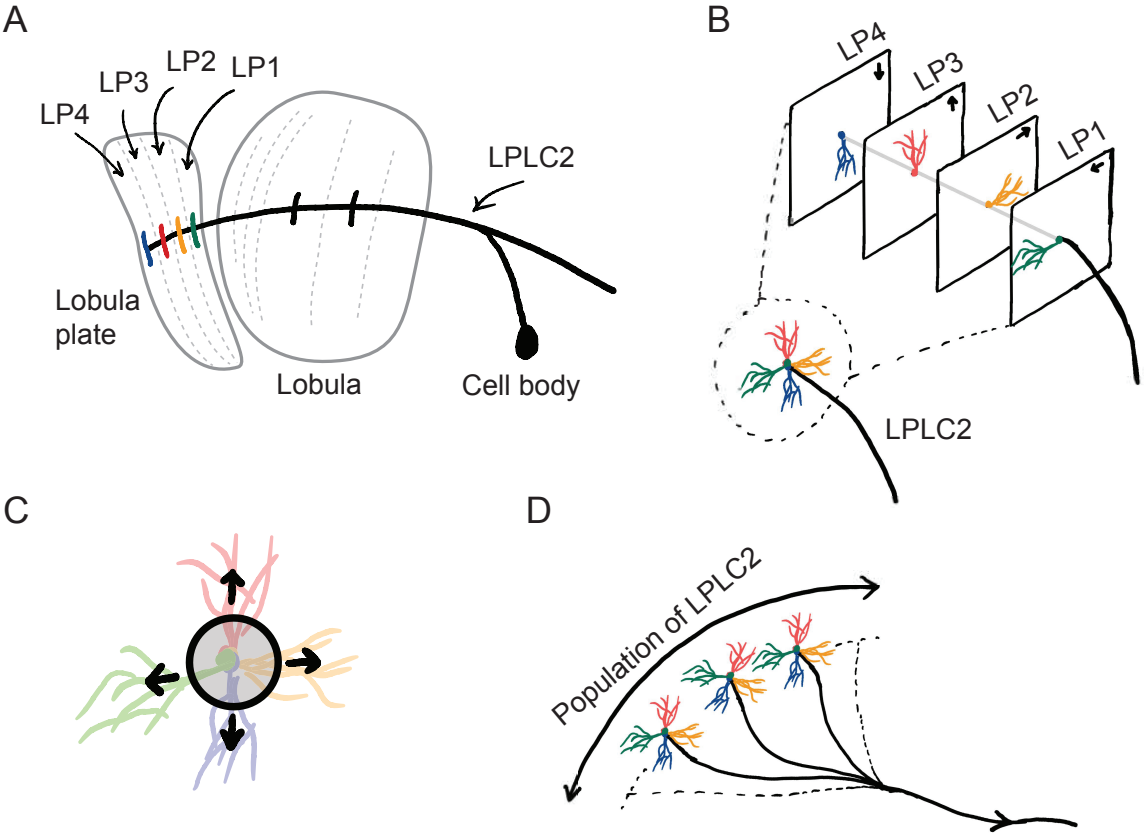

Figure 3.1: Sketches of the anatomy of LPLC2 neurons [Klapoetke et al., 2017]. (A) An LPLC2 neuron has dendrites in lobula and the four layers of the lobula plate (LP): LP1, LP2, LP3 and LP4. (B) Schematic of the four branches of the LPLC2 dendrites in the four layers of the LP. The arrows indicate the preferred direction of motion sensing neurons with axons in each LP layer [Maisak et al., 2013]. (C) The outward dendritic structure of an LPLC2 neuron is selective for the outwardly expanding edges of a looming object (black circle). (D) The axons of a population of more than 200 LPLC2 neurons converge to the GF, a descending neuron, to contribute to signaling for escaping behaviors [Ache et al., 2019b]

visual flow fields created by the animal's own motion (Figure 3.3). Thus, loom detection can be framed as a visual inference problem.

Many sighted animals solve this inference problem with high precision, thanks to robust loom-selective neural circuits evolved over hundreds of millions of years. The neuronal mechanisms for response to looming stimuli have been studied in a wide range of vertebrates, from cats and mice to zebrafish, as well as in humans [King et al., 1992, Hervais-Adelman et al., 2015, Ball and Tronick, 1971, Liu et al., 2011, Salay et al., 2018, Liu et al., 2011, Shang et al., 2015, Wu et al., 2005, Temizer et al., 2015, Dunn et al., 2016, Bhattacharyya et al., 2017]. In invertebrates, detailed anatomical, neurophysiological, behavioral and modeling studies have investigated loom detection, especially for locusts and flies [Oliva and Tomsic, 2014, Sato and Yamawaki, 2014, Santer et al., 2005, Rind and Bramwell, 1996, Card and Dickinson, 2008, De Vries and Clandinin, 2012, Muijres et al., 2014, Klapoetke et al., 2017, Von Reyn et al., 2017, Ache et al., 2019b]. An influential mathematical model of loom detection was derived by studying the responses of the giant descending neurons of locusts, which established a relationship between the timing of the neurons' peak responses and an angular size threshold for the looming object [Gabbiani et al., 1999]. Similar models have been applied to analyze neuronal responses to looming signals in flies, where genetic tools make it possible to precisely dissect neural circuits, revealing various neuron types that are sensitive to looming signals [Von Reyn et al., 2017, Ache et al., 2019b, Morimoto et al., 2020]. However, these computational studies did not directly investigate the relationship between the structure of the loom-sensitive neural circuits and the inference problem they appear to solve. Here, we asked whether we can achieve the properties associated with neural loom detection simply by optimizing shallow neural networks for collision detection.

The starting point for our computational model of loom detection is the known neuroanatomy of the visual system of the fly. In particular, the loom-sensitive neuron LPLC2 (lobula plate/lobula columnar, type 2) [Wu et al., 2016] has been studied in detail. These

neurons tile visual space, sending their axons to a descending neuron called the giant fiber (GF), which triggers the fly's jumping and take-off behaviors [Tanouye and Wyman, 1980, Card and Dickinson, 2008, Von Reyn et al., 2017, Ache et al., 2019b]. Each LPLC2 neuron has four dendritic branches that receive inputs from the four layers of the lobula plate (LP) (Figure 3.1A) [Maisak et al., 2013, Klapoetke et al., 2017]. The retinotopic LP layers host the axon terminals of motion detection neurons, and each layer uniquely receives motion information in one of the four cardinal directions [Maisak et al., 2013]. Moreover, the physical extensions of the LPLC2 dendrites align with the preferred motion directions in the corresponding LP layers (Figure 3.1B) [Klapoetke et al., 2017]. These dendrites form an outward radial structure, which matches the moving edges of a looming object that expands in the visual field (Figure 3.1C). Common stimuli such as the wide-field motion generated by movement of the insect only match part of the radial structure, and strong inhibition for inward-directed motion suppresses responses to such stimuli. Thus, the structure of the LPLC2 dendrites favors responses to objects with edges moving radially outwards, corresponding to motion toward center of the receptive field.

The focus of this chapter is to investigate how loom detection in LPLC2 can be seen as the solution to a computational inference problem. Can the structure of the LPLC2 neurons be explained in terms of optimization—carried out during the course of evolution—for the task of predicting which trajectories will result in collisions? How does coordination among the population of more than 200 LPLC2 neurons tiling a fly's visual system affect this optimization? To answer these questions, we built a simple anatomically-constrained neural network model, which receives motion signals in the four cardinal directions. We trained the model to detect visual objects on a collision course with the observer using artificial stimuli. Surprisingly, optimization finds two distinct types of solutions, with one resembling the LPLC2 neurons and the other having a very different configuration. We analyzed how each of these solutions detects looming events and where they show distinct individual and population behaviors. When the number of units tiling visual space is

increased, the solutions that resemble the actual LPLC2 neurons become favored. When tested on visual stimuli not in the training data, the optimized solutions exhibit response curves that are similar to those of actual LPLC2 neurons as measured by Klapoetke et al. [2017]. Importantly, the optimized model reproduces the canonical linear relationship between the timing of the peak responses and the size-to-speed ratio [Gabbiani et al., 1999]. Although only receiving motion signals, the model shows characteristics of an angular size encoder, which is consistent with many biological loom detectors [Gabbiani et al., 1999, Von Reyn et al., 2017, Ache et al., 2019b]. Our results show that optimizing a neural network to detect looming events gives rise to the properties and tuning of LPLC2 neurons.

3.2 Results

3.2.1 A set of artificial visual stimuli is designed for training models

Our goal is to compare computational models trained to perform loom-detection with the biological computations in LPLC2 neurons. We first created a set of stimuli to act as training data for the inference task (Methods and Materials). We considered the following four types of motion stimuli: loom-and-hit (abbreviated as hit), loom-and-miss (miss), retreat, and rotation (Figure 3.2). The hit stimuli consist of a sphere that moves ballistically towards the origin on a collision course. The miss stimuli consist of a sphere that moves ballistically towards the origin but misses it. The retreat stimuli consist of a sphere moving ballistically away from the origin. The rotation stimuli consist of objects rotating about an axis going through the origin. All stimuli were designed to be isotropic; the first three stimuli could have any orientation in space, while the rotation could be about any axis through the origin. All trajectories were simulated in the frame of reference of the fly at the origin, with distances measured with respect to the origin. For simplicity, the fly is

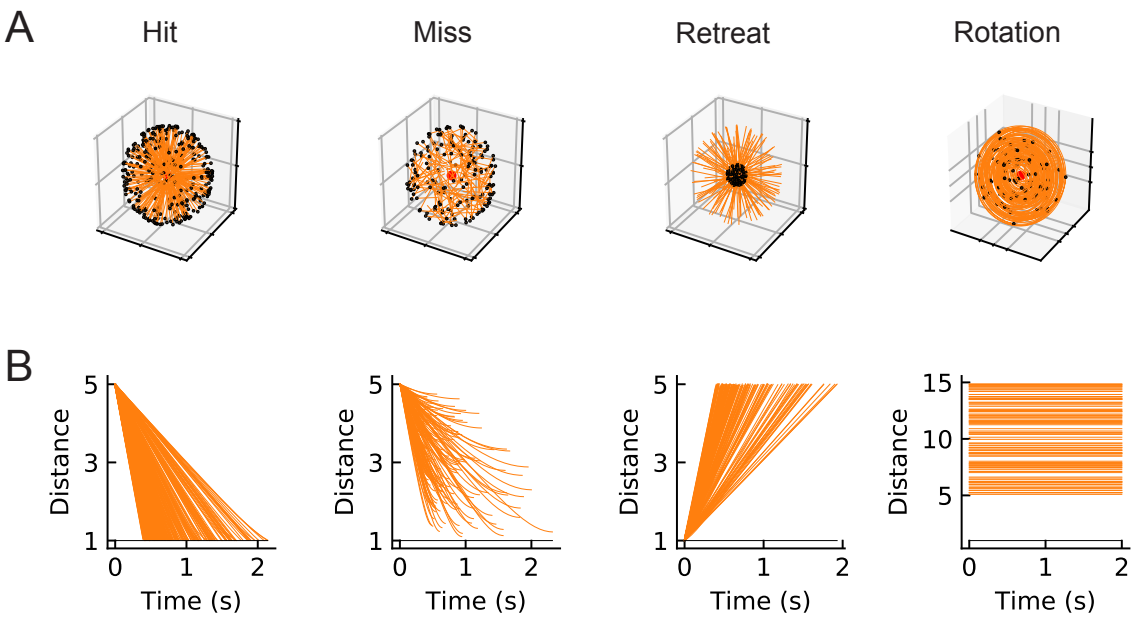

Figure 3.2: Four types of synthetic stimuli (Methods and Materials). (A) Orange lines represent trajectories of the stimuli. The black dots represent the starting points of the trajectories. For hit, miss, and retreat cases, multiple trajectories are shown. For rotation, only one trajectory is shown. (B) Distances of the objects to the fly eye as a function of time. Among misses, only the approaching portion of the trajectory was used. The horizontal black lines indicate the distance of 1.

assumed to be a point particle with no volume (Red dots in Figure 3.2 and the apexes of the cones in Figure 3.3). For hit, miss, and retreat stimuli, the spherical object has unit radius, and for the case of rotation, there were 100 objects of various radii scattered isotropically around the fly (Figure 3.3).

3.2.2 An anatomically-constrained mathematical model

We designed and trained a simple, anatomically-constrained neural network (Figure 3.4 to infer whether or not a moving object will collide with the fly. The features of this network were designed to mirror anatomical features of the fly's LPLC2 neurons (Figure 3.1). Model units receive input from a 60 degree diameter cone of visual space, represented by white cones and grey circles in Figure 3.3, mirroring the receptive field size that has been measured for LPLC2 [Klapoetke et al., 2017]. The four stimulus sets were projected into

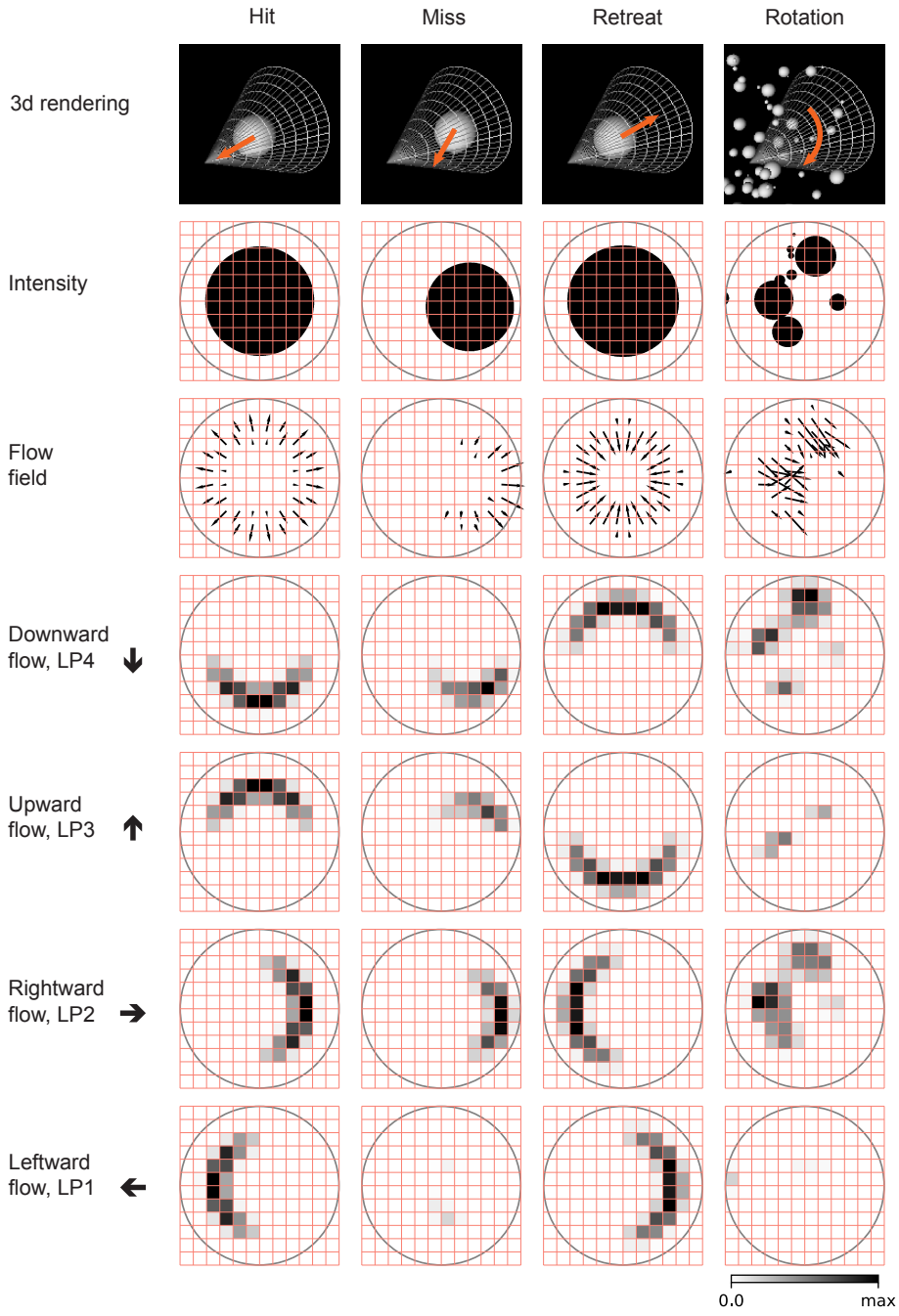

Figure 3.3: Snapshots of optical flows and flow fields calculated by a Hassenstein Reichardt correlator (HRC) model (Methods and Materials) for the 4 types of stimuli (Figure 3.2). First row: 3d rendering of the spherical objects and the LPLC2 receptive field (represented by a cone) at a specific time in the trajectory. The orange arrows indicate the motion direction of each object. Second row: 2d projections of the objects (black shading) within the LPLC2 receptive field (the grey circle). Third row: the thin black arrows indicate flow fields generated by the edges of the moving objects. Forth to seventh rows: decomposition of the flow fields in the four cardinal directions with respect to the LPLC2 neuron under consideration: downward, upward, rightward, and leftward, as indicated by the thick black arrows. These act as models of the motion signal fields in each layer of the lobula plate.

this receptive field for training and evaluating the model. The inputs to the model are local directional signals computed in the four cardinal directions at each point of the visual space: downward , upward, rightward, and leftward (Figure 3.3). These represent the motion signals from T4 and T5 neurons in the four layers of the lobula plate [Maisak et al., 2013]. They are computed as the non-negative components of a Hassenstein-Reichardt correlator model [Hassenstein and Reichardt, 1956] in both horizontal and vertical directions, which acts on the intensities of the projected stimuli (Methods and Materials). The motion signals are computed with a spacing of 5 degrees, roughly matching the spacing of the ommatidia and processing columns in the fly eye [Stavenga, 2003].

Each model unit can weight the motion signals from the four layers using linear spatial filters. There are two sets of non-negative filters, the excitatory filters and the inhibitory filters; these are shown in red and blue, respectively (Figure 3.4A). Each set of filters has four components, or branches, integrating motion signals from the four cardinal directions, respectively. These spatial filters represent excitatory inputs to LPLC2 directly from T4 and T5 in the LP, and inhibitory inputs mediated by local interneurons [Klapoetke et al., 2017, Mauss et al., 2015]. All eight filters act on the 60 degree receptive field of an unit. A 90-degree rotational symmetry is imposed on the filters, so that the filters in each layer are identical. Moreover, each filter is symmetric about the axis of motion (Methods and Materials). No further assumptions were made about the structures of the filters.

The model incorporates a fundamental difference between the excitatory and inhibitory branches: while the integrated signals from each excitatory branch are sent directly to the downstream computations, the integrated signals from each inhibitory branch are rectified before being sent downstream. This difference reflects anatomical constraints of the inputs to an actual LPLC2 neuron, where the excitatory inputs are direct connections with LPLC2 while the inhibitory inputs are mediated by inhibitory interneurons (LPi) between LP layers [Mauss et al., 2015, Klapoetke et al., 2017]. The outputs of the eight branches are summed and rectified to generate the output of a single model unit in response to a

given stimulus (Figure 3.4A.)

In the fly brain, a population of LPLC2 neurons converges onto the GF (Figure 3.1D). Accordingly, in our model there are M replicates of model units, with orientations that are spread uniformly over the 4π steradians of the unit sphere (Figure 3.4C, Methods and Materials). In this way, the receptive fields of the M units roughly tile the whole angular space, with or without overlap, depending on the value of M. The sum of the responses of the M model units is fed into a sigmoid function to generate the predicted probability of collision for a given trajectory (Methods and Materials).

3.2.3 Optimization finds two distinct solutions to the loom-inference problem

The objective of this study is to investigate how the binary classification task shapes the excitatory and inhibitory filters, and how the number of units M affects the results. We begin with the simplest model, which possess only a single unit, i.e., $M = 1$. After training with 200 random initializations of the filters, we find that the converged solutions fall into three broad categories (Figure 3.5A, B). One set of solutions is largely unstructured, with almost all the elements in the filters equal to zero (labeled in black); we will ignore these for the rest of the analysis. The two structured solutions are interesting because, surprisingly, they have spatial structures that are roughly opposite from one another (magenta and green). Based on the configurations of the excitatory filters (Methods and Materials), we call one solution type *outward filters* (magenta), and the other type *inward filters* (green) (Figure 3.5C). In this single-unit model, the inward solutions perform better than the outward solutions on the discrimination task (Figure 3.5D).

As the number of units M increases, the population of units covers a larger angular space, and when M is large enough ($M \geq 16$), the receptive fields of the units begin to overlap with each other (Figure 3.6A). In the fly visual system there are over 200 LPLC2

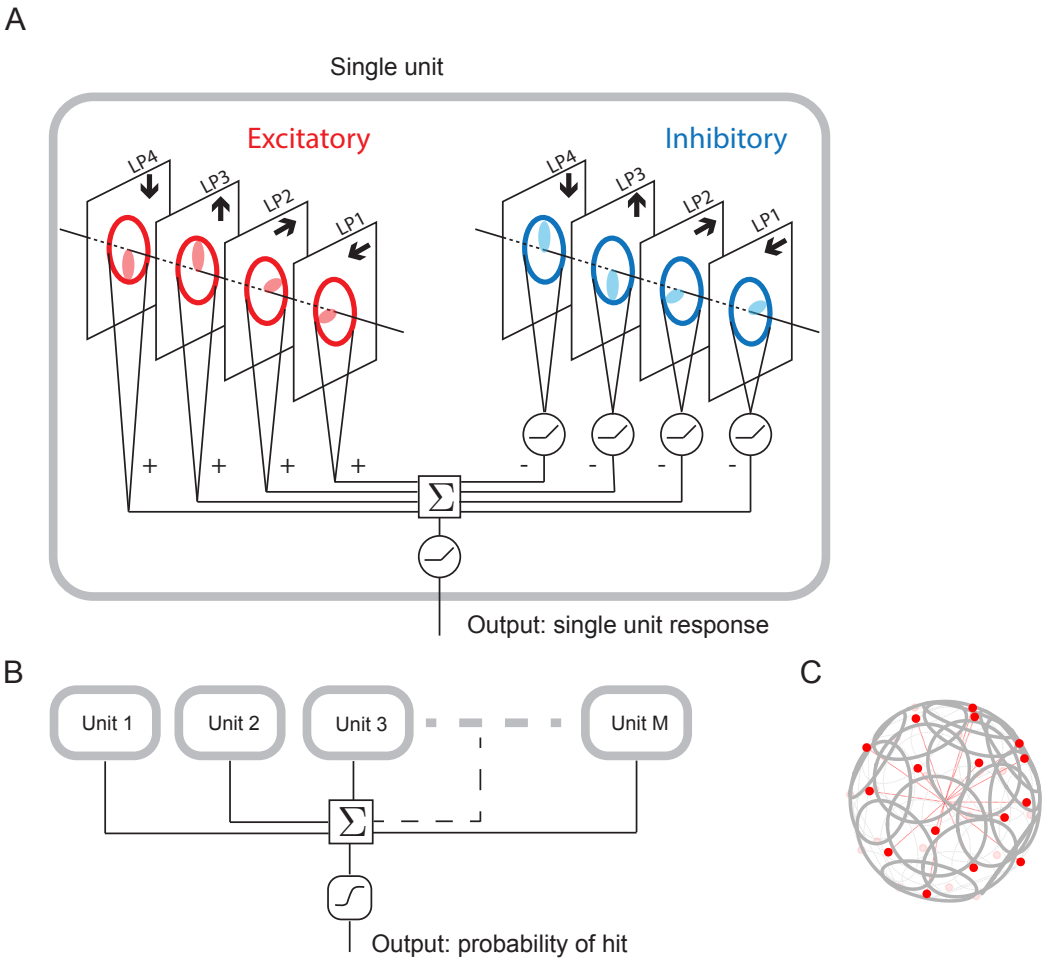

Figure 3.4: Schematic of the model (Methods and Materials). (A) Single unit. There are two sets of nonnegative filters: excitatory (red) and inhibitory (blue). Each set of filters has four branches, and each branch receives a field of motion signals (forth to seventh rows in Figure 3.3) from the corresponding layer of the model LP. The weighted signals from the excitatory branches and the inhibitory branches (rectified) are pooled together to go through a rectifier to produce an output, which is the response of a single unit. (B) The outputs from M units are summed and fed into a sigmoid function to estimate the probability of hit. (C) The M units have their orientations almost evenly distributed in angular space. Red dots represent the centers of the receptive fields and the grey lines represent the boundaries of the receptive fields on unit sphere. The red lines are drawn from the origin to the center of each receptive field.

neurons across both eyes [Ache et al., 2019b], which corresponds to a very dense distribution of the units. This is illustrated by the third row in Figure 3.6A) where $M = 256$. When M is large, approaching objects from any direction are detectable and in fact, such object signals can be detected simultaneously by many neighboring units. Interestingly, the two oppositely structured solutions persist, regardless of the value of M (Figure 3.6). In some outward solutions, structures on the right side of the inhibitory filters are similar to structures of the corresponding excitatory filters. This indicates a degree of redundancy, or non-identifiability in the model.

Units with outward-oriented filters are activated by motion radiating outwards from the center of the receptive field. Thus, these excitatory filters resemble the dendritic structures of the actual LPLC2 neurons observed in experiments, where for example, the rightward motion sensitive branch (LP2) occupies mainly the right side of the receptive field. In the outward solutions, the rightward motion-sensitive inhibitory filter mainly occupies the *left* side of the receptive field. This is also consistent with the properties of the lobula plate intrinsic (LPi) interneurons, which project inhibitory signals roughly retinotopically from one LP layer to an adjacent layer with opposite directional tuning [Mauss et al., 2015, Klapoetke et al., 2017].

The unexpected inward-oriented filters have the opposite structure. In the inward solutions, the rightward sensitive excitatory filter occupies the left side of the receptive field, and the inhibitory filter occupies the right side. Such weightings make the model selective for motion converging towards the receptive field center. At first glance, this is a puzzling structure for a loom detector, so we explored the response properties of the inward and outward solutions in more detail.

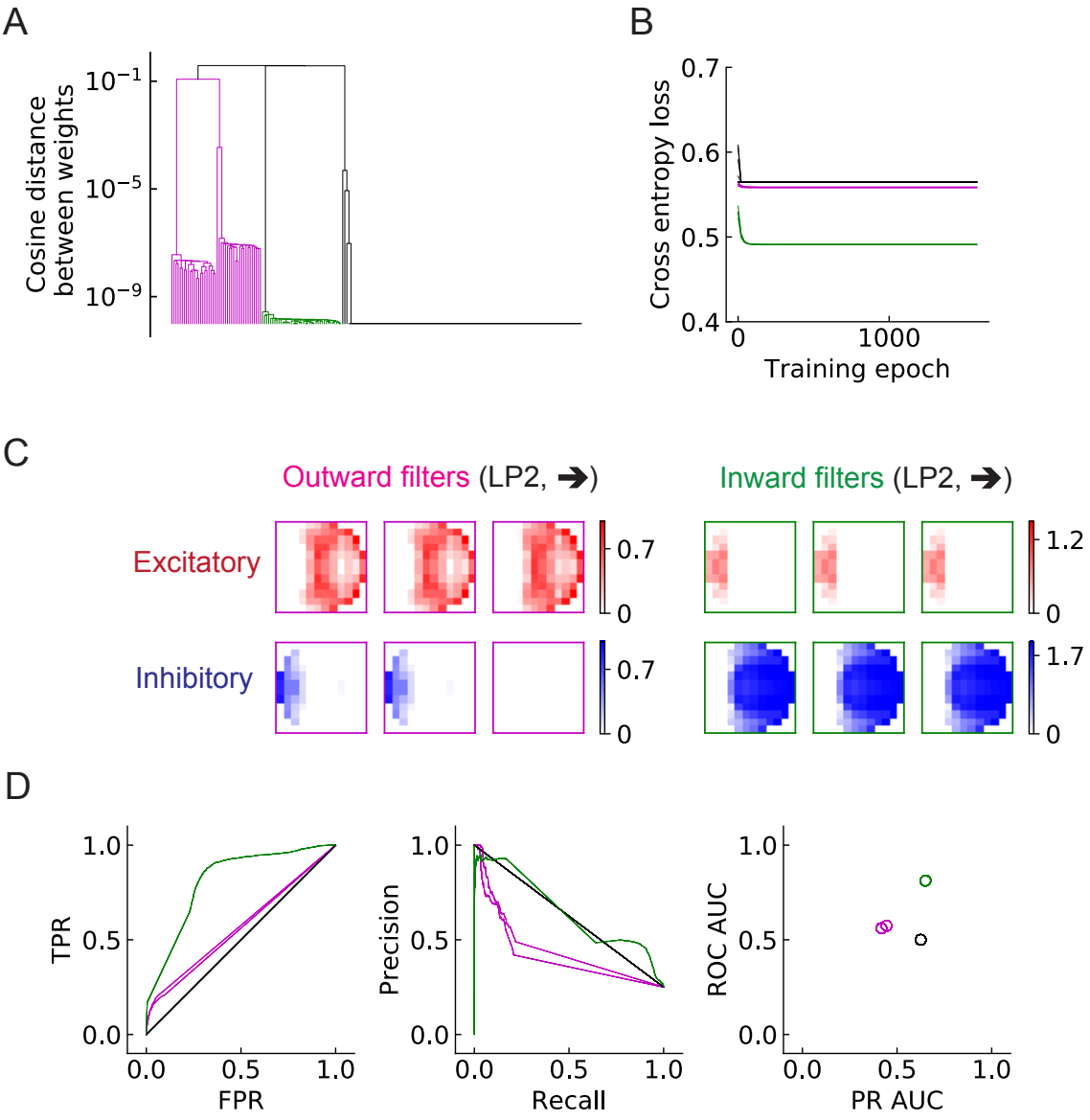

Figure 3.5: Two distinct types of solutions appear from training a single unit on the binary classification task. (A) Clustering of the trained filters/weights shown as a dendrogram (Methods and Materials). Different colors indicate different clusters, which are preserved for the rest of the chapter (see (C)) (B) The trajectories of the loss functions during training. (C) The two distinct types of solutions are represented by two types of filters that have roughly opposing structures: an outward solution (magenta) and an inward solution (green). The excitatory filter weights are shown in red, and the inhibitory filters are shown in blue. (D) Performance of the two solution types (Methods and Materials). TPR: true positive rate; FPR: false positive rate; ROC: receiver operating characteristic; PR: precision recall; AUC: area under the curve.

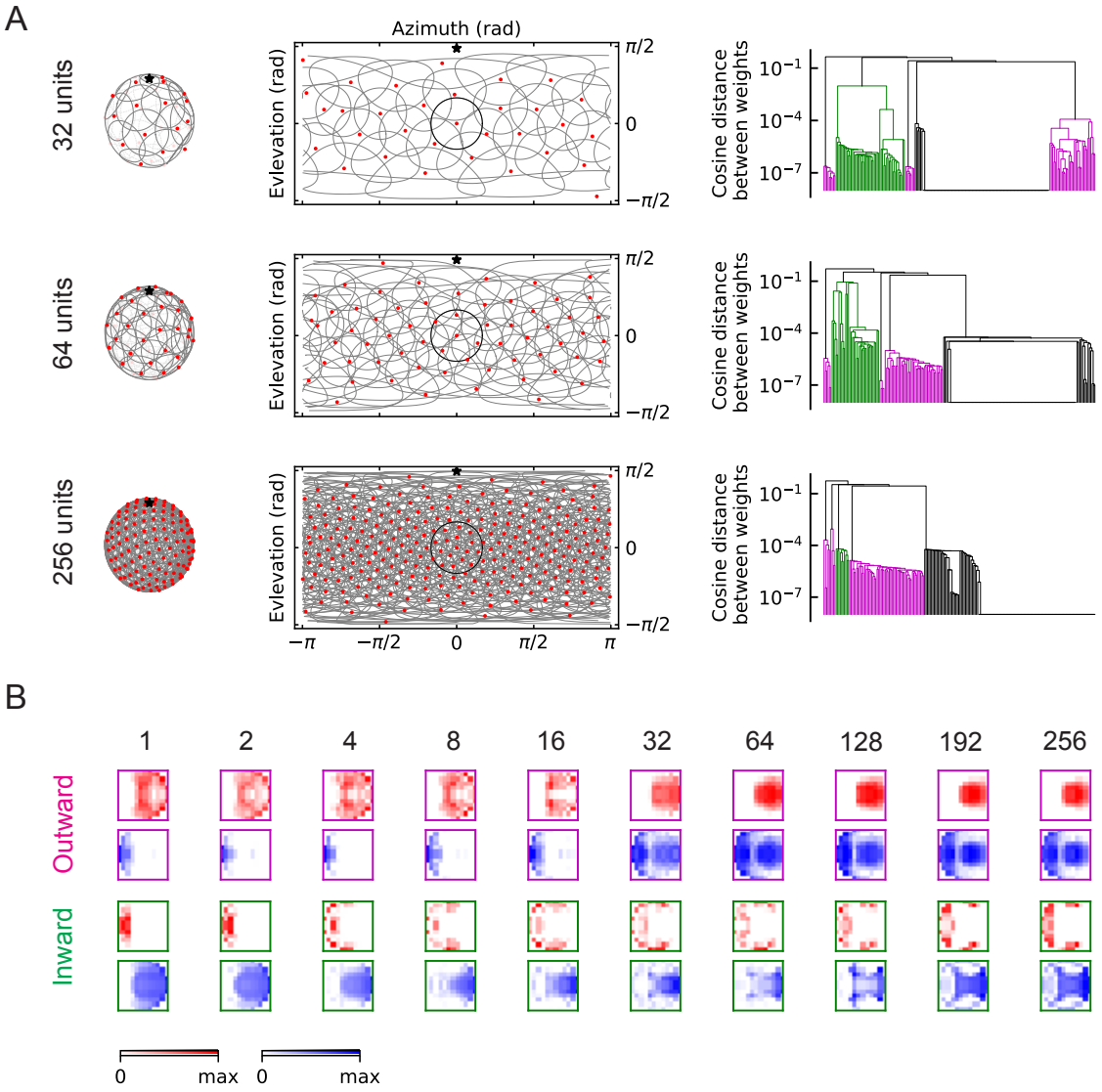

Figure 3.6: The outward and inward solutions also arise for models with multiple units. (A) Left column: angular distribution of the units, where red dots are centers of the receptive fields, the grey circles are the boundaries of the receptive fields, with one field highlighted in black, and the black star indicates the top of the fly head. Middle column: 2d map of the units with the same symbols as in the left column. Right column: clustering results shown as dendrogams with color codes as in Figure 3.5. (B) Examples of the trained excitatory and inhibitory filters for outward and inward solutions with different numbers of units.

3.2.4 Outward and inward filters are selective to signals in different ranges of angles

To understand the differences between the two types of solutions and why the inward filters can predict collisions, we investigated how they respond to hit stimuli from different incoming angles θ (Figure 3.7A). When there is no signal, the baseline activity of outward units is zero; however, the baseline activity of inward units is above zero (grey dashed lines in Figure 3.7B, C). This is because the trained intercepts are negative in the outward case, but positive in the inward case (Methods and Materials). Second, the outward filters respond strongly to stimuli near the center of the receptive field, but do not respond to stimuli having angles larger than $\sim 30^\circ$ (Figure 3.7B, C). In contrast, units with inward filters respond negatively to hit stimuli approaching toward the center and positively to stimuli approaching from the periphery of the receptive field, with angles between $\sim 30^\circ$ and $\sim 90^\circ$ (Figure 3.7B, C). This helps explain why the inward units can act as loom detectors: they are sensitive to hit stimuli originating in a larger solid angle. The hit signals are isotropic (Figure 3.2A), so the number of stimuli within angles $\sim 30^\circ$ and $\sim 90^\circ$ is much larger than the number of stimuli with angles below $\sim 30^\circ$ (Figure 3.7D). Thus, the inward solutions are sensitive to more hit cases than the outward solutions. One may visualize these responses as heat maps of the mean response of the models in terms of object distance to the fly and the incoming angle (Figure 3.7E). For the hit cases, the response patterns are consistent with the intuition about trajectory angles (Figure 3.7C). As expected, the inward solutions respond to the retreating signals with angles near $\sim 180^\circ$, since the motion of edges in that case is radially inward.

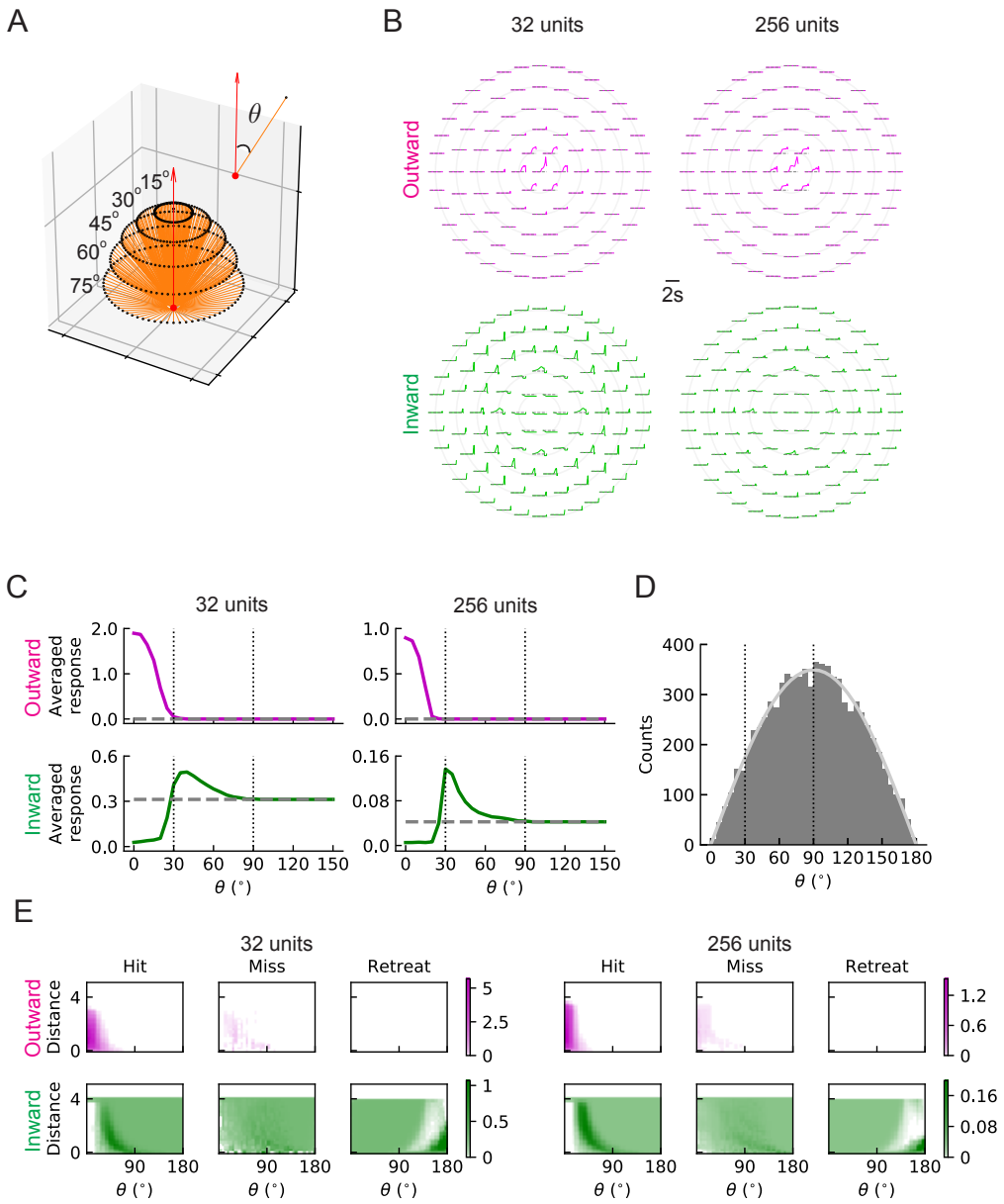

Figure 3.7: The outward and inward filters show distinct behaviors: single unit analysis. (A) Trajectories of hit stimuli with different incoming angles θ. Symbols are the same as in Figure 3.2 except that the upward red arrow represents the orientation of one unit. The numbers with degree units indicate the specific values of the incoming angles. (B) Response patterns of a single unit with either outward (magenta) or inward (green) filters obtained from optimized solutions with 32 and 256 units, respectively. The grey dashed lines show the baseline activity of the unit when there is no stimulus. The solid grey concentric circles correspond to the values of the incoming angles in (A). The scale of the responses in the top left panel is four times the scale in the other three panels. (C) Temporally averaged responses against the incoming angle θ. Symbols and colors are the same as in (B). (D) Histogram of the incoming angles for the hit stimuli in Figure 3.2A. The grey curve represents a scaled sine function equal to the expected probability for isotropic stimuli. (E) Heatmaps of the response of a single unit against the incoming angle θ and the distance to the fly head, for both outward and inward filters obtained from optimized models with 32 and 256 units, respectively.

3.2.5 Outward solutions have sparse codings and populations of units accurately predict hit probabilities

Individual units of the two solutions are very different from each other in their filter structure and response patterns to different stimuli. We decided to investigate how these differences manifest in the activities of populations of units, when units are trained to collectively predict the probability of hit. In populations of units, the outward and inward solutions exhibit very different response patterns for a given hit stimulus (Figure 3.8A, B). In particular, active outward units usually respond more strongly than inward units, but more inward units will be activated. This is consistent with the findings above, in which inward filter shapes responded to hits arriving from a wider distribution of angles.

When a population of units encodes stimuli, at each time point, the sum of the activities of the units is used to infer the probability of hit. In our trained models, the outward and inward solutions give similar probabilities of hit (Figure 3.8A). Both types of solutions can give accurate inferences for the different stimuli (Figure 3.8C). In some cases, misses can be very similar to hits if the object passes near the origin. The models reflect this in their responses to near misses which have higher hit probabilities than far misses (Figure 3.8D).

3.2.6 Large populations of units improve performance and favor outward filters

Since a larger number of units will cover an increasing spatial area of the visual field, the population of units can in principle provide more information about the incoming signals. In general, the models perform better as the number of units M increases (Figure 3.9A). When M is above 32, both the ROC-AUC and PR-AUC scores are almost 1 (Methods and Materials), which indicates that the model is very accurate on the binary classification task presented by the four types of synthetic stimuli.

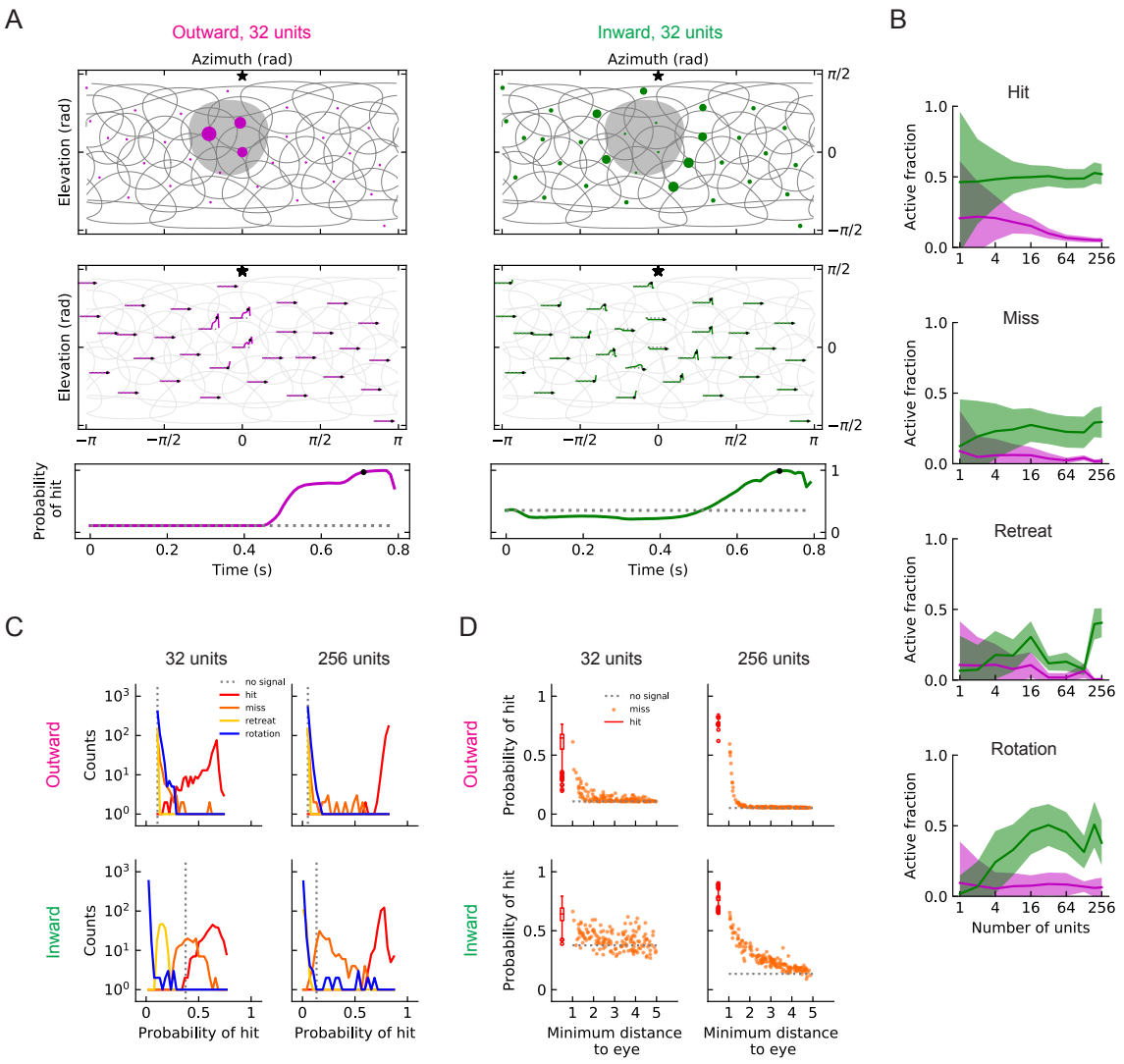

Figure 3.8: Population coding of stimuli. (A) Top row: a snapshot of the responses of outward units (magenta dots) for a hit stimulus (grey shade). Symbols and colors are as in Figure 3.6A. Middle row: the whole trajectories of the responses for the same hit stimulus as in the top row. Bottom row: the entire trajectories of the probability of hit for the same hit stimulus as in the top row (Methods and Materials). Black dots in the middle and bottom rows indicate the time step of the snapshot in the top row. (B) Fractions of the units that are activated by different types of stimuli (hit, miss, retreat, rotation) as a function of the number of units M in the model. The lines represent the mean values averaged across samples, and the shaded areas show one standard deviation (Methods and Materials). (C) Histograms of the probability of hit inferred by models with 32 or 256 units for the four types of synthetic stimuli (Methods and Materials). (D) The inferred probability of hit as a function of the minimum distance of the object to the fly eye for the miss cases. The hit distribution is represented by a box plot (the center line in the box: the median; the upper and lower boundaries of the box: 25% and 75% percentiles; the upper and lower whiskers: the minimum and maximum; the circles: outliers).

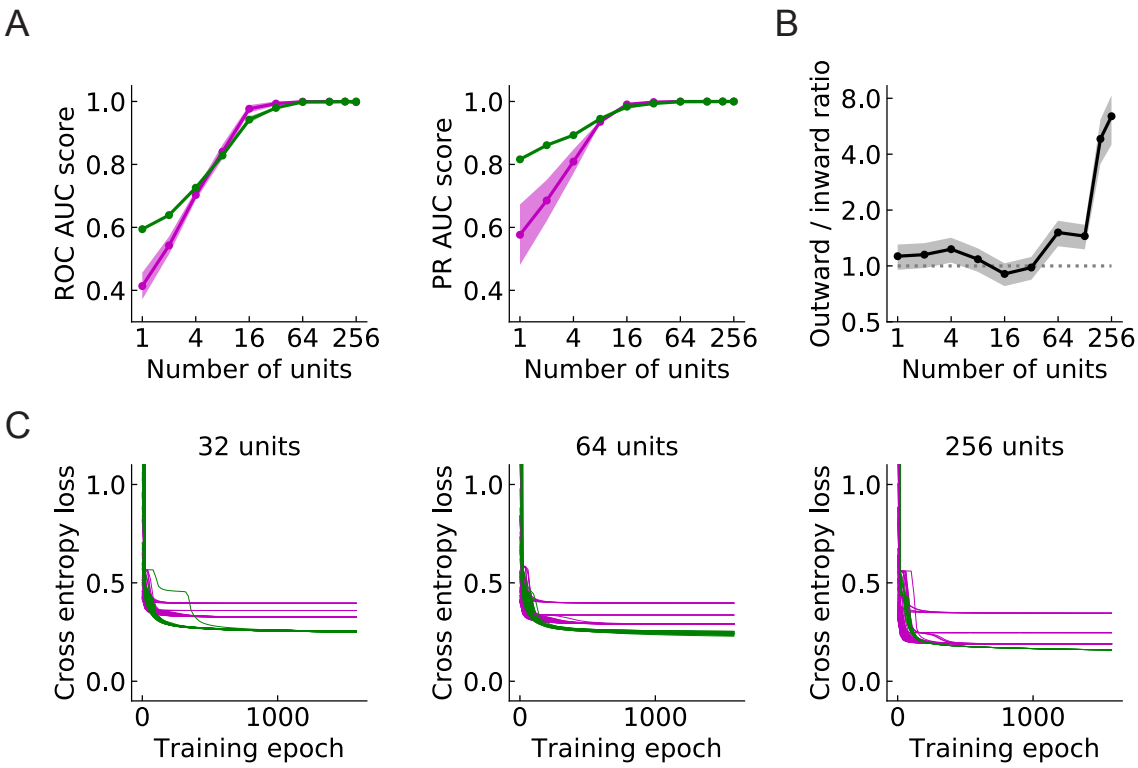

Figure 3.9: Large populations of units improve performances and favor outward solutions (Methods and Materials). (A) Both ROC and PR AUC scores increase as the number of units increases. Lines and dots: average scores; shading: one standard deviation of the scores over the trained models. Magenta: outward solutions; green: inward solutions. (B) The black line and dots show the ratio of the numbers of the two types of the solutions in the set of randomly initialized, trained models. The grey shading is one standard deviation, assuming that the distribution is binomial (Methods and Materials). The dotted horizontal line indicates the ratio of 1. (C) As the population of units increases, cross entropy losses of the outward solutions approach the losses of the inward solutions.

In addition, we calculated the ratio of the number of outward filters to inward filters that arise out of 200 random initializations in models with M units, as we swept M. Interestingly, as the number of units increases, an increasing proportion of solutions have outward filters (Figure 3.9B). For models with 256 units, the chance that an outward filter appears as a solution is almost 90% compared with the roughly 50% when $M = 1$. As M increases, the outward solutions become closer to the inward ones on both AUC scores and cross entropy losses (Figure 3.9A, C). These results suggest that as more units are added, optimization favors the outward solutions.

3.2.7 Activation patterns of computational solutions resemble biological responses

The outward solutions have a receptive field structure that is similar to LPLC2 neurons, based on their anatomy and functional studies. However, it is not clear whether these models possess the functional properties of LPLC2 neurons, which have been studied systematically [Klapoetke et al., 2017, Von Reyn et al., 2017, Ache et al., 2019b]. To see how trained units compare to LPLC2 neuron properties, we presented stimuli to the trained model (Figure 3.10A) to compare its responses to those measured in LPLC2 to similar stimuli.

The model behaves similarly to LPLC2 neurons on many different types of stimuli. Not surprisingly, the model is selective to loom signals and does not have strong responses to non-looming signals (Figure 3.10B). Moreover, the model closely follows the response of LPLC2 neurons to various expanding bar stimuli, including the inhibitory effects of inward motion (Figure 3.10C, D). In addition, in experiments, motion signals that appear at the periphery of the receptive field suppress the activity of the LPLC2 neurons (periphery inhibition) [Klapoetke et al., 2017], and this phenomenon is successfully predicted by the model (Figure 3.10E, F) due to the broad inhibitory filters that the model learns (Figure 3.10A). Interestingly, the model also correctly predicts response patterns of the LPLC2 neurons for expanding bars with different orientations (Figure 3.10G, H).

The ratio of object size to approach velocity, or R/v, is an important parameter for looming stimuli, and many studies have investigated how the response patterns of loom-sensitive neurons depend on this ratio (Top panels in Figure 3.10I, J, K, L) [Gabbiani et al., 1999, Von Reyn et al., 2017, Ache et al., 2019b, De Vries and Clandinin, 2012]. Here, we presented the trained model (Figure 3.10A) with hit stimuli with different R/v ratios, and compared its behaviors with the experimental data (Figure 3.10I-L)). Surprisingly, although our model only has the angular velocities as the inputs (Figure 3.3), it reliably

encodes the angular sizes rather than the angular velocities, indicated by the collapsed response curves (up to different scales) when plotted against the angular sizes (Figure 3.10J) [Von Reyn et al., 2017], though the model curve shapes do not exactly match the experimental ones. On the contrary, for angular velocities, the response curves shift for different R/v ratios, which means they depend on the velocities v of the object (R is fixed to be 1). Both of these response properties are consistent with properties of LPLC2. Meanwhile, a canonical linear relationship between the peak response time relative to the collision and the R/v ratio is also reproduced by the optimized model (Figure 3.10L) [Gabbiani et al., 1999, Ache et al., 2019b].

Importantly, a different outward solution from the same training procedure could reproduce many of the same effects, but it predicts the patterns in the wide expanding bars differently and out of phase from the biological data. This different solution also does a poor job predicting the response curves of the LPLC2 neurons to looming signals with different R/v ratios, although the collapsed and shifted features remain when plotted as functions of angular size and velocity. This shows that even within the family of learned outward solutions, there is variability in the learned response properties. Though solving the inference problem obtains many of the response properties, additional constraints would be required to more precisely reproduce the LPLC2 responses.

3.3 Discussion

In this study, we have shown that training a simple network to detect collisions gives rise to a computation that closely resembles neurons that are sensitive to looming signals. Specifically, we optimized a neural network model to detect whether an object is on a collision course based on the visual motion signals (Figure 3.3), and found that one class of optimized solution matched the anatomy of motion inputs to LPLC2 neurons (Figure 3.1, Figure 3.5, Figure 3.6). Importantly, this solution can reproduce a large range of exper-

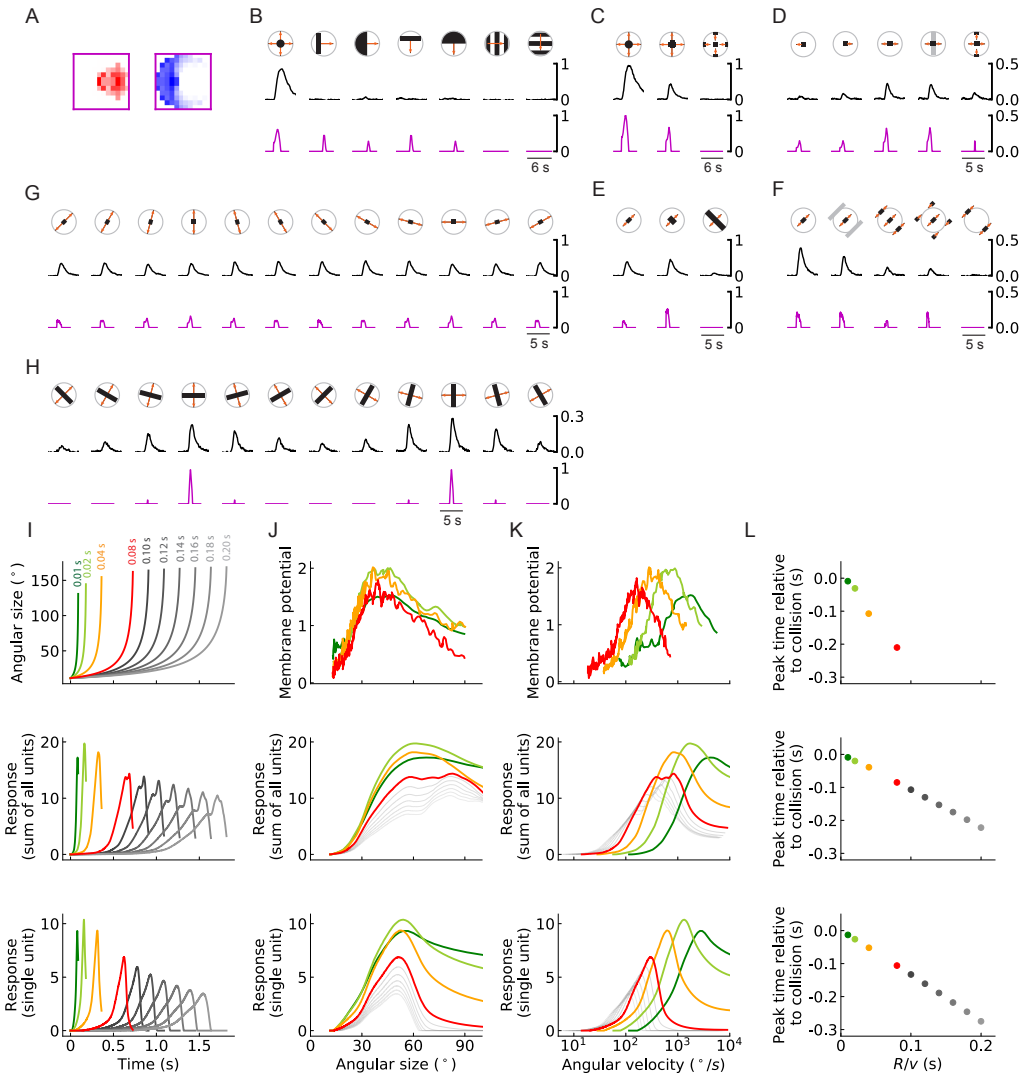

Figure 3.10: Models trained on binary classification tasks exhibit similar responses to LPLC2 neurons observed in experiments. (A) Excitatory and inhibitory filters of an outward solution with 256 units. (B-H) Comparisons of the responses of the solution in (A) and LPLC2 neurons to a variety of stimuli (Methods and Materials). Black lines: data [Klapoetke et al., 2017]; magenta lines: model. Compared with the original plots [Klapoetke et al., 2017], all the stimuli icons here except the ones in (B) have been rotated 45 degrees to match the cardinal directions of LP layers as described in this study. (I) Top: temporal trajectories of the angular sizes for different R/v ratios (color labels apply throughout (I-L)) (Methods and Materials). Middle: response as a function of time for the sum of all 256 units. Bottom: response as a function of time for one of the 256 units. (J-L) Top: experime newtal data (LPLC2/non-LC4 components of GF activity. Data from [Von Reyn et al., 2017, Ache et al., 2019b]). Middle: sum of all 256 units. Bottom: response of one of the 256 units. Responses as function of angular size (J), response as function of angular velocity (K), relationship between peak time relative to collision and R/v ratios (L). We considered the first peak when there were two peaks in the response, such as in the grey curves in the middle panel of (I).

imental observations of LPLC2 neuron responses (Figure 3.10) [Klapoetke et al., 2017, Von Reyn et al., 2017, Ache et al., 2019b].]

The radially structured dendrites of the LPLC2 neuron in the lobula plate can account for its response to motion radiating outward from the receptive field center [Klapoetke et al., 2017]. Our results show that the logic of this computation can be understood in terms of inferential loom detection by the *population* of units. In particular, for an individual detector unit, an inward structure makes a better loom detector than an outward structure, since it is sensitive to colliding objects originating from a wider array of incoming angles (Figure 3.7). As the number of units across visual space increases, the outward-sensitive receptive field structure is represented more often in the optimal solution. The solution depends on the number of detectors, and this is likely related to the increasing overlap in receptive fields as the population grows (Figure 3.6). This result is consistent with prior work showing that populations of neurons often exhibit different and improved coding strategies compared to individual neurons [Pasupathy and Connor, 2002, Georgopoulos et al., 1986, Vogels, 1990, Franke et al., 2016, Zylberberg et al., 2016, Cafaro et al., 2020]. Thus, understanding anatomical, physiological, and algorithmic properties of individual neurons can require considering the population response. The solutions we found to the loom inference problem suggest that individual LPLC2 responses should be interpreted in light of the population of LPLC2 responses.

Surprisingly, the trained outward solutions exhibits the properties of an angular size encoder (Figure 3.10I-L), even though the inputs to the model are a field of motion signals. There are two ways that this tuning arises. First, in a hit stimulus, the angular size and angular velocity are strongly correlated [Gabbiani et al., 1999], which means the angular size affects the magnitude of the motion signals. Second, the angular size is proportional to the length of the outward-moving edges of hitting objects. The angular circumference of the hit stimulus determines how many motion detectors are activated, so that integrated motion signal strength is related to the size. Both of these effects influence the response

patterns of the model units (and the LPLC2 neurons).

Our results shed light on discussions of η-like (encoding angular size) and ρ-like (encoding angular velocity) looming sensitive neurons in the literature [Gabbiani et al., 1999, Wu et al., 2005, Liu et al., 2011, Shang et al., 2015, Temizer et al., 2015, Dunn et al., 2016, Von Reyn et al., 2017, Ache et al., 2019b]. In particular, these optimized models clarify an interesting but puzzling fact: LPLC2 neurons transform their inputs of direction-selective motion signals to computations of angular size [Ache et al., 2019b]. Consistently, our model shows the linear relationship between the peak time relative to collision and the R/v ratio, which looming sensitive neurons that encode angular size should follow [Peek and Card, 2016]. In both cases, these properties appear to be the simple result of training the constrained model to reliably detect looming stimuli.

The units of the outward solution exhibit sparsity in their responses to looming stimuli, in contrast to the denser representations in the inward solution (Figure 3.8). During a looming event, most of the units are quiet and only a few adjacent units have very large activities, reminiscent of sparse codes that seem to be favored, for instance, in cortical encoding of visual scenes [Olshausen and Field, 1996, 1997]. Since the readout of our model is a summation of the activities of the units, sparsity does not directly affect the performance of the model, but is an attribute of the favored solution. For a model with a different loss function or noise, the degree of sparsity might be crucial. For instance, the sparse code of the outward model might make it easier to localize the hit stimulus [Morimoto et al., 2020], or might make the population response more robust to noise [Field, 1994].

Experiments have shown that inhibitory circuits play an important role for the selectivity of LPLC2 neurons. For example, motion signals at the periphery of the receptive field of an LPLC2 neuron inhibit its activity; such peripheral inhibition causes various interesting response patterns of the LPLC2 neurons to different types of stimuli [Klapoetke et al., 2017]. However, the structure of this inhibitory field is not fully understood,

and our model provides a tool to investigate how the inhibitory inputs to LPLC2 neurons affect circuit performance on loom detection tasks. Specifically, strong inhibition on the periphery of the receptive field arises naturally in the outward solutions after optimization [Klapoetke et al., 2017]. The broad inhibition appears in our model to suppress responses to the non-hit stimuli. As in the data, the inhibition is broader than one might expect if the neuron were simply being inhibited by inward motion.

The synthetic stimuli used to train models in this study were unnatural in two ways. The first way was in the proportion of hits and non-hits. We trained with 25% of the training data representing hits. The true fraction of hits among all stimuli encountered by a fly is undoubtedly much less, and this affects how the loss function weights different types of errors. It is also clear that a false-positive hit (in which a fly might jump to escape an object not on collision course) is much less penalized during evolution than a false-negative (in which a fly doesn't jump and an object collides, presumably to the detriment of the fly). It remains unclear how to choose these weights in the training data or in the loss function, but they affect the receptive field weights optimized by the model.

The second issue with the stimuli is that they were caricatures of stimulus types, but did not incorporate the richness of natural stimuli. This richness could include natural textures and spatial statistics [Ruderman and Bialek, 1994], which seem to impact motion detection algorithms [Fitzgerald and Clark, 2015, Leonhardt et al., 2016, Chen et al., 2019]. This richness could also include more natural trajectories for approaching objects. Another way to enrich the stimuli would be to add noise, either in inputs to the model or in the model's units themselves. These aspects of the stimuli were all neglected in this initial study, in part because it is difficult to find characterizations of natural looming events. An interesting future direction will be to investigate the effects of more complex and naturalistic stimuli on the model's filters and performance, as well as on LPLC2 neuron responses themselves.

For simplicity, this model did not impose the hexagonal geometry of the compound eye ommatidia. Instead, we assume that the visual field is separated into a Cartesian lattice

with 5° spacing, each representing a local motion detector with two spatially separated inputs (Figure 3.3). This simplification alters slightly the geometry of the motion signals compared to the real motion detector receptive fields [Shinomiya et al., 2019]. This could potentially affect the learned spatial weightings and reproduction of the LPLC2 responses to various stimuli, since the specific shapes of the filters matter (Figure 3.10). Thus, the hexagonal ommatidial structure and the full extent of inputs to T4 and T5 might be crucial if one wants to make comparisons with the dynamics and detailed responses of LPLC2 neurons. However, this geometric distinction seems unlikely to affect the main results of how to infer the presence of hit stimuli.

Our model requires a field of estimates of the local motion. Here, we used the simplest model – the Hassenstein-Reichardt correlator model Equation 3.3 (Methods and Materials) [Hassenstein and Reichardt, 1956] – but the model could be extended by replacing it with a more sophisticated model for motion estimation. Some biophysically realistic ones might take into account synaptic conductances [Gruntman et al., 2018, 2019, Badwan et al., 2019, Zavatone-Veth et al., 2020]. Alternatively, in natural environments, contrasts fluctuate in time and space. Thus, if one includes more naturalistic spatial and temporal patterns, one might consider a motion detection model that can adapt to changing contrasts in time and space [Drews et al., 2020, Matulis et al., 2020].

Our neural network model is highly constrained by the specific anatomy of LPLC2 circuits, and no unnecessary layers were added. The resulting model is a shallow neural network (Figure 3.1 and Figure 3.4. This shallowness leads to limited dimensionality of the model, which will be prone to finding non-optimal local minima during training. Indeed, in many cases, the training resulted in models with poor performance and filters with weights very close to zero. We minimized this problem by choosing the initialization scales for the filters so that optimization resulted in meaningful models with structured filters about half of the time. However, the ratio of the outward and inward solutions (Figure 3.9B) was not affected by the initialization scales.

Although the outward filter of the unit emerges naturally from our gradient descent training protocol, that does not mean that the structure is learned by LPLC2 neurons in the fly. There is some experience dependent plasticity in the fly eye [Kikuchi et al., 2012], but these visual computations are likely to be primarily genetically determined. Thus, one could think of the computation of the LPLC2 neuron as being shaped through millions of years of evolution. Interestingly, optimization algorithms similar to evolution may be able to avoid getting stuck in local optima [Stanley et al., 2019], and thus work well with the sort of shallow neural network found in the fly eye.

In this study, we focused on the motion signal inputs to LPLC2 neurons, and we neglected other inputs to LPLC2 neurons, such as inputs coming from the lobula that likely report non-motion visual features. It would be interesting to investigate how this additional non-motion information would affect the performance and optimal solutions of the inference units. For instance, another lobula columnar neurons, LC4, is loom sensitive and receives inputs in the lobula [Von Reyn et al., 2017]. The LPLC2 and LC4 neurons are the primary excitatory inputs to the GF, which mediates the escape behavior of a fly [Von Reyn et al., 2014, Ache et al., 2019b]. The inference framework set out here would allow one to incorporate of parallel non-motion intensity channels, either by adding them into the inputs to the LPLC2-like units, or by adding in a parallel population of LC4-like units. This would require a reformulation of the probabilistic model in Equation 3.5. Notably, one of the most studied loom detecting neurons, the lobula giant movement detector (LGMD) in locusts, does not appear to receive direction-selective inputs, as LPLC2 does [Rind and Bramwell, 1996, Gabbiani et al., 1999]. Thus, the inference framework set out here can be flexibly modified to investigate loom detection under a wide variety of constraints and inputs, which allow it to be applied to other neurons, beyond LPLC2.

3.4 Methods and Materials

3.4.1 Code availability

Code to perform all simulations in this chapter and to reproduce all figures is available at http://www.github.com/ClarkLabCode/LoomDetectionANN.

3.4.2 Coordinate system and stimuli

We designed a suite of visual stimuli to simulate looming objects, retreating objects, and rotational visual fields. In this section, we describe the suite of stimuli and the coordinate systems used in our simulations.

In our simulations and training, the fly is at rest on a horizontal plane, with its head pointing in a specific direction. The fly head is modeled to be a point particle with no volume. A three dimensional right-handed frame of reference Σ is set up and attached to the fly head at the origin. The z axis points in the anterior direction from the fly head, perpendicular to the line that connects the two eyes, and in the horizontal plane of the fly; the y axis points toward the right eye, also in the horizontal plane; and the x axis points upward and perpendicular to the horizontal plane. Looming or retreating objects are represented in this space by a sphere with radius $R = 1$, and the coordinates of an object's center at time t are denoted as $\mathbf{r}(t) = (x(t), y(t), z(t))$. Thus, the distance between the object center and the fly head is $D(t) = \|\mathbf{r}(t)\| = \sqrt{x^2(t) + y^2(t) + z^2(t)}$.

Within this coordinate system, we set up cones to represent individual units. The receptive field of LPLC2 neurons is measured at roughly 60° in diameter [Klapoetke et al., 2017]. Thus, we here model each unit as a cone with its vertex at the origin and with half-angle of 30°. For each unit m ($m = 1, 2, \ldots, M$), we set up a local frame of reference Σ_m: the z_m axis is the axis of the cone and its positive direction points outward from the origin. The local Σ_m can be obtained from Σ by two rotations: around x of Σ and around

the new $y^{'}$ after the rotation around x. For each unit, its cardinal directions are defined as: upward (positive direction of x_m), downward (negative direction of x_m), leftward (negative direction of y_m) and rightward (positive direction of y_m). To get the signals that are received by a specific unit m, the coordinates of the object in Σ are rotated to the local frame of reference Σ_m.

Within this coordinate system, we can set up cones representing the extent of a spherical object moving in the space. The visible outline of a spherical object spans a cone with its point at the origin. The half-angle of this cone is a function of time and can be denoted as $\theta_{\mathrm{s}}(t)$:

$$\theta_{\mathrm{s}}(t) = \arcsin \frac{R}{D(t)}. \tag{3.1}$$

One can calculate how the cone of the object overlaps with the receptive field cones of each unit.

There are multiple layers in the fly visual system [Takemura et al., 2017], but here we focus on two coarse grained stages of processing: (1) the estimation of local motion direction from optical intensities by motion detection neurons T4 and T5 and (2) the integration of the flow fields by LPLC2 neurons. In our simulations, the interior of the mth unit cone is represented by a N-by-N matrix, so that each element in this matrix (except the ones at the four corners) indicates a specific direction in the angular space within the unit cone. If an element also falls within the object cone, then its value is set to 1; otherwise it is 0. Thus, at each time t, this matrix is an optical contrast signal and can be represented by $C(x_m, y_m, t)$, where (x_m, y_m) are the coordinates in Σ_m. In general, N should be large enough to provide good angular resolutions. Then, K^2 ($K < N$) motion detectors are evenly distributed within the unit cone, with each occupying an L-by-L grid in the N-by-N matrix, where $L = N/K$. This L-by-L grid represents a 5°-by-5° square in the angular space, consistent with the approximate spacing of the inputs of motion detectors T4 and T5. This arrangement effectively upsamples the spatial resolution of the intensity data be-

fore it is discretized into motion signals with a resolution of 5°. Since the receptive field of an LPLC2 neuron is roughly 60°, the value of K is chosen to be 12. To get sufficient angular resolution for the local motion detectors, L is set to be 4, so that N is set to 48.

Each motion detector is assumed to be a Hassenstein Reichardt Correlator (HRC) and calculates local flow fields from $C(x_m, y_m, t)$ [Hassenstein and Reichardt, 1956, Potters and Bialek, 1994]. The HRC used here has two inputs, separated by 5° in angular space. Each input applies first a spatial filter on the contrast $C(x_m, y_m, t)$ and then temporal filters:

$$I_j(t; x_m, y_m) = \sum_{t'=0}^{t} \sum_{x'_m=-N}^{N} \sum_{y'_m=-N}^{N} f_j(t') G(x'_m, y'_m) C(x_m - x'_m, y_m - y'_m, t - t'), \quad (3.2)$$

where f_j $(j \in 1, 2)$ is a temporal filter and G is a discrete 2d Gaussian kernel with mean 0° and standard deviation of 2.5° to approximate the acceptance angle of the fly photoreceptors [Stavenga, 2003]. The temporal filter f_1 was chosen to be an exponential function $f_1(t') = (1/\tau)\exp(-t'/\tau)$ with τ set to 0.03 seconds [Salazar-Gatzimas et al., 2016], and f_2 a delta function $f_2 = \delta(t')$. This leads to

$$F(t; x_{m1}, y_{m1}, x_{m2}, y_{m2}) = I_1(t; x_{m1}, y_{m1}) I_2(t; x_{m2}, y_{m2}) - I_1(t; x_{m2}, y_{m2}) I_2(t; x_{m1}, y_{m1}). \quad (3.3)$$

as the local flow field at time t between two inputs located at (x_{m1}, y_{m1}) and (x_{m2}, y_{m2}).

Four types of T4 and T5 neurons have been found that project to layers 1, 2, 3, and 4 of the lobula plate. Each type is sensitive to one of the cardinal directions: down, up, left, right [Maisak et al., 2013]. Thus, in our model, there are four non-negative, local flow fields that serve as the only inputs to the model: $U_-(t)$ (downward, corresponding LP layer 4), $U_+(t)$ (upward, LP layer 3), $V_-(t)$ (leftward, LP layer 1) and $V_+(t)$ (rightward, LP layer 2), each of which is a K-by-K matrix. To calculate these matrices, two sets

of motion detectors are needed, one for the vertical directions and one for the horizontal directions. The HRC model in Equation 3.3 is direction sensitive and is opponent, meaning that for motion in the preferred (null) direction, the output of the HRC model is positive (negative). Thus, assuming that upward (rightward) is the preferred vertical (horizontal) direction, we obtain the non-negative elements of the four flow fields as

$$
\begin{aligned}
[U_-(t)]_{k_1k_2} &= \max(0, -F(t; x_{m1}, y_m, x_{m2}, y_m)) \\
[U_+(t)]_{k_1k_2} &= \max(0, F(t; x_{m1}, y_m, x_{m2}, y_m)) \\
[V_-(t)]_{k_1k_2} &= \max(0, -F(t; x_m, y_{m1}, x_m, y_{m2})) \\
[V_+(t)]_{k_1k_2} &= \max(0, F(t; x_m, y_{m1}, x_m, y_{m2})),
\end{aligned}
$$

where $k_1, k_2 \in \{1, 2, \ldots, K\}$ and $|\cdot|$ represents the absolute value. In the above expressions, it implies, for $[U_-(t)]_{k_1k_2}$ and $[U_+(t)]_{k_1k_2}$, the vertical motion detector at (k_1, k_2) has its two inputs located at (x_{m1}, y_m) and (x_{m2}, y_m), respectively. Similarly, for for $[V_-(t)]_{k_1k_2}$ and $[V_+(t)]_{k_1k_2}$, the horizontal motion detector at (k_1, k_2) has its two inputs located at (x_m, y_{m1}) and (x_m, y_{m2}). Using the opponent HRC output as the motion signals for each layer is reasonable because the motion detectors T4 and T5 are highly direction-selective over a large range of inputs [Maisak et al., 2013, Creamer et al., 2018] and synaptic, 3-input models for T4 are approximately equivalent to opponent HRC models [Zavatone-Veth et al., 2020].

We simulated the trajectories $\mathbf{r}(t)$ of the object in the frame of reference Σ at a time resolution of 0.01 seconds. For hit, miss, and retreat cases, the trajectories of the object are always straight lines (i.e., ballistic motion), and the velocities of the object were randomly sampled from a range $[2R, 10R](s^{-1})$ with the trajectories confined to be within a sphere of $5R$ centered at the fly head. The radius of the object, R, is always set to be 1 except in the rotational stimuli. To generate rotational stimuli, we placed 100 objects with various

radii randomly selected from $[0, 1]$ at random distances ([5, 15]) and positions around the fly, and rotated them all around a randomly chosen axis. The rotational speed was chosen from a Gaussian distribution with mean $0°/s$ and standard deviation $200°/s$, a reasonable rotational velocity for walking flies [DeAngelis et al., 2019].

We reproduced a range of stimuli used in a previous study [Klapoetke et al., 2017] and tested our trained model on them (Figure 3.10B-H). To match the cardinal directions of LP layers (Figure 3.1), we have rotated the stimuli (except in Figure 3.10B) 45 degrees compared with the ones displayed in the figures in [Klapoetke et al., 2017]. The disc (Figure 3.10B, C) expands from $20°$ to $60°$ with an edge speed of $10°/s$. All the bar and edge motions have an edge speed of $20°/s$. The width of the bars are $60°$ (right panel of Figure 3.10E, and Figure 3.10H), $20°$ (middle panel of Figure 3.10E), and $10°$ (all the rest). All the responses of the models (except in Figure 3.10B) have been normalized by the peak of the response to the expanding disc (Figure 3.10B).

We created a range of hit stimuli with various R/v ratios: $0.01\ s$, $0.02\ s$, $0.04\ s$, $0.08\ s$, $0.10\ s$, $0.12\ s$, $0.14\ s$, $0.16\ s$, $0.18\ s$, $0.20\ s$. The radius R of the spherical object is fixed to be 1, and the velocity is changed accordingly to achieve different R/v ratios.

3.4.3 Models

Experiments have shown that an LPLC2 neuron has four dendritic structures in the four LP layers, and that they receive direct excitatory inputs from T4/T5 motion detection neurons [Maisak et al., 2013, Klapoetke et al., 2017]. It has been proposed that each dendritic structure also receives inhibitory inputs mediated by lobulate plate intrinsic interneurons, such as LPi4-3 [Klapoetke et al., 2017]. Accordingly, our models have two types of non-negative filters, one excitatory and one inhibitory (Figure 3.4, represented by W^{e} and W^{i}, respectively. Each filter is a 12-by-12 matrix. We rotate W^{e} and W^{i} counterclockwise by multiples of $90°$ to obtain the filters that are used to integrate the four motion signals:

$U_-(t)$, $U_+(t)$, $V_-(t)$, $V_+(t)$. Specifically, we define the corresponding four excitatory filters as: $W^{\mathrm{e}}_{U_-} = \text{rotate}(W^{\mathrm{e}}, 270°)$, $W^{\mathrm{e}}_{U_+} = \text{rotate}(W^{\mathrm{e}}, 90°)$, $W^{\mathrm{e}}_{V_-} = \text{rotate}(W^{\mathrm{e}}, 180°)$, $W^{\mathrm{e}}_{V_+} = \text{rotate}(W^{\mathrm{e}}, 0°)$, and the inhibitory filters as: $W^{\mathrm{i}}_{U_-} = \text{rotate}(W^{\mathrm{i}}, 270°)$, $W^{\mathrm{i}}_{U_+} = \text{rotate}(W^{\mathrm{i}}, 90°)$, $W^{\mathrm{i}}_{V_-} = \text{rotate}(W^{\mathrm{i}}, 180°)$, $W^{\mathrm{i}}_{V_+} = \text{rotate}(W^{\mathrm{i}}, 0°)$. In addition, we impose mirror symmetry to the filters, and with the above definitions of the rotated filters, the upper half of W^{e} is a mirror image of the lower half of W^{e}. The same mirror symmetry applies to W^{i}. Thus, there are in total 144 parameters in the two sets of filters. In fact, since only the elements within a 60 degree cone contribute to the filter for the units, the corners are excluded, resulting in only 112 trainable parameters in the excitatory and inhibitory filters.

In computer simulations, the weights and flow fields are flattened to be one-dimensional column vectors. The responses of the inhibitory units are:

$$
\begin{aligned}
r^{\mathrm{i}}_{U_-}(t) &= \phi\left((W^{\mathrm{i}}_{U_-})^T U_-(t) + b^{\mathrm{i}}\right) \\
r^{\mathrm{i}}_{U_+}(t) &= \phi\left((W^{\mathrm{i}}_{U_+})^T U_+(t) + b^{\mathrm{i}}\right) \\
r^{\mathrm{i}}_{V_+}(t) &= \phi\left((W^{\mathrm{i}}_{V_+})^T V_+(t) + b^{\mathrm{i}}\right) \\
r^{\mathrm{i}}_{V_-}(t) &= \phi\left((W^{\mathrm{i}}_{V_-})^T V_-(t) + b^{\mathrm{i}}\right),
\end{aligned}
$$

where $\phi(\cdot) = \max(\cdot, 0)$ is the rectified linear activation function, and $b^{\mathrm{i}} \in \mathbb{R}$ is the intercept. The response of a single unit m is

$$
\begin{aligned}
r_m(t) = \phi\Bigg(&(W^{\mathrm{e}}_{U_-})^T U_-(t) + (W^{\mathrm{e}}_{U_+})^T U_+(t) + (W^{\mathrm{e}}_{V_+})^T V_+(t) + (W^{\mathrm{e}}_{V_-})^T V_-(t) - \\
&\left(r^{\mathrm{i}}_{U_-}(t) + r^{\mathrm{i}}_{U_+}(t) + r^{\mathrm{i}}_{V_+}(t) + r^{\mathrm{i}}_{V_-}(t)\right) + b^{\mathrm{e}}\Bigg), \quad (3.4)
\end{aligned}
$$

where $b^{\mathrm{e}} \in \mathbb{R}$ is the intercept (Figure 3.4). The inferred probability of hit for a specific

trajectory is

$$\hat{P}_{\text{hit}} = \frac{1}{T}\sum_{t=1}^{T}\sigma\left(\sum_{m} r_m(t) + b\right), \tag{3.5}$$

where T is the total number of time steps in the trajectory and $\sigma(\cdot)$ is the sigmoid function. Since we are adding three intercepts b^{i}, b^{e}, and b, there are 115 parameters to train in this model.

3.4.4 Training and testing

We created a synthetic data set containing four types of motion: *loom-and-hit*, *loom-and-miss*, *retreat*, and *rotation*. The proportions of these types were 0.25, 0.125, 0.125, and 0.5, respectively. In total, there were 5200 trajectories, with 4,000 for training and 1,200 for testing. Trajectories with motion type *loom-and-hit* are labeled as hit or $y_n = 1$ (probability of hit is 1), while trajectories of other motion types are labeled as non-hit or $y_n = 0$ (probability of hit is 0), where n is the index of each specific sample. Models with smaller M have fewer trajectories in the receptive field of any unit. For stability of training, we therefore increased the number of trajectories by factors of eight, four, and two for $M = 1, 2, 4$, respectively.

The loss function to be minimized in our training was the cross entropy between the label y_n and the inferred probability of hit $\hat{P}_{\text{hit}}$, and averaged across all samples, together with a regularization term:

$$\text{loss} = -\frac{1}{N}\sum_{n=1}^{N}\Big\{y_n \log \hat{P}_{\text{hit}}(n) + (1-y_n)\log(1-\hat{P}_{\text{hit}}(n))\Big\} + \beta\sum_{W}\|W\|^2, \tag{3.6}$$

where $\hat{P}_{\text{hit}}(n)$ is the inferred probability of hit for sample n, β is the strength of the ℓ_2 regularization, and W represents all the effective parameters in the two excitatory and inhibitory filters.

The strength of the regularization β was set to be 10^{-4}, which was obtained by grad-

ually increasing β until the performance of the model on test data started to drop. The regularization sped up convergence of solutions, but the regularization strength did not strongly influence the main results in the chapter.

To speed up training, rather than taking a temporal average as shown in Equation 3.5, a snapshot was sampled randomly from each trajectory, and the probability of hit of this snapshot was used to represent the whole trajectory, i.e., $\hat{P}_{\text{hit}} = \sigma\left(\sum_m r_m(t) + b\right)$, where t is a random sample from $\{1, 2, \ldots, T\}$. Mini-batch gradient descent was used in training, and the learning rate was 0.001.

After training, the models were tested on the entire trajectories with the probability of hit defined in Equation 3.5. Models trained only on snapshots performed well on the test data. During testing, the performance of the model was evaluated by the area under the curve (AUC) of the receiver operating characteristic (ROC) and precision-recall (PR) curves [Hanley and McNeil, 1982, Davis and Goadrich, 2006]. TensorFlow [Abadi et al., 2016] was used to train all models.

3.4.5 Clustering the solutions

We used the following procedure to cluster the solutions. Each solution had an excitatory and an inhibitory filter. We flattened these two filters, and concatenated them into a single vector. (The elements at the corners were deleted since they are outside of the receptive field.) Thus, each solution was represented by a vector, from which we calculated the cosine distance for each pair of solutions. The obtained distance matrix was then fed into a hierarchical clustering algorithm [Virtanen et al., 2020]. After obtaining the hierarchical clustering, the outward and inward filters were identified by their shape. We counted the non-zero filter elements corresponding to flow fields with components radiating outward and subtracted the number of non-zero filter elements corresponding to flow fields with components directed inward. If the resulted value was positive, the filters were labeled as

outward; otherwise, the filters were labeled as inward. If the elements in the concatenated vector were all close to zero, then the corresponding filters were labeled as unstructured.

3.4.6 Statistics

To calculate the fraction of active units for the model with $M = 256$ (Figure 3.8B), we looked at the response curves of each unit to all trajectories of a specific type of stimuli, and if the unit response is above the baseline (dotted lines in Figure 3.7B), then the unit is counted as active. So, for each trajectory/stimulus, we obtained the number of active units. After this, we calculated the mean and standard deviation across all the trajectories within each type of stimuli (hit, miss, retreat, rotation).

For a model with M units, where $M \in \{1, 2, 4, 8, 16, 32, 64, 128, 192, 256\}$, 200 random initializations were used to train it. Within these 200 training runs, the number of outward solutions N_{out} were (starting from smaller values of M) 44, 46, 48, 50, 48, 50, 53, 55, 58, 64, and the number of inward solutions N_{in} were 39, 40, 39, 46, 53, 51, 35, 38, 12, 10. The average score curves and dots in Figure 3.9A were obtained by taking the average among each type of solution, with the shading indicating two standard deviations. The curve and dots in Figure 3.9B are the ratio of the number of outward solutions to the number of inward solutions. To obtain error bars (grey shade), we considered the training results as a binomial distribution, with the probability of obtaining an outward solution being $N_{\text{out}}/(N_{\text{out}} + N_{\text{in}})$, and with the probability of obtaining an inward solution being $N_{\text{in}}/(N_{\text{out}} + N_{\text{in}})$. Thus, the standard deviation of this binomial distribution is $\sigma_{\text{b}} = \sqrt{N_{\text{out}}N_{\text{in}}/(N_{\text{out}} + N_{\text{in}})}$. From this, we calculate the error bar as the propagated error [Morgan et al., 1990]:

$$\text{propagated error} = \frac{N_{\text{out}}}{N_{\text{in}}}\sqrt{\left(\frac{\sigma_{\text{b}}}{N_{\text{out}}}\right)^2 + \left(\frac{\sigma_{\text{b}}}{N_{\text{in}}}\right)^2}. \tag{3.7}$$

Chapter 4

Sketched Least-Squares Value Iteration for Linear Markov Decision Processes

4.1 Introduction

We consider a basic reinforcement learning model – the Markov Decision Process (MDP), consisting of a tuple $(\mathcal{S}, \mathcal{A}, \mathbb{P}, r)$ where $\mathcal{S}, \mathcal{A}, \mathbb{P}, r$ denote the state space, action space, transition probability, and reward function, respectively. In MDP, an agent at a state $s \in \mathcal{S}$ plays an action $a \in \mathcal{A}$. After receiving a reward $r(s, a)$, the agent transitions to a new state $s' \in \mathcal{S}$ according to the unknown transition probability $\mathbb{P}(s'|s, a)$. The goal of the agent is to maximize the long-term rewards by interacting with the environment.

The performance of an algorithm for MDP is usually measured by *regret*, which refers to the difference between the cumulative rewards obtained using the best policy and the cumulative reward obtained by the algorithm. In the tabular case where both $\mathcal{S}$ and $\mathcal{A}$ are finite sets, algorithms that achieve $\mathcal{O}(\sqrt{|\mathcal{S}||\mathcal{A}|T})$ regret have been proposed [Jaksch et al., 2010, Azar et al., 2017, Agrawal and Jia, 2017, Dann et al., 2019, Zanette and Brunskill, 2019, Efroni et al., 2019]. However, the polynomial dependency on the size of the state and action space sometimes can be prohibitive, e.g., in the game of Go where the number of states is on the order of 3^{361}. But the worst case $\mathcal{O}(\sqrt{|\mathcal{S}||\mathcal{A}|T})$ regret cannot be improved

in general [Jaksch et al., 2010, Osband et al., 2019].

To break this curse of dimensionality, function approximation is commonly employed to approximate the (action-)value function of the policy. In fact, there have been numerous practical successes of reinforcement learning with function approximation based on deep neural networks [Mnih et al., 2015, Silver et al., 2017, Kober et al., 2012]. However, theoretical guarantees on reinforcement learning with function approximation have been scarce. Recently, a number of theoretical works considered reinforcement learning with linear function approximation [Jin et al., 2019, Zanette et al., 2019, Yang and Wang, 2019a,b, Cai et al., 2019, Du et al., 2019] or generalized linear function approximation [Wang et al., 2019]. In particular, Jin et al. [2019], Zanette et al. [2019] focused on MDPs with certain linear structure that include all tabular MDPs as special case to motivate linear approximation, as optimal value functions are linear under the structural assumption. Two different algorithms based on the idea of Least-Squares Value Iteration (LSVI) were proposed, one utilizing the *optimism-in-the-face-of-uncertainty* principle and the other bearing resemblance to Thompson Sampling. For both algorithms, it is shown that the regret only depends polynomially on the feature dimension instead of dimensions of state or action space.

While the curse of dimensionality on regret was overcome, computational efficiency emerges as a potential concern for both LSVI algorithms. The time and space complexity of both algorithms are $\mathcal{O}(d^2)$, which could be prohibitively large when feature dimension d is large. In fact, it is common to find reinforcement learning applications with high-dimensional features, e.g., when the features are pixels of raw images from Atari game [Mnih et al., 2015]. Therefore, improving the computational efficiency of reinforcement learning algorithms is of significant practical interest.

Matrix sketching is a powerful dimensionality reduction technique that has been studied extensively for kernel regression and other problems [Woodruff, 2014, Cohen et al., 2015, Alaoui and Mahoney, 2015, Yang et al., 2017]. Recently, it has found applications in

online learning [Luo et al., 2016, Calandriello et al., 2017, Luo et al., 2019], including (kernelized) linear bandits [Kuzborskij et al., 2019, Calandriello et al., 2019]. Ghavamzadeh et al. [2010] and Pan et al. [2017] applied matrix sketching to learn the value function of policies that are approximated by linear functions. However, they focused on the problem of policy evaluation rather than regret minimization.

In this chapter, we show that an online deterministic matrix sketching procedure called *Frequent Directions* [Liberty, 2013, Ghashami et al., 2016] can be used to make LSVI algorithms more efficient. Perhaps surprisingly, in addition to the improved computational efficiency, we show that the asymptotic regret bound of sketched LSVI algorithms is smaller than the bound of the non-sketched counterpart in some regime, depending on a tradeoff governed by the sketch size m and the tail spectra not covered by the sketch. This is in stark contrast to the results from many previous works on sketching for regret minimization in online learning problems, where the regret bounds are only inflated by some factors without any gains [Luo et al., 2016, Calandriello et al., 2017, Luo et al., 2019, Kuzborskij et al., 2019, Calandriello et al., 2019]. In fact, without any low-rank assumption on the problem, it might not be intuitive to see why matrix sketching could lead to regret improvement. On a high level, it is possible here because the proof relies on uniform concentration of self-normalized process with random value functions, and hence depends on the covering number of the class of random value functions parameterized by some matrices. The reduced covering number of the class of functions parameterized by low-rank sketched matrices might compensate for the spectral error introduced by sketching.

Mathematically, the regret bounds of Least-Squares Value Iteration with Upper Confidence Bounds (LSVI-UCB) algorithm proposed by Jin et al. [2019] and Randomized Least-Squares Value Iteration (RLSVI) algorithm proposed by Zanette et al. [2019] are $\widetilde{\mathcal{O}}(\sqrt{d^3H^3T})$ and $\widetilde{\mathcal{O}}(\sqrt{d^4H^4T})$ [1], respectively. Applying Frequent Directions with a sketch of size m, sketched LSVI-UCB and RLSVI algorithms incur only $\mathcal{O}(md)$ time

[1] Here $\widetilde{\mathcal{O}}$ hides only constants and poly-logarithmic factors.

and space complexity, while having regret upper bounded as $\widetilde{\mathcal{O}}((1+\varepsilon_m)^{3/2}\sqrt{md^2H^3T})$ and $\widetilde{\mathcal{O}}((1+\varepsilon_m)^{3/2}\sqrt{md^3H^4T})$, respectively. Even though the regret bounds of sketched algorithms are inflated by a factor of $(1+\varepsilon_m)^{3/2}$, where ε_m is bounded by the sum of tail eigenvalues not covered by the sketch, they are also reduced by a factor of $\sqrt{d/m}$. Thus, the regret bound of the sketched algorithms could be better than the regret bound of the non-sketched counterpart if sketch size m is properly chosen, i.e., when $(1+\varepsilon_m)^3 m < d$.

Our main contribution in this chapter is three-fold. Firstly, we propose two algorithms for the linear MDP problem that enjoy improved computational efficiency by combining Frequent Directions sketching with Least-Squares Value Iteration algorithms. Secondly, in contrast to previous works, we theoretically show that the regret bound of the sketched algorithms is better than the regret bound of the non-sketched counterpart in certain regimes. Lastly, we present the first simulation study on the linear MDP problem and verify our theoretical result by observing improved regret for sketched LSVI algorithms in two different environments. Remarkably, in a high-dimensional environment, sketched LSVI algorithms can yield superior performance while using only 30% of the space and time compared to the non-sketched counterpart.

4.2 Preliminaries

In an episodic finite-horizon MDP denoted by the tuple $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$, there is a nonempty set of states $\mathcal{S}$ (measurable but possibly infinite) and a finite set of actions $\mathcal{A}$ with cardinality $A > 0$. $H \in \mathbb{Z}_+$ denotes the length of each episode. $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$ and $r = \{r_h\}_{h=1}^H$ are the transition probabilities and the reward functions. For each $h \in [H]$, let $\mathbb{P}_h(s'|s_h, a_h)$ and $r_h(s_h, a_h) \in [0,1]$ [2] denote the probability of transitioning to state s' and reward if action a_h is taken at state s_h and step h.

[2] Even though reward is assumed to be deterministic, results can be easily generalized to stochastic reward setting.

The agent aims to learn the optimal policy by interacting with the environment for a fixed number of K episodes. A policy of an agent is a function $\pi : \mathcal{S} \times [H] \to \mathcal{A}$ where $\pi_h(s)$ denotes the action that the agent will take at state s and step h. We use $V_h^\pi : \mathcal{S} \to \mathbb{R}$ to denote the value function of a policy π, i.e.,

$$V_h^\pi(s) \stackrel{\text{def}}{=} \mathbb{E}\left[\sum_{t=h}^{H} r_t(s_t, \pi_t(s_t)) \middle| s_h = s\right].$$

We use $Q_h^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ to denote the action-value function of a policy π, i.e.,

$$Q_h^\pi(s,a) \stackrel{\text{def}}{=} r_h(s,a) + \mathbb{E}\left[\sum_{t=h+1}^{H} r_t(s_t, \pi_t(s_t)) \middle| s_h = s, a_h = a\right].$$

As both the episode length H and the action set $\mathcal{A}$ are finite, there always exists an optimal policy $\pi^\star$ which gives the optimal value $V_h^\star(s) \stackrel{\text{def}}{=} \sup_\pi V_h^\pi(s)$ and $Q_h^\star(s,a) \stackrel{\text{def}}{=} \sup_\pi Q_h^\pi(s,a)$ [Puterman, 1994].

For each $k \geq 1$, the initial state s_{1k} is set by an adversary at the beginning of the kth episode. Without knowledge of the reward function or the transition probabilities, the agent chooses a policy π_k based on past observations. For a given positive number of episode K, the performance of the agent is evaluated in terms of (expected) regret, defined as

$$\text{Regret}(K) \stackrel{\text{def}}{=} \sum_{k=1}^{K} \left[V_1^\star(s_{1k}) - V_1^{\pi_k}(s_{1k})\right].$$

We focus on linear MDPs, i.e., MDPs in which both the transition dynamics and the reward are linear in some features of state and action pair. Formally, we make the following assumption:

Assumption 4.2.1 ([Jin et al., 2019, Zanette et al., 2019])**.** *MDP($\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r$) is a* linear MDP *if for any $h \in [H]$, there exists feature map $\phi_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, $\psi_h : \mathcal{S} \to \mathbb{R}^d$, and*

parameter $\theta_h^r \in \mathbb{R}^d$ such that for any $(s, s', a) \in \mathcal{S} \times \mathcal{S} \times \mathcal{A}$, we have

$$\mathbb{P}_h(s'|s,a) = \phi_h(s,a)^\top \psi_h(s'), \quad r_h(s,a) = \phi_h(s,a)^\top \theta_h^r.$$

Note that ϕ_h is known but ψ_h and θ_h^r are unknown. Moreover, we assume $\|\phi_h(s,a)\| \le L_\phi$ for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, $\int_s \|\psi_h(s)\| \le L_\psi$ for all $h \in [H]$, and $\|\theta_h^r\| \le L_r$ for all $h \in [H]$.

Note that such linear structure exists for all tabular MDPs by setting ϕ to be the canonical basis in $\mathbb{R}^d$ where $d = |\mathcal{S}| \times |\mathcal{A}|$. A crucial property of the linear MDP is that, for all policies including the optimal policy, the action-value functions are always linear in the feature map ϕ (Lemma 4.10.1). Under mild conditions, the linear transition assumption is necessary for all policies with linear action-value functions to have zero Bellman error [Jin et al., 2019].

Notation We use T to denote the total number of time steps, i.e., $T \stackrel{\text{def}}{=} KH > 0$. We use s_{hk} and a_{hk} to denote the state encountered and action taken at step h in episode k. We write $\phi_{hk} \stackrel{\text{def}}{=} \phi_h(s_{hk}, a_{hk})$, $r_{hk} \stackrel{\text{def}}{=} r_h(s_{hk}, a_{hk})$. Given a positive definite $d \times d$ matrix $\mathbf{A}$ and $x \in \mathbb{R}^d$, we denote $\|x\|_{\mathbf{A}} \stackrel{\text{def}}{=} \sqrt{x^\top \mathbf{A} x}$. Sometimes we use the shorthand $\mathbb{E}_{s'|s,a}[\cdot] \stackrel{\text{def}}{=} \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot|s,a)}[\cdot]$ where h can be deduced from context.

4.3 LSVI Algorithms

Value iteration is a simple algorithm that finds the optimal policy by solving the Bellman optimality equation recursively: $Q_h^\star(s,a) \leftarrow r_h(s,a) + \mathbb{E}_{s'|s,a}\left[\max_{a' \in \mathcal{A}} Q_{h+1}^\star(s', a')\right]$. When $Q_h^\star(s,a)$ is parameterized by a linear form $\phi_h(s,a)^\top \theta_h^\star$, we can apply the idea from Least-Squares Value Iteration (LSVI) [Bradtke and Barto, 1996], and solve for θ_{hk} in a

regularized least-squares problem:

$$\theta_{hk} \leftarrow \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{k-1} \left[r_{hi} + V_{h+1,k}(s_{h+1,i}) - \phi_{hi}^\top \theta \right]^2 + \lambda \|\theta\|^2 \tag{4.1}$$

where $V_{h+1,k}$ is some estimate of the value function V_{h+1} during episode k. The solution to the least square problem (4.1) is given as

$$\widehat{\theta}_{hk} \leftarrow \mathbf{\Sigma}_{h,k-1}^{-1} \left(\sum_{i=1}^{k-1} \phi_{hi} \left[r_{hi} + V_{h+1,k}(s_{h+1,i}) \right] \right), \tag{4.2}$$

where $\mathbf{\Sigma}_{h,k-1} \stackrel{\text{def}}{=} \sum_{i=1}^{k-1} \phi_{hi} \phi_{hi}^\top + \lambda \mathbf{I}_d$. Using $\phi_h(s,a)^\top \widehat{\theta}_{hk}$ as an estimate of the optimal action-value function, a greedy algorithm would simply follow the policy that chooses an action to maximize the estimated action-value function at each state. But reinforcement learning problems typically require a careful balance of exploration and exploitation, and it turns out that pure exploitation without exploration is not sufficient to guarantee low regret. To enforce exploration, LSVI-UCB additionally adds an exploration bonus term so that with high probability, the estimated action-value is an overestimation of the optimal action-value. In contrast, RLSVI adds perturbation, and exploration is achieved by carefully tuning the scale of the perturbation.

4.3.1 LSVI-UCB

LSVI-UCB (Algorithm 1) enforces exploration by enforcing optimism from exploration bonus (Upper-Confidence Bounds) of the form $\beta(\phi \mathbf{\Sigma}_{h,k-1}^{-1} \phi)^{1/2}$ to compensate for the uncertainty along the ϕ direction. With properly chosen scalar β, it can be shown that our estimated action-value function $Q_{h,k}$ will be an upper bound of Q_h^* for all state and action pairs. The regret guarantee of LSVI-UCB is given in Theorem 4.3.1.

Theorem 4.3.1 ([Jin et al., 2019]). *Under Assumption 4.2.1 with $L_\phi = 1$ and $L_\psi = L_r =$*

Algorithm 1: LSVI-UCB [Jin et al., 2019]

Define $V_{H+1,k}(s) \stackrel{\text{def}}{=} 0, Q_{H+1,k}(s,a) \stackrel{\text{def}}{=} 0$ and $V_{hk}(s) \stackrel{\text{def}}{=} \max_a Q_{hk}(s,a)$, with

$$Q_{hk}(s,a) \stackrel{\text{def}}{=} \min\left\{\phi_h(s,a)^\top \widehat{\theta}_{hk} + \beta\left[\phi_h(s,a)^\top \mathbf{\Sigma}_{h,k-1}^{-1} \phi_h(s,a)\right]^{1/2}, H\right\} \forall (s,a,h,k).$$

Initialize $\mathbf{\Sigma}_{h0} \leftarrow \lambda \mathbf{I}_d, \forall h$.
for episode $k = 1, 2, \ldots, K$ **do**
  Receive the initial state s_{1k}.
  for step $h = H, H-1 \ldots, 1$ **do**
    Estimate $\widehat{\theta}_{hk}$ as in (4.2) and let $\widehat{\theta}_{hk}$ be $\mathbf{\Sigma}_{h,k-1}^{-1}\left(\sum_{i=1}^{k-1} \phi_{hi}\left[r_{hi} + V_{h+1,k}(s_{h+1,i})\right]\right)$.
  end for
  for step $h = 1, \ldots, H$ **do**
    Take action $a_{hk} \leftarrow \arg\max_{a \in \mathcal{A}} Q_{hk}(s_{hk}, a)$, and observe $s_{h+1,k}$.
    Update $\mathbf{\Sigma}_{h,k} \leftarrow \mathbf{\Sigma}_{h,k-1} + \phi_{hk}\phi_{hk}^\top$.
  end for
end for

$\sqrt{d}$, *there exists an absolute constant* $C > 0$ *such that, for any* $\delta \in (0,1)$, *if we set* $\lambda = 1$ *and* $\beta = c_\beta \cdot dH\sqrt{\iota}$ *for* $c_\beta > C$ *in Algorithm 1 with* $\iota \overset{\text{def}}{=} \log(2dT/\delta)$, *then with probability at least* $1-\delta$, *the total regret of LSVI-UCB is at most* $\widetilde{\mathcal{O}}(\sqrt{d^3H^3T})$.

4.3.2 RLSVI

Similar to the linear bandit setting [Agrawal and Goyal, 2013, Abeille and Lazaric, 2017], in RLSVI (Algorithm 2) the variance of the perturbations, controlled by scalar σ^2, is carefully chosen to guarantee that the value estimates are optimistic with at least constant probability. We have the following guarantee on the regret of RLSVI.

Theorem 4.3.2 ([Zanette et al., 2019]). *Under Assumption 4.2.1, if we set* $\sigma = \sqrt{H}\left(\widetilde{\mathcal{O}}(Hd) + L_\phi(3HL_\psi + L_r)\right), \lambda = 1, \alpha_U = 1/\widetilde{\mathcal{O}}(\sigma\sqrt{d})$ *and* $\alpha_L = \alpha_U/2$ *in Algorithm 2, then for any* $0 < \delta < \Phi(-1)/2$ [3], *with probability at least* $1-\delta$, *the total regret of RLSVI is at most* $\widetilde{\mathcal{O}}(\sigma dH\sqrt{K})$. *If we further assume that* $L_\phi = \widetilde{\mathcal{O}}(1)$ *and* $L_r, L_\psi = \widetilde{\mathcal{O}}(d)$, *then the bound reduces to* $\widetilde{\mathcal{O}}(H^2d^2\sqrt{T})$.

Definition 4.3.3 (RLSVI Q function). *For some constants* α_L, α_U *where* $\alpha_L < \alpha_U$ *and using shorthand* $q \overset{\text{def}}{=} \left(\|\phi_h(s,a)\|_{\mathbf{\Sigma}_{h,k-1}^{-1}} - \alpha_L\right)/(\alpha_U - \alpha_L)$, $B_h \overset{\text{def}}{=} H - h + 1$, *define:*

$$\overline{Q}_{hk}(s,a) \overset{\text{def}}{=} \begin{cases} \phi_h(s,a)^\top \overline{\theta}_{hk}, & \text{if } \|\phi_h(s,a)\|_{\mathbf{\Sigma}_{h,k-1}^{-1}} \le \alpha_L, \\ B_h, & \text{if } \|\phi_h(s,a)\|_{\mathbf{\Sigma}_{h,k-1}^{-1}} \ge \alpha_U, \\ q \cdot \phi_h(s,a)^\top \overline{\theta}_{hk} + (1-q)B_h & \text{otherwise.} \end{cases}$$

Complexity Analysis Algorithm 1 and Algorithm 2 only need to store $\mathbf{\Sigma}_{h,k-1}$ at episode k and $r_{hk}, \{\phi_h(s_{hk},a)\}_{a\in\mathcal{A}}$ for all (h,k), which takes $\mathcal{O}(d^2H + dAT)$ space. The time complexity is dominated by computing $\widehat{\theta}_{hk}$ which takes $\mathcal{O}(d^2AK)$ time per step using Sherman-Morrison formula to update $\mathbf{\Sigma}_{hk}^{-1}$. Thus, the total runtime is $\mathcal{O}(d^2AKT)$.

[3] $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable

Algorithm 2: RLSVI [Zanette et al., 2019]

Define $\overline{V}_{H+1,k}(s) \stackrel{\text{def}}{=} 0, \overline{Q}_{H+1,k}(s,a) \stackrel{\text{def}}{=} 0$ and $\overline{V}_{hk}(s) \stackrel{\text{def}}{=} \max_a \overline{Q}_{hk}(s,a)$, with $\overline{Q}_{hk}(s,a)$ defined in Definition 4.3.3, $\forall (s,a,h,k)$.
Initialize $\mathbf{\Sigma}_{h0} \leftarrow \lambda \mathbf{I}_d, \forall h$.
for episode $k = 1, 2, \ldots, K$ **do**
  Receive the initial state s_{1k}.
  for step $h = H, H-1, \ldots, 1$ **do**
    Estimate $\widehat{\theta}_{hk}$ as in (4.2) with $\overline{V}_{h+1,k}$ in place of $V_{h+1,k}$. In other words, let $\widehat{\theta}_{hk}$ be $\mathbf{\Sigma}_{h,k-1}^{-1}\left(\sum_{i=1}^{k-1} \phi_{hi}\left[r_{hi} + \overline{V}_{h+1,k}(s_{h+1,i})\right]\right)$.
    Sample $\overline{\xi}_{hk} \sim \mathcal{N}(0, \sigma^2 \mathbf{\Sigma}_{h,k-1}^{-1})$.
    $\overline{\theta}_{hk} \leftarrow \widehat{\theta}_{hk} + \overline{\xi}_{hk}$.
  end for
  for step $t = 1, \ldots, H$ **do**
    Take action $a_{hk} \leftarrow \arg\max_{a \in \mathcal{A}} \overline{Q}_{hk}(s_{hk}, a)$, and observe $s_{h+1,k}$.
    Update $\mathbf{\Sigma}_{hk} \leftarrow \mathbf{\Sigma}_{h,k-1} + \phi_{hk}\phi_{hk}^\top$.
  end for
end for

4.4 Matrix Sketching

The $\mathcal{O}(d^2)$ complexity of both algorithms is due to computing the inverse correlation matrix $\mathbf{\Sigma}_{hk}^{-1}$. Let $\mathbf{\Phi}_{hk} \stackrel{\text{def}}{=} [\phi_{h1}, \ldots, \phi_{hk}]^\top \in \mathbb{R}^{k \times d}$ so that $\mathbf{\Sigma}_{hk} = \sum_{i=1}^{k} \phi_{hi}\phi_{hi}^\top + \lambda \mathbf{I}_d = \mathbf{\Phi}_{hk}^\top \mathbf{\Phi}_{hk} + \lambda \mathbf{I}_d$. To reduce the computational complexity, matrix sketching techniques can be applied to maintain an appropriate matrix $\mathbf{S}_{hk}$ such that $\mathbf{S}_{hk}^\top \mathbf{S}_{hk}$ is close to $\mathbf{\Phi}_{hk}^\top \mathbf{\Phi}_{hk}$.

4.4.1 Frequent Directions

Frequent Directions (FD) [Liberty, 2013, Ghashami et al., 2016] is a deterministic matrix sketching algorithm that maintains a matrix $\mathbf{S}_{hk} \in \mathbb{R}^{m \times d}$, where constant $m < d$ is called the sketch size. It has the property that $\mathbf{S}_{hk}$ can be efficiently updated in time $\mathcal{O}(md)$. Moreover, using the Woodbury identity we may write,

$$\widetilde{\mathbf{\Sigma}}_{hk}^{-1} = \left(\mathbf{S}_{hk}^\top \mathbf{S}_{hk} + \lambda \mathbf{I}_d\right)^{-1} = \frac{1}{\lambda}\left(\mathbf{I}_d - \mathbf{S}_{hk}^\top \mathbf{H}_{hk} \mathbf{S}_{hk}\right)$$

where $\mathbf{H}_{hk} \overset{\text{def}}{=} \left(\mathbf{S}_{hk}\mathbf{S}_{hk}^\top + \lambda \mathbf{I}_m\right)^{-1}$. Note that matrix product involving $\mathbf{S}_{hk}$ and $\mathbf{H}_{hk}$ requires time $\mathcal{O}(md)$ and $\mathcal{O}(m^2)$. Therefore, matrix product involving $\widetilde{\mathbf{\Sigma}}_{hk}^{-1}$ takes $\mathcal{O}(md)$ time to compute. The updates of $\mathbf{S}_{hk}$ and $\mathbf{H}_{hk}$ are summarized in Algorithm 3. Since $\mathbf{S}_{hk}$ and $\mathbf{H}_{hk}$ can be updated in $\mathcal{O}(md)$ time (see section 3.2 of [Ghashami et al., 2016]), the total run time will be reduced from $\mathcal{O}(d^2)$ to $\mathcal{O}(md)$.

Algorithm 3: FD Sketching

1: **Input:** $\mathbf{S}_{h,k-1} \in \mathbb{R}^{m\times d}, \phi_{hk} \in \mathbb{R}^d, \lambda > 0$.
2: Compute eigendecomposition
$\mathbf{U}_{hk}^\top \text{diag}(a_1, \ldots, a_m)\mathbf{U}_{hk} = (\mathbf{S}_{h,k-1})^\top \mathbf{S}_{h,k-1} + \phi_{hk}\phi_{hk}^\top$.
3: $\mathbf{S}_{hk} \leftarrow \text{diag}(\sqrt{a_1 - a_m}, \ldots, \sqrt{a_{m-1} - a_m}, 0)\mathbf{U}_{hk}$, $\mathbf{H}_{hk} \leftarrow \text{diag}(\frac{1}{a_1 - a_m + \lambda}, \ldots, \frac{1}{\lambda})$.
4: **Output:** $\mathbf{S}_{hk}, \mathbf{H}_{hk}$.

The sacrificed accuracy by running FD sketching is related to the tail spectrum of the correlation matrix $\mathbf{\Phi}_{hK}^\top \mathbf{\Phi}_{hK}$. Mathematically, it is captured by the spectral error $\varepsilon_{h,m}$ defined as

$$\varepsilon_{h,m} \overset{\text{def}}{=} \min_{k=0,\ldots,m-1} \frac{\gamma_{k+1} + \gamma_{k+2} + \cdots + \gamma_d}{\lambda(m-k)} \tag{4.3}$$

where $\gamma_1 \geq \cdots \geq \gamma_d$ are the eigenvalues of the correlation matrix $\mathbf{\Phi}_{hK}^\top \mathbf{\Phi}_{hK}$. Observing that $\varepsilon_{h,m} \leq (\gamma_m + \cdots + \gamma_d)/\lambda$, if the matrix $\mathbf{\Phi}_{hK}^\top \mathbf{\Phi}_{hK}$ has low rank or light-tailed spectrum, we would expect this spectral error to be small.

In the sequel, we will use a shorthand

$$\widetilde{m}_h \overset{\text{def}}{=} d\log(1 + \varepsilon_{h,m}) + m\log\left(1 + \frac{KL_\phi^2}{m\lambda}\right). \tag{4.4}$$

Note that if $\varepsilon_{h,m} = 0$, $\widetilde{m}_h$ grows nearly linearly in m, whereas if $\varepsilon_{h,m}$ is large, $\widetilde{m}_h$ grows nearly linearly in d.

4.5 Sketched LSVI Algorithms

4.5.1 Sketched LSVI-UCB

Algorithm 4: S-LSVI-UCB

Define $V_{H+1,k}(s) \stackrel{\text{def}}{=} 0, Q_{H+1,k}(s,a) \stackrel{\text{def}}{=} 0$ and $V_{hk}(s) \stackrel{\text{def}}{=} \max_a Q_{hk}(s,a)$, with

$$Q_{hk}(s,a) \stackrel{\text{def}}{=} \min\left\{\phi_h(s,a)^\top \widehat{\theta}_{hk} + \beta_h \sqrt{\phi_h(s,a)^\top \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1} \phi_h(s,a)}, H\right\}, \ \forall (s,a,h,k).$$

Initialize $\widetilde{\mathbf{\Sigma}}_{h0}^{-1} \leftarrow \frac{1}{\lambda}\mathbf{I}_d, \ \mathbf{S}_{h0} \leftarrow 0, \ \forall h.$
for episode $k = 1, 2, \ldots, K$ **do**
  Receive the initial state s_{1k}.
  for step $h = H, H-1 \ldots, 1$ **do**
    Let $\widehat{\theta}_{hk}$ be estimated as $\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\left(\sum_{i=1}^{k-1} \phi_{hi}\left[r_{hi} + V_{h+1,k}(s_{h+1,i})\right]\right)$.
  end for
  for step $h = 1, \ldots, H$ **do**
    Take action $a_{hk} \leftarrow \arg\max_{a \in \mathcal{A}} Q_{hk}(s_{hk}, a)$, and observe $s_{h+1,k}$.
    Compute $\mathbf{S}_{hk}, \mathbf{H}_{hk}$ given $\mathbf{S}_{h,k-1}, \phi_{hk}$ with Alg 3. Update $\widetilde{\mathbf{\Sigma}}_{hk}^{-1} \leftarrow \frac{1}{\lambda}\left(\mathbf{I}_d - \mathbf{S}_{hk}^\top \mathbf{H}_{hk} \mathbf{S}_{hk}\right)$.
  end for
end for

S-LSVI-UCB, the FD-sketched counterpart of LSVI-UCB, is shown in Algorithm 4. S-LSVI-UCB enjoys the following regret bound, characterized in terms of the spectral error $\varepsilon_{h,m}$.

Theorem 4.5.1. *Under Assumption 4.2.1, there exists an absolute constant $C > 0$ such that, for any $\delta \in (0,1)$, if we set $\beta_h = c_\beta \cdot HL_dL_\lambda(1+\varepsilon_{h,m})\sqrt{m\iota}$ for any $c_\beta > C$ in Algorithm 4 with $L_d \stackrel{\text{def}}{=} \max\left\{\sqrt{d}, L_\psi, \frac{L_r}{H}\right\}, L_\lambda \stackrel{\text{def}}{=} \max\{1, \sqrt{\lambda}\}$, $\iota \stackrel{\text{def}}{=} \log\left(\left(2L_dL_\lambda \max\{1, L_\phi\} dT\right)/(\sqrt{\lambda}\delta)\right)$, then with probability at least $1-\delta$, the total regret of S-LSVI-UCB is at most $\widetilde{\mathcal{O}}\left(\sum_{h=1}^H L_dL_\lambda\sqrt{(1+\varepsilon_{h,m})^3 m\widetilde{m}_h \cdot HT}\right)$. If we further*

assume that $\lambda = 1, L_\phi = 1$ *and* $L_\psi = L_r = \sqrt{d}$*, and denote* $\varepsilon_m = \max_h\{\varepsilon_{h,m}\}$, $\widetilde{m} = \max_h\{\widetilde{m}_h\}$*, then the regret bound can be simplified as* $\widetilde{\mathcal{O}}\left(\sqrt{(1+\varepsilon_m)^3 md\widetilde{m} \cdot H^3T}\right)$.

Note that the regret of S-LSVI-UCB is not directly comparable to LSVI-UCB in its current form because of the presence of $\widetilde{m}$. However, the motivation behind the introduction of $\widetilde{m}_h$ in Equation (4.4) was to show potentially better dependency on the dimension (linear in m rather than d) when the spectral error is small. By Lemma 4.7.2, we can trivially replace any occurrence of $\widetilde{m}_h$ with d (after ignoring logarithmic dependency on other quantities) in the asymptotic regret bound. In other words, $\widetilde{\mathcal{O}}\left(\sqrt{(1+\varepsilon_m)^3 md^2 \cdot H^3T}\right)$ is also a valid regret bound for S-LSVI-UCB. Comparing the two regret bounds, we see that the sketching technique inflates the regret bound by a factor of $(1+\varepsilon_m)^{3/2}$. However, it also reduces the dependency on the dimension by a factor of $\sqrt{d/m}$. Therefore, as long as we are in the regime of $(1+\varepsilon_m)^3 m < d$, the sketching technique would bring a reduction in regret, on top of the improved time and space complexity. The potential improvement in regret is not observed in the results for many previous works on sketching for online regret minimization problems, where the regret bounds are only inflated by some factors [Luo et al., 2016, Calandriello et al., 2017, Luo et al., 2019, Kuzborskij et al., 2019, Calandriello et al., 2019]. The potential improvement comes from the fact that the regret bound depends on the covering number of a class of function that contains the random value function $V_{hk}(\cdot)$, which is parameterized by $\widehat{\theta}_{hk}$ and the correlation matrix $\boldsymbol{\Sigma}_{h,k-1}$. When the function class is parameterized by the sketched correlation matrix $\widetilde{\boldsymbol{\Sigma}}_{h,k-1}$ instead, the inherent low-rank structure reduces the covering number and saves a factor of $\sqrt{d/m}$ in the dimension.

4.5.2 Sketched RLSVI

S-RLSVI, the FD-sketched counterpart of RLSVI, is shown in Algorithm 5. Note that in addition to $\widetilde{\boldsymbol{\Sigma}}_{hk}^{-1}$, the algorithm also needs to compute $\widetilde{\boldsymbol{\Sigma}}_{h,k-1}^{-1/2}$ to efficiently sample from

Algorithm 5: S-RLSVI

Define $\overline{V}_{H+1,k}(s) \stackrel{\text{def}}{=} 0, \overline{Q}_{H+1,k}(s,a) \stackrel{\text{def}}{=} 0$ and $\overline{V}_{hk}(s) \stackrel{\text{def}}{=} \max_a \overline{Q}_{hk}(s,a)$, with $\overline{Q}_{hk}(s,a)$ defined in Definition 4.3.3 but with $\widetilde{\mathbf{\Sigma}}_{h,k-1}$ in place of $\mathbf{\Sigma}_{h,k-1}$, $\forall(s,a,h,k)$.
Initialize $\widetilde{\mathbf{\Sigma}}_{h0}^{-1} \leftarrow \frac{1}{\lambda}\mathbf{I}_d,\ \mathbf{S}_{h0} \leftarrow 0,\ \forall h$.
for episode $k = 1, 2, \ldots, K$ **do**
  Receive the initial state s_{1k}.
  for step $h = H, H-1, \ldots, 1$ **do**
    Let $\widehat{\theta}_{hk}$ be estimated as $\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\left(\sum_{i=1}^{k-1}\phi_{hi}\left[r_{hi} + \overline{V}_{h+1,k}(s_{h+1,i})\right]\right)$.
    Sample $\overline{\xi}_{hk} \sim \mathcal{N}(0, \sigma^2\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1})$.
    $\overline{\theta}_{hk} \leftarrow \widehat{\theta}_{hk} + \overline{\xi}_{hk}$.
  end for
  for step $t = 1, \ldots, H$ **do**
    Take action $a_{hk} \leftarrow \arg\max_{a\in\mathcal{A}} \overline{Q}_{hk}(s_{hk}, a)$, and observe $s_{h+1,k}$.
    Compute $\mathbf{S}_{hk}, \mathbf{H}_{hk}$ given $\mathbf{S}_{h,k-1}, \phi_{hk}$ with Alg 3. Update $\widetilde{\mathbf{\Sigma}}_{hk}^{-1} \leftarrow \frac{1}{\lambda}\left(\mathbf{I}_d - \mathbf{S}_{hk}^\top \mathbf{H}_{hk}\mathbf{S}_{hk}\right)$.
  end for
end for

$\mathcal{N}(0, \sigma^2\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1})$. This can be done through the generalized Woodbury identity (Corollary 1 of [Kuzborskij et al., 2019]).

The regret bound of S-RLSVI is stated in the Theorem 4.5.2. Note that the dependence of $\sigma, \alpha_U, \alpha_L$ on h in notation are omitted to reduce clutter.

Theorem 4.5.2. *Under Assumption 4.2.1, if we set* $\sigma = \widetilde{\mathcal{O}}\left(H^{3/2}L_d L_\lambda(1+\varepsilon_{h,m})\right)$ *where* $L_d = \max\left\{\sqrt{md}, L_\psi, \frac{L_r}{H}\right\}$, $L_\lambda = \max\{1, \sqrt{\lambda}\}$ $\alpha_U = 1/\widetilde{\mathcal{O}}(\sigma\sqrt{d})$ *and* $\alpha_L = \alpha_U/2$ *in Algorithm 5, then for any* $0 < \delta < \Phi(-1)/2$, *with probability at least* $1-\delta$, *the total regret of S-RLSVI is at most* $\widetilde{\mathcal{O}}\left(\sum_{h=1}^H \sigma\sqrt{(1+\varepsilon_{h,m})d\widetilde{m}_h K}\right)$. *If we further assume that* $\lambda = 1$ *and* $L_r, L_\psi = \widetilde{\mathcal{O}}(\sqrt{md})$, *and denote* $\varepsilon_m = \max_h\{\varepsilon_{h,m}\}$, $\widetilde{m} = \max_h\{\widetilde{m}_h\}$. *Then the regret bound can be simplified as* $\widetilde{\mathcal{O}}\left(\sqrt{(1+\varepsilon_m)^3 md^2\widetilde{m}\cdot H^4 T}\right)$.

Similar to the situation for S-LSVI-UCB, we can use the regret bound of $\widetilde{\mathcal{O}}\left(\sqrt{(1+\varepsilon_m)^3 md^3\cdot H^4 T}\right)$ for straightforward comparison. Comparing to the bound of RLSVI, we see that the sketching technique inflates the regret by a factor of $(1+$

$\varepsilon_m)^{3/2}$ but also reduces the dependency on the dimension by a factor of $\sqrt{d/m}$, just like the comparison between bounds for LSVI-UCB and S-LSVI-UCB. Again, as long as the spectral error ε_m is relatively small compared to $\sqrt{d/m}$, the sketching technique would result in decreased regret, in addition to improved efficiency.

Complexity Analysis Algorithm 4 and Algorithm 5 only need to store $\mathbf{S}_{h,k-1}$ and $\mathbf{H}_{h,k-1}$ at episode k and $r_{hk}, \{\phi_h(s_{hk}, a)\}_{a\in\mathcal{A}}$ for all (h, k), which takes $\mathcal{O}(mdH + dAT)$ space. The total runtime is improved from $\mathcal{O}(d^2AKT)$ to $\mathcal{O}(mdAKT)$ as discussed in Section 4.4.

4.6 Experiments

We empirically evaluate the performance of the two LSVI algorithms and their sketched counterparts in the two environments described below. For both environments, we consider two configurations to demonstrate the applicability of the sketched algorithms. The first environment is a commonly studied environment in the literature that has a finite number of states and low feature dimensions. The second environment is a carefully designed environment with an infinite number of states and a high feature dimension. While many previous works have studied the linear MDP and its closely related variants, most of them are purely theoretical [Jin et al., 2019, Zanette et al., 2019, Yang and Wang, 2019a, Cai et al., 2019]. To the best of our knowledge, this chapter is the first that presents simulation experiments of algorithms designed to solve the linear MDP problem or similar problems.

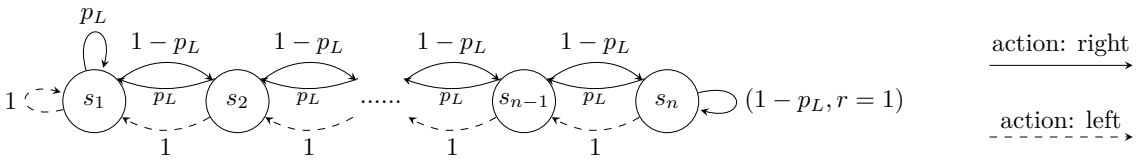

Figure 4.1: RiverSwim Environment.

4.6.1 Setup

RiverSwim Environment This environment has been studied by Strehl and Littman [2008], Osband et al. [2013], Osband and Van Roy [2017], Efroni et al. [2019]. The environment consists of n states arranged in a chain. Starting from the leftmost state, the agent can choose to swim left or right at each step. Swimming left always succeeds and results in moving to the state on the left, but swimming right fails with probability p_L, in which case the agent moves to the left. If the agent is at the right side of the chain and tries to move right, it will receive a reward $r = 1$. A graphical illustration of the environment can be found in Figure 4.1. When the episode length is not smaller than the number of states, the optimal policy in this environment is to keep swimming right. Instead of the canonical tabular encoding of the environment's linear structure with a feature dimension $d = 2n$ to indicate each state and action pair, we devised a compact encoding of the linear structure to reduce the feature dimension from $2n$ to $n + 1$.

Labyrinth Environment Consider an agent walking on $\mathbb{N}_0 = \{0, 1, 2, \dots\}$ in a game with episode length equal to H where each number indicates a different state. Starting at $s_{\text{start}} = 0$, upon choosing an action the agent will optionally receive a diminutive auxiliary reward, and then transit to a randomly chosen state indicated by a non-negative integer at each step. However, the agent will incur a substantial reward only if it arrives at a specific state s_{goal} and take a specific action a_{goal} here. The environment is designed such that there is a shortcut (a sequence of state-action pairs) for the agent to arrive at s_{goal} deterministically within $H - 1$ steps. The agent's optimal solution is to find the shortcut in this Labyrinth of infinite states and exploit it.

Details of the encodings of the linear structure in both environments are deferred to Section 4.11.

Implementation The regularization parameter λ is always set to 1. The hyper-parameters β and σ are selected by cross validation. We test the sketched algorithms with sketch size approximating 80%, 60%, and 50% or 40% of the feature dimension for the RiverSwim environment; 90%, 70%, 50%, and 30% for the Labyrinth environment. Results are averaged over 20 random seeds.

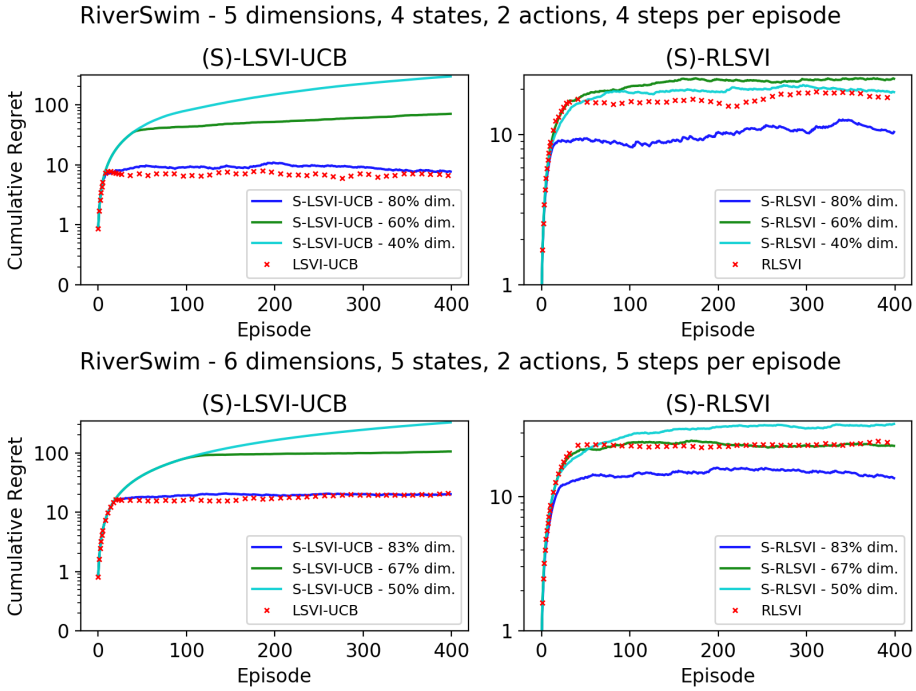

Figure 4.2: Cumulative Regret for RiverSwim Environment.

4.6.2 Results

Figure 4.2 and 4.3 illustrate the evolution of the cumulative regret for the two LSVI algorithms and their sketched counterparts, with various sketch sizes, using a logarithmic scale.

Riverswim Environment The RiverSwim environment is a low-dimensional environment. Combined with our compact design of the encoding, it is plausible that all feature dimensions are important to preserve information crucial for exploration and exploitation.

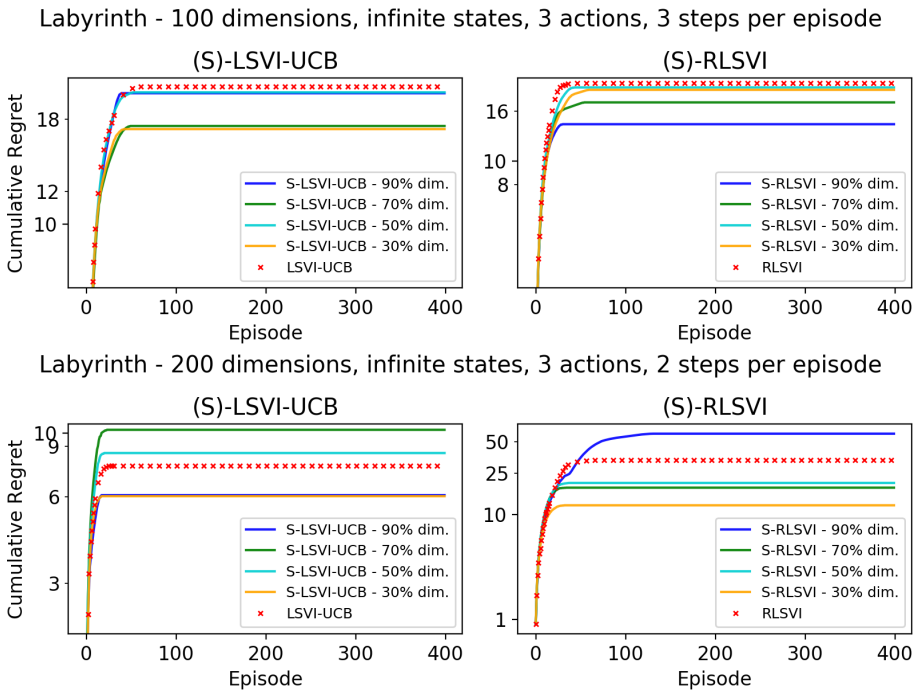

Figure 4.3: Cumulative Regret for Labyrinth Environment. First row displays the environment with no auxiliary reward; second row displays the environment with auxiliary reward of 3×10^{-4} for each action.

Therefore, aggressive sketching that uses a sketch size of only 40% of the dimension yields noticeably worse performance. However, even in this scenario, comparable or better performance can still be achieved by using a sketch size of around 80% of the dimension, and sometimes even around 60% suffices.

Labyrinth Environment The Labyrinth environment is high-dimensional and is in the regime where sketching is both more likely to be necessary and more likely to yield huge efficiency gains. As can be observed from the figure, for the configuration that no auxiliary reward is given, all sketched algorithms have superior performance compared to the non-sketched counterparts, regardless of the sketch size. Remarkably, for both configurations of the Labyrinth Environment, computational efficiency can be tripled by maintaining a sketch size of only 30% of the feature dimension, while at the same time improvement in performance can be consistently observed.

Summary The simulation results demonstrate that the sketched algorithms are particularly advantageous in the high-dimensional environment, achieving better performance using only 30% of space and time compared to the non-sketched counterpart. On the other hand, even in a low-dimensional environment, sketching can still achieve comparable performance using only 60% or 80% of the space and time.

4.7 Sketching Lemmas and Proofs

This section contains lemmas related to the correlation matrix $\mathbf{\Phi}_{hk}^\top \mathbf{\Phi}_{hk}$ and its FD-sketched estimates $\mathbf{S}_{hk}^\top \mathbf{S}_{hk}$ with sketch size m. In the rest of this section we drop the subscript h because all results hold for a generic $h \in [H]$.

Let ρ_k be the smallest eigenvalue of the FD-sketched correlation matrix $\mathbf{S}_k^\top \mathbf{S}_k$ and let $\bar{\rho}_k = \rho_1 + \cdots + \rho_k$. Recall that $\mathbf{\Sigma}_k = \mathbf{\Phi}_k^\top \mathbf{\Phi}_k + \lambda \mathbf{I}_d$ and $\widetilde{\mathbf{\Sigma}}_k = \mathbf{S}_k^\top \mathbf{S}_k + \lambda \mathbf{I}_d$. Also, ε_m is defined in (4.3).

All following results hold for any $k = 0, \ldots, K$, any $\lambda > 0$, and any sketch size $m = 1, \ldots, d$ unless explicitly stated otherwise.

Lemma 4.7.1.

$$\varepsilon_m \leq \frac{K L_\phi^2}{m\lambda}.$$

Proof. Let $\lambda_1 \geq \cdots \geq \lambda_d$ be the eigenvalues of the correlation matrix $\mathbf{\Phi}_K^\top \mathbf{\Phi}_K$. By (4.3),

$$\varepsilon_m \leq \frac{\lambda_1 + \ldots \lambda_d}{m\lambda} = \frac{\mathrm{tr}(\mathbf{\Phi}_K^\top \mathbf{\Phi}_K)}{m\lambda} = \frac{\mathrm{tr}\left(\sum_{i=1}^K \phi_i \phi_i^\top\right)}{m\lambda} = \frac{\sum_{i=1}^K \mathrm{tr}\left(\phi_i \phi_i^\top\right)}{m\lambda} = \frac{\sum_{i=1}^K \|\phi_i\|^2}{m\lambda} \leq \frac{K L_\phi^2}{m\lambda}.$$

□

Lemma 4.7.2.

$$\widetilde{m}_h = \mathcal{O}(d).$$

Proof.

$$\begin{aligned}\widetilde{m}_h &= d\log(1+\varepsilon_{h,m}) + m\log\left(1+\frac{KL_\phi^2}{m\lambda}\right)\\ &\leq d\log\left(1+\frac{KL_\phi^2}{m\lambda}\right) + d\log\left(1+\frac{KL_\phi^2}{m\lambda}\right) = 2d\log\left(1+\frac{KL_\phi^2}{m\lambda}\right) = \mathcal{O}(d).\end{aligned}$$

□

Lemma 4.7.3 (Ghashami et al. [2016])**.**

$$\frac{\bar{\rho}_k}{\lambda} \leq \varepsilon_m.$$

Lemma 4.7.4 ([Kuzborskij et al., 2019, Proposition 3])**.**

$$\mathbf{\Phi}_k^\top \mathbf{\Phi}_k = \mathbf{S}_k^\top \mathbf{S}_k + \bar{\rho}_k \mathbf{I}_d.$$

Lemma 4.7.5 ([Kuzborskij et al., 2019, Lemma 8])**.**

$$\begin{aligned}\log\left(\frac{\det(\mathbf{\Sigma}_k)}{\det(\lambda\mathbf{I}_d)}\right) &\leq d\log\left(1+\frac{\bar{\rho}_k}{\lambda}\right) + m\log\left(1+\frac{kL_\phi^2}{m\lambda}\right)\\ &\leq d\log\left(1+\varepsilon_m\right) + m\log\left(1+\frac{kL_\phi^2}{m\lambda}\right).\end{aligned}$$

Lemma 4.7.6 ([Kuzborskij et al., 2019, Lemma 9])**.**

$$\sum_{k=1}^{K} \min\{1, \|\phi_k\|^2_{\widetilde{\mathbf{\Sigma}}_{k-1}^{-1}}\} \leq 2(1+\varepsilon_m)\left[d\log\left(1+\varepsilon_{h,m}\right) + m\log\left(1+\frac{KL_\phi^2}{m\lambda}\right)\right].$$

Lemma 4.7.7 ([Kuzborskij et al., 2019, Proof of Lemma 9])**.** $\forall x \in \mathbb{R}^d$,

$$\|x\|^2_{\widetilde{\mathbf{\Sigma}}_k^{-1}} \leq \frac{\lambda+\bar{\rho}_k}{\lambda}\|x\|^2_{\mathbf{\Sigma}_k^{-1}} \leq (1+\varepsilon_m)\|x\|^2_{\mathbf{\Sigma}_k^{-1}}.$$

4.8 Proof of Theorem 4.5.1

The proof in this section follows Jin et al. [2019], with modifications to account for the spectral error $\varepsilon_{h,m}$ because of the sketched correlation matrix $\widetilde{\Sigma}$. One key innovation in the proof lies in Lemma 4.8.2, where the bound is improved from $\mathcal{O}(d^2)$ to $\mathcal{O}(md)$ due to the reduced covering number. The covering number is reduced because the space of low rank matrices has smaller covering number than the counterpart without low rank assumption.

We first show that the weight vectors $\widehat{\theta}_{hk}$ in Algorithm 4 has bounded norm.

Lemma 4.8.1 (Bound on Weights of S-LSVI-UCB)**.** *For any $(h,k) \in [H] \times [K]$, the weights $\widehat{\theta}_{hk}$ in Algorithm 4 satisfies*

$$\|\widehat{\theta}_{hk}\| \le 2H\sqrt{\frac{(1+\varepsilon_{h,m})dk}{\lambda}}.$$

Proof. For any vector $v \in \mathbb{R}^d$, we have

$$\begin{aligned}
|v^\top \widehat{\theta}_{hk}| &= \left| v^\top \widetilde{\Sigma}_{h,k-1}^{-1} \sum_{i=1}^{k-1} \phi_{hi} \left[r_{hi} + V_{h+1,k}(s_{h+1,i}) \right] \right| \\
&\le \sum_{i=1}^{k-1} 2H \left| v^\top \widetilde{\Sigma}_{h,k-1}^{-1} \phi_{hi} \right| \\
&\le 2H \cdot \sqrt{\left[\sum_{i=1}^{k-1} v^\top \widetilde{\Sigma}_{h,k-1}^{-1} v \right] \cdot \left[\sum_{i=1}^{k-1} \phi_{hi}^\top \widetilde{\Sigma}_{h,k-1}^{-1} \phi_{hi} \right]} \\
&\overset{(a)}{\le} 2H \cdot \sqrt{\frac{k\|v\|^2}{\lambda} \cdot \left[(1+\varepsilon_{h,m}) \sum_{i=1}^{k-1} \phi_{hi}^\top \Sigma_{h,k-1}^{-1} \phi_{hi} \right]} \\
&\overset{(b)}{\le} 2H \sqrt{\frac{(1+\varepsilon_{h,m})k\|v\|^2}{\lambda} \cdot d} \\
&= 2H\|v\| \sqrt{\frac{(1+\varepsilon_{h,m})dk}{\lambda}}.
\end{aligned}$$

In the derivation, step (a) is due to the fact that the maximum eigenvalue of $\widetilde{\Sigma}_{h,k-1}^{-1}$ is no

more than $1/\lambda$ and Lemma 4.7.7. Step (b) is due to Lemma 4.10.3. The remainder of the proof follows from $\|\widehat{\theta}_{hk}\| = \max_{\|v\|=1} |v^\top \widehat{\theta}_{hk}|$. □

Second, we present a concentration lemma that is critical for bounding the difference between the estimated and true action-value functions. Importantly, the bound has improved dependence on dimension compared to [Jin et al., 2019, Lemma B.3].

Lemma 4.8.2. *Under the setting of Theorem 4.5.1, let c_β be the constant in our definition of β_h, i.e., $\beta_h = c_\beta \cdot HL_dL_\lambda(1+\varepsilon_{h,m})\sqrt{m\iota}$ where $L_d = \max\left\{\sqrt{d}, L_\psi, \frac{L_r}{H}\right\}, L_\lambda = \max\{1, \sqrt{\lambda}\}, \iota = \log\left(\frac{2dL_dL_\lambda \max\{1,L_\phi\}T}{\sqrt{\lambda}\delta}\right)$. There exists absolute constant c_0 that is independent of c_β such that for any fixed $\delta \in (0,1)$, the following event, denoted as $\mathcal{G}$,*

$$\forall (h,k) \in [H]\times[K] : \left\|\sum_{i=1}^{k-1} \phi_{hi}\left(V_{h+1,k}(s_{h+1,i}) - \mathbb{E}_{s'|s_{hi},a_{hi}}\left[V_{h+1,k}(s')\right]\right)\right\|_{\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}} \le c_0 \cdot HL_d\sqrt{(1+\varepsilon_{h,m})m\chi},$$

where $\chi \overset{\text{def}}{=} \log\left(\frac{2(c_\beta+1)L_dL_\lambda \max\{1,L_\phi\}dT}{\sqrt{\lambda}\delta}\right)$ happens with probability at least $1-\delta$.

Proof. We can write $V_{h+1,k}(\cdot)$ as

$$V_{h+1,k}(\cdot) = \min\Bigg\{\max_a\{\phi_{h+1}(\cdot,a)^\top \widehat{\theta}_{h+1,k} + \beta_{h+1}\sqrt{\phi_{h+1}(\cdot,a)^\top\left(\mathbf{S}_{h+1,k-1}^\top \mathbf{S}_{h+1,k-1} + \lambda \mathbf{I}_d\right)^{-1}\phi_{h+1}(\cdot,a)}\}, H\Bigg\}.$$

By Lemma 4.8.1 and Lemma 4.7.1, we have $\|\widehat{\theta}_{h+1,k}\| \le 2H\sqrt{(1+\varepsilon_{h+1,m})dk/\lambda} \le 2H\sqrt{(\lambda + KL_\phi^2)dk/\lambda^2}$ and $\beta_h = c_\beta \cdot HL_dL_\lambda(1+\varepsilon_{h,m})\sqrt{m\iota} \le c_\beta \cdot HL_dL_\lambda(1+KL_\phi^2/\lambda)\sqrt{m\iota}$. Moreover, $\mathbf{S}_{h+1,k-1}^\top \mathbf{S}_{h+1,k-1}$ is positive semi-definite and has rank no greater than m. Furthermore, by Lemma 4.7.4 we have

$$\|\mathbf{S}_{h+1,k-1}^\top \mathbf{S}_{h+1,k-1}\|_F \le \|\mathbf{S}_{h+1,k-1}\|_F^2 = \text{tr}(\mathbf{S}_{h+1,k-1}^\top \mathbf{S}_{h+1,k-1})$$

$$\leq \mathrm{tr}(\mathbf{\Phi}_{h+1,k-1}^{\top}\mathbf{\Phi}_{h+1,k-1}) = \mathrm{tr}\left(\sum_{i=1}^{k-1}\phi_{h+1,i}\phi_{h+1,i}^{\top}\right) = \sum_{i=1}^{k-1}\|\phi_{h+1,i}\|^2 \leq kL_\phi^2. \quad (4.5)$$

Therefore, Lemma 4.10.6 implies that $V_{h+1,k}(\cdot)$ is within a class of function $\mathcal{V}$ whose ϵ-covering number $\mathcal{N}_\epsilon$ with respect to the distance $\mathrm{dist}(V, V') = \sup_x |V(x) - V'(x)|$ is upper bounded by

$$\log \mathcal{N}_\epsilon \leq d\log\left(\frac{12H\sqrt{(\lambda + KL_\phi^2)dk}}{\lambda\epsilon}\right) + (2d+1)m\log\left(\frac{36kL_\phi^4 c_\beta^2 H^2 L_d^2 L_\lambda^2(1 + KL_\phi^2/\lambda)^2 m\iota}{\lambda^2\epsilon^2}\right). \quad (4.6)$$

So, there exists $\widetilde{V}_{h+1,k} \in \mathcal{V}$ in the ϵ-covering such that

$$V_{h+1,k} = \widetilde{V}_{h+1,k} + \Delta_{h+1,k}, \quad \sup_x |\Delta_{h+1,k}(x)| \leq \epsilon.$$

Therefore,

$$\begin{aligned}
&\left\|\sum_{i=1}^{k-1}\phi_{hi}\left(V_{h+1,k}(s_{h+1,i}) - \mathbb{E}_{s'|s_{hi},a_{hi}}\left[V_{h+1,k}(s')\right]\right)\right\|_{\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}}^2 \\
&\leq 2\left\|\sum_{i=1}^{k-1}\phi_{hi}\left(\widetilde{V}_{h+1,k}(s_{h+1,i}) - \mathbb{E}_{s'|s_{hi},a_{hi}}\left[\widetilde{V}_{h+1,k}(s')\right]\right)\right\|_{\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}}^2 \\
&\quad + 2\left\|\sum_{i=1}^{k-1}\phi_{hi}\left(\Delta_{h+1,k}(s_{h+1,i}) - \mathbb{E}_{s'|s_{hi},a_{hi}}\left[\Delta_{h+1,k}(s')\right]\right)\right\|_{\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}}^2 \\
&\leq 2\left\|\sum_{i=1}^{k-1}\phi_{hi}\left(\widetilde{V}_{h+1,k}(s_{h+1,i}) - \mathbb{E}_{s'|s_{hi},a_{hi}}\left[\widetilde{V}_{h+1,k}(s')\right]\right)\right\|_{\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}}^2 \\
&\quad + 8\epsilon^2 k^2 \sup_{\|\phi\|\leq L_\phi}\|\phi\|_{\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}}^2 \\
&\leq 2(1+\varepsilon_{h,m})\left\|\sum_{i=1}^{k-1}\phi_{hi}\left(\widetilde{V}_{h+1,k}(s_{h+1,i}) - \mathbb{E}_{s'|s_{hi},a_{hi}}\left[\widetilde{V}_{h+1,k}(s')\right]\right)\right\|_{\mathbf{\Sigma}_{h,k-1}^{-1}}^2
\end{aligned}$$

$$
\begin{aligned}
&+ 8\epsilon^2 k^2 \cdot (1+\varepsilon_{h,m}) \sup_{\|\phi\| \le L_\phi} \|\phi\|^2_{\mathbf{\Sigma}^{-1}_{h,k-1}} \\
&\le 2(1+\varepsilon_{h,m}) \left\| \sum_{i=1}^{k-1} \phi_{hi} \left(\widetilde{V}_{h+1,k}(s_{h+1,i}) - \mathbb{E}_{s'|s_{hi},a_{hi}} \left[\widetilde{V}_{h+1,k}(s') \right] \right) \right\|^2_{\mathbf{\Sigma}^{-1}_{h,k-1}} \\
&\quad + \frac{8\epsilon^2 k^2 (1+\varepsilon_{h,m}) L_\phi^2}{\lambda}.
\end{aligned}
$$

Using Lemma 4.10.2 and Lemma 4.7.5 to control the first term and a union bound over the $\mathcal{N}_\epsilon$ possible $\widetilde{V}_{h+1,k}$ yields that with probability at least $1-\delta$, for all $k \in [K]$ we have

$$
\begin{aligned}
&\left\| \sum_{i=1}^{k-1} \phi_{hi} \left(V_{h+1,k}(s_{h+1,i}) - \mathbb{E}_{s'|s_{hi},a_{hi}} \left[V_{h+1,k}(s') \right] \right) \right\|^2_{\widetilde{\mathbf{\Sigma}}^{-1}_{h,k-1}} \\
&\le 4H^2(1+\varepsilon_{h,m}) \left[\frac{1}{2} \log \left(\frac{\det(\mathbf{\Sigma}_{h,k-1})}{\det(\lambda \mathbf{I}_d)} \right) + \log \frac{\mathcal{N}_\epsilon}{\delta} \right] + \frac{8k^2\epsilon^2(1+\varepsilon_{h,m})^2 L_\phi^2}{\lambda} \\
&\le 4H^2(1+\varepsilon_{h,m}) \Bigg[\frac{d}{2} \log(1+\varepsilon_{h,m}) + \frac{m}{2} \log \left(1 + \frac{(k-1)L_\phi^2}{m\lambda} \right) \\
&\quad + d \log \left(\frac{12H\sqrt{(\lambda + KL_\phi^2)dk}}{\lambda\epsilon} \right) \\
&\quad + (2d+1)m \log \left(\frac{36kL_\phi^4 c_\beta^2 H^2 L_d^2 L_\lambda^2 (1+KL_\phi^2/\lambda)^2 m\iota}{\lambda^2\epsilon^2} \right) + \log \frac{1}{\delta} \Bigg] \\
&\quad + \frac{8k^2\epsilon^2(1+\varepsilon_{h,m})L_\phi^2}{\lambda}.
\end{aligned}
$$

Taking another union bound over $h \in [H]$, we have that with probability at least $1-\delta$, $\forall (h,k) \in [H] \times [K]$,

$$
\begin{aligned}
&\left\| \sum_{i=1}^{k-1} \phi_{hi} \left(V_{h+1,k}(s_{h+1,i}) - \mathbb{E}_{s'|s_{hi},a_{hi}} \left[V_{h+1,k}(s') \right] \right) \right\|^2_{\widetilde{\mathbf{\Sigma}}^{-1}_{h,k-1}} \\
&\qquad \le 4H^2(1+\varepsilon_{h,m}) \left[\frac{d}{2} \log(1+\varepsilon_{h,m}) + \frac{m}{2} \log \left(1 + \frac{(k-1)L_\phi^2}{m\lambda} \right) \right.
\end{aligned}
$$

$$+ d \log \left(\frac{12H\sqrt{(\lambda + KL_\phi^2)dk}}{\lambda\epsilon} \right)$$
$$+(2d+1)m \log \left(\frac{36kL_\phi^4 c_\beta^2 H^2 L_d^2 L_\lambda^2 (1 + KL_\phi^2/\lambda)^2 m\iota}{\lambda^2\epsilon^2} \right) + \log \frac{H}{\delta} \Bigg] + \frac{8k^2\epsilon^2(1+\varepsilon_{h,m})L_\phi^2}{\lambda}.$$

Picking $\epsilon = \frac{\sqrt{md\lambda}H}{KL\phi}$, we see that there exists a absolute constant $c_0 > 0$ independent of c_β such that

$$\left\| \sum_{i=1}^{k-1} \phi_{hi} \left(V_{h+1,k}(s_{h+1,i}) - \mathbb{E}_{s'|s_{hi},a_{hi}} \left[V_{h+1,k}(s') \right] \right) \right\|^2_{\widetilde{\mathbf{\Sigma}}^{-1}_{h,k-1}}$$
$$\leq c_0 \cdot (1 + \varepsilon_{h,m}) mdH^2 \log \left(\frac{2(c_\beta + 1)L_d L_\lambda \max\{1, L_\phi\} dT}{\sqrt{\lambda}\delta} \right).$$

□

Equipped with the concentration lemma Lemma 4.8.2, we can recursively bound the difference between the estimated and true action-value functions of any policy by the expected difference at next step and an additional error term that is upper bounded by our exploration bonus.

Lemma 4.8.3. *Under the setting of Theorem 4.5.1, there exists an absolute constant $C > 0$ such that for $\beta_h = c_\beta \cdot HL_d L_\lambda (1 + \varepsilon_{h,m})\sqrt{m\iota}$ where $L_d = \max\left\{\sqrt{d}, L_\psi, \frac{L_r}{H}\right\}, L_\lambda = \max\{1, \sqrt{\lambda}\}, \iota = \log\left(\frac{2L_d L_\lambda \max\{1, L_\phi\} dT}{\sqrt{\lambda}\delta}\right), c_\beta > C$, and for any fixed policy π, on the event $\mathcal{G}$ defined in Lemma 4.8.2, we have for all $(s, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$,*

$$\left| \phi_h(s,a)^\top \widehat{\theta}_{hk} - Q_h^\pi(s,a) - \mathbb{E}_{s'|s,a} \left[V_{h+1,k}(s') - V_{h+1}^\pi(s') \right] \right|$$
$$\leq \beta_h \sqrt{\phi_h(s,a)^\top \widetilde{\mathbf{\Sigma}}^{-1}_{h,k-1} \phi_h(s,a)}. \quad (4.7)$$

Proof. By Lemma 4.10.1 and the Bellman equation,

$$Q_h^\pi(s,a) = \phi_h(s,a)^\top \theta_h^\pi = r_h(s,a) + \mathbb{E}_{s'|s,a} \left[V_{h+1}^\pi(s') \right].$$

As a result,

$$
\begin{aligned}
\widehat{\theta}_{hk} - \theta_h^\pi =& \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1} \left(\sum_{i=1}^{k-1} \phi_{hi} \left[r_{hi} + V_{h+1,k}(s_{h+1,i}) \right] \right) - \theta_h^\pi \\
=& \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1} \left(\sum_{i=1}^{k-1} \phi_{hi} \left[r_{hi} + V_{h+1,k}(s_{h+1,i}) \right] - \left(\sum_{i=1}^{k-1} \phi_{hi}\phi_{hi}^\top + (\lambda - \overline{\rho}_{h,k-1}) \mathbf{I}_d \right) \theta_h^\pi \right) \\
=& -(\lambda - \overline{\rho}_{h,k-1}) \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1} \theta_h^\pi + \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1} \left(\sum_{i=1}^{k-1} \phi_{hi} \left[r_{hi} + V_{h+1,k}(s_{h+1,i}) - \phi_{hi}^\top \theta_h^\pi \right] \right) \\
=& -(\lambda - \overline{\rho}_{h,k-1}) \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1} \theta_h^\pi \\
& + \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1} \left(\sum_{i=1}^{k-1} \phi_{hi} \left(V_{h+1,k}(s_{h+1,i}) - \mathbb{E}_{s'|s_{hi},a_{hi}} \left[V_{h+1}^\pi(s') \right] \right) \right) \\
=& -(\lambda - \overline{\rho}_{h,k-1}) \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1} \theta_h^\pi \\
& + \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1} \left(\sum_{i=1}^{k-1} \phi_{hi} \left(V_{h+1,k}(s_{h+1,i}) - \mathbb{E}_{s'|s_{hi},a_{hi}} \left[V_{h+1,k}(s') \right] \right) \right) \\
& + \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1} \left(\sum_{i=1}^{k-1} \phi_{hi} \mathbb{E}_{s'|s_{hi},a_{hi}} \left[V_{h+1,k}(s') - V_{h+1}^\pi(s') \right] \right).
\end{aligned}
$$

Here $\overline{\rho}_{h,k-1}$ is the sum of the smallest eigenvalue of the FD-sketched correlation matrix $\mathbf{S}_{h,i}^\top \mathbf{S}_{h,i}$ for $i = 1, \ldots, k-1$, as defined in Section 4.7. Now we bound the three terms' inner product with $\phi_h(s,a)$ separately. For the first term,

$$
\begin{aligned}
& \left| \phi_h(s,a)^\top (\lambda - \overline{\rho}_{h,k-1}) \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1} \theta_h^\pi \right| \\
& \leq (\lambda + \overline{\rho}_{h,k-1}) \sqrt{(\theta_h^\pi)^\top \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1} \theta_h^\pi} \cdot \sqrt{\phi_h(s,a)^\top \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1} \phi_h(s,a)} \\
& \leq (1 + \varepsilon_{h,m}) \sqrt{\lambda} \|\theta_h^\pi\| \cdot \sqrt{\phi_h(s,a)^\top \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1} \phi_h(s,a)} \\
& \leq (1 + \varepsilon_{h,m}) \sqrt{\lambda} \cdot (L_r + H L_\psi) \cdot \sqrt{\phi_h(s,a)^\top \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1} \phi_h(s,a)}.
\end{aligned}
$$

The second inequality uses the fact that the eigenvalues of $\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}$ is no larger than $1/\lambda$.

The third inequality uses Lemma 4.10.1. For the second term,

$$\begin{aligned}
&\left|\phi_h(s,a)^\top \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\left(\sum_{i=1}^{k-1}\phi_{hi}\left(V_{h+1,k}(s_{h+1,i})-\mathbb{E}_{s'|s_{hi},a_{hi}}\left[V_{h+1,k}(s')\right]\right)\right)\right| \\
&\leq \left\|\sum_{i=1}^{k-1}\phi_{hi}\left(V_{h+1,k}(s_{h+1,i})-\mathbb{E}_{s'|s_{hi},a_{hi}}\left[V_{h+1,k}(s')\right]\right)\right\|_{\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}} \cdot \sqrt{\phi_h(s,a)^\top \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\phi_h(s,a)} \\
&\leq c_0 \cdot H\sqrt{(1+\varepsilon_{h,m})md\chi}\cdot\sqrt{\phi_h(s,a)^\top \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\phi_h(s,a)}.
\end{aligned}$$

The last step is condition on the event $\mathcal{G}$ defined in Lemma 4.8.2 where c_0 is an absolute constant that is independent of c_β and $\chi = \log\left(\frac{2(c_\beta+1)L_d L_\lambda \max\{1,L_\phi\}dT}{\sqrt{\lambda}\delta}\right)$. For the third term,

$$\begin{aligned}
&\phi_h(s,a)^\top \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\left(\sum_{i=1}^{k-1}\phi_{hi}\mathbb{E}_{s'|s_{hi},a_{hi}}\left[V_{h+1,k}(s')-V_{h+1}^{\pi}(s')\right]\right) \\
&= \phi_h(s,a)^\top \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\left(\sum_{i=1}^{k-1}\phi_{hi}\phi_{hi}^\top\int_{s'}\psi_h(s')\left[V_{h+1,k}(s')-V_{h+1}^{\pi}(s')\right]\right) \\
&= \phi_h(s,a)^\top\int_{s'}\psi_h(s')\left[V_{h+1,k}(s')-V_{h+1}^{\pi}(s')\right] \\
&\quad -(\lambda-\overline{\rho}_{h,k-1})\phi_h(s,a)^\top \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\int_{s'}\psi_h(s')\left[V_{h+1,k}(s')-V_{h+1}^{\pi}(s')\right] \\
&= \mathbb{E}_{s'|s,a}\left[V_{h+1,k}(s')-V_{h+1}^{\pi}(s')\right] \\
&\quad -(\lambda-\overline{\rho}_{h,k-1})\phi_h(s,a)^\top \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\int_{s'}\psi_h(s')\left[V_{h+1,k}(s')-V_{h+1}^{\pi}(s')\right].
\end{aligned}$$

The first part appears in the left hand side of (4.7), while the second part can be bounded by

$$\begin{aligned}
&\left|(\lambda-\overline{\rho}_{h,k-1})\phi_h(s,a)^\top \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\int_{s'}\psi_h(s')\left[V_{h+1,k}(s')-V_{h+1}^{\pi}(s')\right]\right| \\
&\leq 2(\lambda+\overline{\rho}_{h,k-1})H\phi_h(s,a)^\top \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\int_{s'}\psi_h(s') \\
&\leq 2(1+\varepsilon_{h,m})\lambda H\sqrt{\left(\int_{s'}\psi_h(s')\right)^\top \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\int_{s'}\psi_h(s')}\cdot\sqrt{\phi_h(s,a)^\top \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\phi_h(s,a)}
\end{aligned}$$

$$\leq 2(1+\varepsilon_{h,m})H\sqrt{\lambda}\int_{s'}\|\psi_h(s')\|\cdot\sqrt{\phi_h(s,a)^\top\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\phi_h(s,a)}$$
$$\leq 2(1+\varepsilon_{h,m})HL_\psi\sqrt{\lambda}\cdot\sqrt{\phi_h(s,a)^\top\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\phi_h(s,a)}.$$

Putting things together, we see that

$$\left|\phi_h(s,a)^\top\widehat{\theta}_{hk}-Q_h^\pi(s,a)-\mathbb{E}_{s'|s,a}\left[V_{h+1,k}(s')-V_{h+1}^\pi(s')\right]\right|$$
$$\leq c'\cdot(1+\varepsilon_{h,m})H\sqrt{m\chi}\cdot\max\{\sqrt{\lambda},1\}\cdot\max\left\{\sqrt{d},L_\psi,\frac{L_r}{H}\right\}\cdot\sqrt{\phi_h(s,a)^\top\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\phi_h(s,a)}.$$

for some absolute constant c' that is independent of c_β. Finally, we need to prove the existence of C so that for any $c_\beta > C$, we have

$$c'\sqrt{\chi}=c'\sqrt{\log(c_\beta+1)+\iota}\leq c_\beta\sqrt{\iota}$$

where $\iota=\log\left(\frac{2L_dL_\lambda\max\{1,L_\phi\}dT}{\sqrt{\lambda}\delta}\right)\geq\log 2$. Since c' is independent of c_β, evidently as long as c_β is large enough, $c'\sqrt{\log(c_\beta+1)+x}\leq c_\beta\sqrt{x}$ holds for all $x\geq\log 2$. This concludes the proof.

□

An immediate consequence of Lemma 4.8.3 is that the estimated action-value fucntion is an over-estimation compared to the true action-value function with high probability.

Lemma 4.8.4. *Under the settting of Theorem 4.5.1, on the event $\mathcal{G}$ defined in Lemma 4.8.2, we have $Q_{hk}(s,a)\geq Q_h^\star(s,a)$ for all $(s,a,h,k)\in\mathcal{S}\times\mathcal{A}\times[H+1]\times[K]$.*

Proof. The lemma is proved by induction. In the base case where $h=H+1$, the lemma trivially holds as $Q_{H+1,k}(s,a)=Q_{H+1}^\star(s,a)=0$. Now for a generic h, By Lemma 4.8.3

$$\left|\phi_h(s,a)^\top\widehat{\theta}_{hk}-Q_h^\star(s,a)-\mathbb{E}_{s'|s,a}\left[V_{h+1,k}(s')-V_{h+1}^\star(s')\right]\right|$$
$$\leq\beta_h\sqrt{\phi_h(s,a)^\top\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\phi_h(s,a)}.$$

Denote $\pi_{h+1,k}(s') \stackrel{\text{def}}{=} \arg\max_a Q_{h+1,k}(s',a)$ and $\pi^\star_{h+1}(s') \stackrel{\text{def}}{=} \arg\max_a Q^\star_{h+1}(s',a)$. Using the inductive assumption at $h+1$, $\forall s' \in \mathcal{S}$,

$$\begin{aligned}\left[V_{h+1,k}(s') - V^\star_{h+1}(s')\right] &= Q_{h+1,k}(s', \pi_{h+1,k}(s')) - Q^\star_{h+1}(s', \pi^\star_{h+1}(s')) \\ &\geq Q_{h+1,k}(s', \pi^\star_{h+1}(s')) - Q^\star_{h+1}(s', \pi^\star_{h+1}(s')) \geq 0.\end{aligned}$$

Therefore,

$$Q^\star_h(s,a) \leq \min\left\{\phi_h(s,a)^\top \widehat{\theta}_{hk} + \beta_h \sqrt{\phi_h(s,a)^\top \widetilde{\boldsymbol{\Sigma}}^{-1}_{h,k-1}\phi_h(s,a)}, H\right\} = Q_{hk}(s,a).$$

This concludes the proof. □

Now we are ready to prove the main theorem on regret of S-LSVI-UCB, which is restated as follows.

Theorem 4.5.1. *Under Assumption 4.2.1, there exists an absolute constant $C > 0$ such that, for any $\delta \in (0,1)$, if we set $\beta_h \stackrel{\text{def}}{=} c_\beta \cdot HL_dL_\lambda(1+\varepsilon_{h,m})\sqrt{m\iota}$ for any $c_\beta > C$ in Algorithm 4 with $L_d \stackrel{\text{def}}{=} \max\left\{\sqrt{d}, L_\psi, \frac{L_r}{H}\right\}$, $L_\lambda \stackrel{\text{def}}{=} \max\{1, \sqrt{\lambda}\}$, $\iota \stackrel{\text{def}}{=} \log\left(\left(2L_dL_\lambda \max\{1, L_\phi\}dT\right)/(\sqrt{\lambda}\delta)\right)$, then with probability at least $1-\delta$, the total regret of S-LSVI-UCB is at most $\widetilde{\mathcal{O}}\left(\sum_{h=1}^H L_dL_\lambda\sqrt{(1+\varepsilon_{h,m})^3 m\widetilde{m}_h \cdot HT}\right)$. If we further assume that $\lambda = 1, L_\phi = 1$ and $L_\psi = L_r = \sqrt{d}$, and denote $\varepsilon_m = \max_h\{\varepsilon_{h,m}\}$, $\widetilde{m} = \max_h\{\widetilde{m}_h\}$, then the regret bound can be simplified as $\widetilde{\mathcal{O}}\left(\sqrt{(1+\varepsilon_m)^3 md\widetilde{m} \cdot H^3T}\right)$.*

Proof. Define $\delta_{hk} \stackrel{\text{def}}{=} V_{hk}(s_{hk}) - V^{\pi_k}_h(s_{hk})$ where π_k is the policy according to S-LSVI-UCB before episode k. and $\zeta_{h,k} \stackrel{\text{def}}{=} \mathbb{E}\left[\delta_{h+1,k}|s_{hk}, a_{hk}\right] - \delta_{h+1,k}$. We condition on the event $\mathcal{G}$ define in Lemma 4.8.2 with probability $1-\delta/2$. Lemma 4.8.3 and Lemma 4.8.4 gives

$$\text{Regret}(K) = \sum_{k=1}^K \left[V^\star_1(s_{1k}) - V^{\pi_k}_1(s_{1k})\right]$$

$$
\begin{aligned}
&\leq \sum_{k=1}^{K} \left[V_{1k}(s_{1k}) - V_1^{\pi_k}(s_{1k})\right] \\
&= \sum_{k=1}^{K} \delta_{1k} \\
&\leq \sum_{k=1}^{K} \left[\delta_{2k} + \zeta_{1k} + \min\left\{H, \beta_1 \sqrt{\phi_{1k}^\top \widetilde{\boldsymbol{\Sigma}}_{1,k-1}^{-1} \phi_{1k}}\right\}\right] \\
&= \sum_{k=1}^{K} \delta_{2k} + \sum_{k=1}^{K} \left[\zeta_{1k} + \min\left\{H, \beta_1 \sqrt{\phi_{1k}^\top \widetilde{\boldsymbol{\Sigma}}_{1,k-1}^{-1} \phi_{1k}}\right\}\right] \\
&\leq \dots \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H} \zeta_{hk} + \sum_{k=1}^{K}\sum_{h=1}^{H} \min\left\{H, \beta_h \sqrt{\phi_{hk}^\top \widetilde{\boldsymbol{\Sigma}}_{h,k-1}^{-1} \phi_{hk}}\right\}. \qquad (4.8)
\end{aligned}
$$

For the first term, observe that $\{\zeta_{hk}\}$ is a martingale difference sequence that is bounded by $2H$. Therefore, Azuma-Hoeffding inequality yields that with probability at least $1 - \delta/2$,

$$
\sum_{k=1}^{K}\sum_{h=1}^{H} \zeta_{hk} \leq 2H\sqrt{T\log(2/\delta)} \leq 2H\sqrt{T\iota}. \qquad (4.9)
$$

For the second term, with our choice of $\beta_h = c_\beta \cdot HL_dL_\lambda\sqrt{(1+\varepsilon_{h,m})\iota}$, we write

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H} \min\left\{H, \beta_h \sqrt{\phi_{hk}^\top \widetilde{\boldsymbol{\Sigma}}_{h,k-1}^{-1} \phi_{hk}}\right\} \\
&= \sum_{k=1}^{K}\sum_{h=1}^{H} \min\left\{H, c_\beta HL_dL_\lambda(1+\varepsilon_{h,m})\sqrt{m\iota \cdot \phi_{hk}^\top \widetilde{\boldsymbol{\Sigma}}_{h,k-1}^{-1} \phi_{hk}}\right\} \\
&= H\sum_{h=1}^{H}\sum_{k=1}^{K} \min\left\{1, c_\beta L_dL_\lambda(1+\varepsilon_{h,m})\sqrt{m\iota \cdot \phi_{hk}^\top \widetilde{\boldsymbol{\Sigma}}_{h,k-1}^{-1} \phi_{hk}}\right\} \\
&\leq H\sum_{h=1}^{H}\left(\max\left\{1, c_\beta L_dL_\lambda(1+\varepsilon_{h,m})\sqrt{m\iota}\right\} \cdot \sum_{k=1}^{K} \min\left\{1, \|\phi_{hk}\|_{\widetilde{\boldsymbol{\Sigma}}_{h,k-1}^{-1}}\right\}\right).
\end{aligned}
$$

Next, we apply Cauchy-Schwarz and Lemma 4.7.6 to continue

$$
\begin{aligned}
&H\sum_{h=1}^{H}\left(\max\left\{1, c_\beta L_d L_\lambda(1+\varepsilon_{h,m})\sqrt{m\iota}\right\}\cdot\sum_{k=1}^{K}\min\left\{1, \|\phi_{hk}\|_{\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}}\right\}\right)\\
&\le H\sum_{h=1}^{H}\left(\max\left\{1, c_\beta L_d L_\lambda(1+\varepsilon_{h,m})\sqrt{m\iota}\right\}\cdot\sqrt{K}\cdot\left[\sum_{k=1}^{K}\min\left\{1, \|\phi_{hk}\|^2_{\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}}\right\}\right]^{1/2}\right)\\
&\le 2H\sum_{h=1}^{H}\left(\max\left\{1, c_\beta L_d L_\lambda(1+\varepsilon_{h,m})\sqrt{m\iota}\right\}\cdot\sqrt{(1+\varepsilon_{h,m})K}\right.\\
&\quad\left.\cdot\left[d\log\left(1+\varepsilon_{h,m}\right)+m\log\left(1+\frac{KL_\phi^2}{m\lambda}\right)\right]^{1/2}\right)\\
&\le 2H\sum_{h=1}^{H}\sqrt{(1+\varepsilon_{h,m})K\widetilde{m}_h}\max\left\{1, c_\beta L_d L_\lambda(1+\varepsilon_{h,m})\sqrt{m\iota}\right\}. \qquad (4.10)
\end{aligned}
$$

Combining (4.8), (4.9), and (4.10), we conclude that with probability at least $1-\delta$,

$$
\text{Regret}(K) \overset{\widetilde{\mathcal{O}}}{=} \sum_{h=1}^{H} L_d L_\lambda\sqrt{(1+\varepsilon_{h,m})^3 m\cdot\widetilde{m}_h HT}.
$$

□

4.9 Proof of Theorem 4.5.2

The proof in this section is based on Zanette et al. [2019]. A notable deviation is Lemma 4.9.8, where the dependence on dimension is improved by a factor of $\sqrt{d/m}$ due to the reduced covering number from low rank property of sketched correlation matrix.

4.9.1 Definitions

In this section, we formally define the filtrations, the parameters used in the algorithm, and the good events.

Definition 4.9.1. *For any $(h,k) \in [H] \times [K]$, define the filtrations*

$$\begin{aligned}
\mathcal{H}_{hk} &\stackrel{\text{def}}{=} \{s_{ij}, a_{ij}, r_{ij} : j \le k, i \le h \text{ if } j = k \text{ else } i \le H\} \\
\mathcal{H}_k &\stackrel{\text{def}}{=} \mathcal{H}_{H,k} \\
\overline{\mathcal{H}}_{hk} &\stackrel{\text{def}}{=} \mathcal{H}_k \cup \{\overline{\xi}_{ik}, i \ge h\} \\
\overline{\mathcal{H}}_k &\stackrel{\text{def}}{=} \overline{\mathcal{H}}_{1k}
\end{aligned}$$

Definition 4.9.2. *For any $0 < \delta < 1$, any $h \in [H]$, and some absolute constants $c_\beta > 0$, let*

$$\begin{aligned}
L_d &\stackrel{\text{def}}{=} \max\{\sqrt{md}, L_r, L_\psi/H\} \\
L_\lambda &\stackrel{\text{def}}{=} \max\{1, \sqrt{\lambda}\} \\
\iota &\stackrel{\text{def}}{=} \log\left(\frac{2L_d L_\lambda \max\{L_\phi, 1\} dT}{\sqrt{\lambda}\delta}\right) \\
\sqrt{\beta_h(\delta)} &\stackrel{\text{def}}{=} c_\beta \cdot H L_d L_\lambda (1 + \varepsilon_{h,m})\sqrt{\iota} \\
\sqrt{\nu_h(\delta)} &\stackrel{\text{def}}{=} 2\sqrt{\beta_h(\delta)} \\
\sqrt{\gamma_h(\delta)} &\stackrel{\text{def}}{=} \sqrt{2dH\nu_h(\delta)\log(4dT/\delta)}
\end{aligned}$$

Definition 4.9.3. *Set*

$$\begin{aligned}
\sigma^2 &\stackrel{\text{def}}{=} H\nu_h(\delta) \\
\alpha_U &\stackrel{\text{def}}{=} \frac{1}{4\sqrt{\gamma_h(\delta)}} \le \frac{1}{2\left(\sqrt{\nu_h(\delta)} + \sqrt{\gamma_h(\delta)}\right)} \\
\alpha_L &\stackrel{\text{def}}{=} \alpha_U/2
\end{aligned}$$

Definition 4.9.4. *For any $(h,k) \in [H] \times [K]$, let $\overline{\eta}_{h,k}$ be defined in (4.11) and $\overline{\lambda}^{\pi}_{hk}$ be*

defined in (4.12), define the following set of good events.

$$\mathcal{G}_{hk}^{\overline{\xi}} \stackrel{\text{def}}{=} \left\{ \|\overline{\xi}_{hk}\|_{\widetilde{\mathbf{\Sigma}}_{h,k-1}} \leq \sqrt{\gamma_h(\delta)} \right\}$$

$$\mathcal{G}_{hk}^{\overline{\eta}} \stackrel{\text{def}}{=} \left\{ \forall (s,a) \in \mathcal{S} \times \mathcal{A},\ |\phi_h(s,a)^\top \overline{\eta}_{hk}| \leq \sqrt{\beta_h(\delta)} \|\phi_h(s,a)\|_{\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}} \right\}$$

$$\mathcal{G}_{hk}^{\overline{\lambda}} \stackrel{\text{def}}{=} \Bigg\{ \forall \text{ policy } \pi \text{ and } \forall (s,a) \in \mathcal{S} \times \mathcal{A},\ |\phi_h(s,a)^\top \overline{\lambda}_{hk}^{\pi}| \leq (1+\varepsilon_{h,m})(2HL_\psi + L_r)\sqrt{\lambda} \|\phi_h(s,a)\|_{\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}} \Bigg\}$$

$$\mathcal{G}_{hk}^{\overline{Q}} \stackrel{\text{def}}{=} \left\{ \forall (s,a) \in \mathcal{S} \times \mathcal{A},\ |\overline{Q}_{hk}(s,a) - Q_h^\star(s,a)| \leq H - h + 1 \right\}$$

$$\overline{\mathcal{G}}_{hk} \stackrel{\text{def}}{=} \{\mathcal{G}_{hk}^{\overline{\xi}} \cap \mathcal{G}_{hk}^{\overline{\eta}} \cap \mathcal{G}_{hk}^{\overline{\lambda}} \cap \mathcal{G}_{hk}^{\overline{Q}}\}$$

$$\overline{\mathcal{G}}_{k} \stackrel{\text{def}}{=} \cap_{h \in [H]} \overline{\mathcal{G}}_{hk}$$

4.9.2 Concentration

The purpose of this section is to show that good events happen with high probability. The high level idea is as follows. Lemma 4.9.5 decomposes the difference of the unclipped estimated Q-value defined by $\overline{\theta}_{hk}$ and the true Q-value Q^π of any policy π into four parts. One term is the recursive term, while the three other terms are bounded separately in Lemma 4.9.6, 4.9.7, and 4.9.8, either independently or conditioned on bounded $\overline{Q}_{t+1}$. Then, in Lemma 4.9.10, we show that $\overline{Q}_t$ is bounded as long as $\overline{Q}_{t+1}$ is bounded and the three other terms are bounded. Finally, in Lemma 4.9.11, we inductively show that good events happen with high probability.

By Lemma 4.10.1, there exists θ_h^π such that $Q_h^\pi(s,a) = \phi_h(s,a)^\top \theta_h^\pi\ \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$. Define $\overline{\eta}_{h,k}$ and $\overline{\lambda}_{hk}^\pi$ as follows.

$$\overline{\eta}_{h,k} \stackrel{\text{def}}{=} \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1} \left(\sum_{i=1}^{k-1} \phi_{hi} \left(\overline{V}_{h+1,k}(s_{h+1,i}) - \mathbb{E}_{s' | s_{hi}, a_{hi}} \left[\overline{V}_{h+1,k}(s') \right] \right) \right) \tag{4.11}$$

$$\overline{\lambda}_{hk}^\pi \stackrel{\text{def}}{=} -(\lambda - \overline{\rho}_{h,k-1}) \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1} \left(\int_{s'} \psi_h(s') \left[\overline{V}_{h+1,k}(s') - V_{h+1}^\pi(s') \right] + \theta_h^\pi \right) \tag{4.12}$$

Recall that $\overline{\rho}_{h,k-1}$ is the sum of the minimum eigenvalue of the FD-sketched correlation matrix $\mathbf{S}_{h,i}^\top \mathbf{S}_{h,i}$ for $i = 1, \ldots, k-1$, as defined in Section 4.7.

Lemma 4.9.5 (Decomposition of Unclipped Q-value)**.** *For any* $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ *and any policy* π*,*

$$\phi_h(s,a)^\top \overline{\theta}_{hk} - Q_h^\pi(s,a) = \mathbb{E}_{s'|s,a}\left[\overline{V}_{h+1,k}(s') - V_{h+1}^\pi(s')\right] + \phi_h(s,a)^\top \left(\overline{\eta}_{hk} + \overline{\xi}_{hk} + \overline{\lambda}_{hk}^\pi\right).$$

Proof. By definition of $\overline{\theta}_{hk}$ and $Q_h^\pi(s,a)$,

$$\phi_h(s,a)^\top \overline{\theta}_{hk} - Q_h^\pi(s,a) = \phi_h(s,a)^\top \left(\widehat{\theta}_{hk} + \overline{\xi}_{hk} - \theta_h^\pi\right).$$

We rewrite $\widehat{\theta}_{hk} - \theta_h^\pi$ as follows.

$$\begin{aligned}
\widehat{\theta}_{hk} - \theta_h^\pi =& \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1} \left(\sum_{i=1}^{k-1} \phi_{hi}\left[r_{hi} + \overline{V}_{h+1,k}(s_{h+1,i})\right]\right) - \theta_h^\pi \\
=& \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1} \left(\sum_{i=1}^{k-1} \phi_{hi}\left[r_{hi} + \overline{V}_{h+1,k}(s_{h+1,i})\right] \right. \\
& \left. - \left(\sum_{i=1}^{k-1} \phi_{hi}\phi_{hi}^\top + (\lambda - \overline{\rho}_{h,k-1})\mathbf{I}_d\right)\theta_h^\pi\right) \\
=& -(\lambda - \overline{\rho}_{h,k-1})\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\theta_h^\pi + \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\left(\sum_{i=1}^{k-1} \phi_{hi}\left[r_{hi} + \overline{V}_{h+1,k}(s_{h+1,i}) - \phi_{hi}^\top\theta_h^\pi\right]\right) \\
=& -(\lambda - \overline{\rho}_{h,k-1})\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\theta_h^\pi \\
& + \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\left(\sum_{i=1}^{k-1} \phi_{hi}\left(\overline{V}_{h+1,k}(s_{h+1,i}) - \mathbb{E}_{s'|s_{hi},a_{hi}}\left[V_{h+1}^\pi(s')\right]\right)\right) \\
=& -(\lambda - \overline{\rho}_{h,k-1})\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\theta_h^\pi \\
& + \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\left(\sum_{i=1}^{k-1} \phi_{hi}\left(\overline{V}_{h+1,k}(s_{h+1,i}) - \mathbb{E}_{s'|s_{hi},a_{hi}}\left[\overline{V}_{h+1,k}(s')\right]\right)\right) \\
& + \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\left(\sum_{i=1}^{k-1} \phi_{hi}\mathbb{E}_{s'|s_{hi},a_{hi}}\left[\overline{V}_{h+1,k}(s') - V_{h+1}^\pi(s')\right]\right)
\end{aligned}$$

$$
\begin{aligned}
=&\overline{\eta}_{hk} - (\lambda - \overline{\rho}_{h,k-1})\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\theta_h^{\pi} \\
&+ \widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\left(\sum_{i=1}^{k-1}\phi_{hi}\mathbb{E}_{s'|s_{hi},a_{hi}}\left[\overline{V}_{h+1,k}(s') - V_{h+1}^{\pi}(s')\right]\right).
\end{aligned}
$$

Next, we expand the inner product of the third term with $\phi_h(s,a)$ as follows

$$
\begin{aligned}
&\phi_h(s,a)^{\top}\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\left(\sum_{i=1}^{k-1}\phi_{hi}\mathbb{E}_{s'|s_{hi},a_{hi}}\left[\overline{V}_{h+1,k}(s') - V_{h+1}^{\pi}(s')\right]\right) \\
&= \phi_h(s,a)^{\top}\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\left(\sum_{i=1}^{k-1}\phi_{hi}\phi_{hi}^{\top}\int_{s'}\psi_h(s')\left[\overline{V}_{h+1,k}(s') - V_{h+1}^{\pi}(s')\right]\right) \\
&= \phi_h(s,a)^{\top}\int_{s'}\psi_h(s')\left[\overline{V}_{h+1,k}(s') - V_{h+1}^{\pi}(s')\right] \\
&\quad - (\lambda - \overline{\rho}_{h,k-1})\phi_h(s,a)^{\top}\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\int_{s'}\psi_h(s')\left[\overline{V}_{h+1,k}(s') - V_{h+1}^{\pi}(s')\right] \\
&= \mathbb{E}_{s'|s,a}\left[\overline{V}_{h+1,k}(s') - V_{h+1}^{\pi}(s')\right] \\
&\quad - (\lambda - \overline{\rho}_{h,k-1})\phi_h(s,a)^{\top}\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\int_{s'}\psi_h(s')\left[\overline{V}_{h+1,k}(s') - V_{h+1}^{\pi}(s')\right].
\end{aligned}
$$

Putting things together, we get the claimed result. □

Lemma 4.9.6 (Regularization)**.** *For any policy π and any $(s,a,h,k) \in \mathcal{S}\times\mathcal{A}\times[H]\times[K]$, if $|\overline{Q}_{h+1,k}(s,a) - Q_{h+1}^{\pi}(s,a)| \le H - h$, then*

$$
\left|\phi_h(s,a)^{\top}\overline{\lambda}_{hk}^{\pi}\right| \le (1+\varepsilon_{h,m})(2HL_{\psi} + L_r)\sqrt{\lambda}\|\phi_h(s,a)\|_{\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}}.
$$

Proof.

$$
\begin{aligned}
&\left|\phi_h(s,a)^{\top}\overline{\lambda}_{hk}^{\pi}\right| \\
&= \left|\phi_h(s,a)^{\top}(\lambda - \overline{\rho}_{h,k-1})\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\left(\int_{s'}\psi_h(s')\left[\overline{V}_{h+1,k}(s') - V_{h+1}^{\pi}(s')\right] + \theta_h^{\pi}\right)\right| \\
&\le (\lambda + \overline{\rho}_{h,k-1})\left\|\phi_h(s,a)^{\top}\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\right\|\left(\left\|\int_{s'}\psi_h(s')\left[\overline{V}_{h+1,k}(s') - V_{h+1}^{\pi}(s')\right]\right\| + \|\theta_h^{\pi}\|\right)
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(a)}{\leq} (1+\varepsilon_{h,m})\sqrt{\lambda}\,\|\phi_h(s,a)\|_{\widetilde{\mathbf{\Sigma}}^{-1}_{h,k-1}} \left(\left\| \int_{s'} \psi_h(s') \left[\overline{V}_{h+1,k}(s') - V^{\pi}_{h+1}(s') \right] \right\| + \|\theta^{\pi}_h\| \right) \\
&\leq (1+\varepsilon_{h,m})\sqrt{\lambda}\,\|\phi_h(s,a)\|_{\widetilde{\mathbf{\Sigma}}^{-1}_{h,k-1}} \left(\int_{s'} \left\| \psi_h(s') \left[\overline{V}_{h+1,k}(s') - V^{\pi}_{h+1}(s') \right] \right\| + \|\theta^{\pi}_h\| \right) \\
&\leq (1+\varepsilon_{h,m})\sqrt{\lambda}\,\|\phi_h(s,a)\|_{\widetilde{\mathbf{\Sigma}}^{-1}_{h,k-1}} \left(\int_{s'} \|\psi_h(s')\| \left| \overline{V}_{h+1,k}(s') - V^{\pi}_{h+1}(s') \right| + \|\theta^{\pi}_h\| \right) \\
&\leq (1+\varepsilon_{h,m})\sqrt{\lambda}\,\|\phi_h(s,a)\|_{\widetilde{\mathbf{\Sigma}}^{-1}_{h,k-1}} \left((H-t) \int_{s'} \|\psi_h(s')\| + \|\theta^{\pi}_h\| \right) \\
&\overset{(b)}{\leq} (1+\varepsilon_{h,m})\sqrt{\lambda}\,\|\phi_h(s,a)\|_{\widetilde{\mathbf{\Sigma}}^{-1}_{h,k-1}} \left((H-t)L_{\psi} + (L_r + HL_{\psi}) \right) \\
&\leq (1+\varepsilon_{h,m})(2HL_{\psi} + L_r)\sqrt{\lambda}\,\|\phi_h(s,a)\|_{\widetilde{\mathbf{\Sigma}}^{-1}_{h,k-1}}.
\end{aligned}
$$

In the derivation step (a) uses the fact that the maximum eigenvalue of $\widetilde{\mathbf{\Sigma}}^{-1}_{h,k-1}$ is no more than $1/\lambda$, while step (b) uses Lemma 4.10.1. □

Lemma 4.9.7 (Gaussian Randomness)**.** *Fix $(h,k) \in [H] \times [K]$. For any $\delta \in (0,1)$, with probability at least $1 - \delta/2T$, the event $\mathcal{G}^{\overline{\xi}}_{hk}$ happens, i.e.,*

$$
\|\overline{\xi}_{hk}\|_{\widetilde{\mathbf{\Sigma}}_{h,k-1}} \leq \sqrt{\gamma_h(\delta)}.
$$

Proof. Given that $\overline{\xi}_{hk} \sim N(0, H\nu_h(\delta)\widetilde{\mathbf{\Sigma}}^{-1}_{h,k-1})$, Lemma 4.10.8 implies that with probability at least $1 - \delta/2T$,

$$
\|\overline{\xi}_{hk}\|_{\widetilde{\mathbf{\Sigma}}_{h,k-1}} \leq \sqrt{2H\nu_h(\delta) d \log(4dT/\delta)} \overset{\text{def}}{=} \sqrt{\gamma_h(\delta)}.
$$

□

Lemma 4.9.8 (Concentration Induction)**.** *Fix $(h,k) \in [H] \times [K]$. There exists absolute constant C such that for any $c_\beta > C$ and $\delta \in (0,1)$, if we set the parameter as in Section*

4.9.1, then condition on $\mathcal{G}^{\overline{Q}}_{h+1,k}$ and $\mathcal{G}^{\overline{\xi}}_{h+1,k}$, with probability at least $1-\delta/2T$,

$$\left|\phi_h(s,a)^\top \overline{\eta}_{hk}\right| \le \sqrt{\beta_h(\delta)}\,\|\phi_h(s,a)\|_{\widetilde{\mathbf{\Sigma}}^{-1}_{h,k-1}}.$$

Proof. By definition of $\overline{\eta}_{hk}$ in (4.11) and Cauchy-Schwarz,

$$\begin{aligned}&\left|\phi_h(s,a)^\top \overline{\eta}_{hk}\right| \\ &\le \|\phi_h(s,a)\|_{\widetilde{\mathbf{\Sigma}}^{-1}_{h,k-1}} \cdot \left\|\sum_{i=1}^{k-1}\phi_{hi}\left(\overline{V}_{h+1,k}(s_{h+1,i}) - \mathbb{E}_{s'|s_{hi},a_{hi}}\left[\overline{V}_{h+1,k}(s')\right]\right)\right\|_{\widetilde{\mathbf{\Sigma}}^{-1}_{h,k-1}}. \end{aligned} \tag{4.13}$$

By conditioning on $\mathcal{G}^{\overline{Q}}_{h+1,k}$, we can follow the same reasoning as in the proof of Lemma 4.8.1 and, combined with Lemma 4.7.1, deduce that

$$\|\widehat{\theta}_{h+1,k}\| \le 2H\sqrt{\frac{(1+\varepsilon_{h,m})kd}{\lambda}} \le 2H\sqrt{\frac{(\lambda+KL_\phi^2)kd}{\lambda^2}}.$$

Combined with the event $\mathcal{G}^{\overline{\xi}}_{h+1,k}$ we condition on, we have

$$\begin{aligned}\|\overline{\theta}_{h+1,k}\| = \|\widehat{\theta}_{h+1,k} + \overline{\xi}_{h+1,k}\| &\le \|\widehat{\theta}_{h+1,k}\| + \|\overline{\xi}_{h+1,k}\| \\ &\le 2H\sqrt{\frac{(\lambda+KL_\phi^2)kd}{\lambda^2}} + \sqrt{\frac{\gamma_h(\delta)}{\lambda}}. \end{aligned} \tag{4.14}$$

To get rid of the randomness in $\gamma_h(\delta)$ from dependence on $\varepsilon_{h,m}$, we define

$$\begin{aligned}\sqrt{\beta_{\max}(\delta)} &\overset{\text{def}}{=} c_\beta \cdot HL_dL_\lambda(1+kL_\phi^2/\lambda)\sqrt{\iota} && \ge \sqrt{\beta_h(\delta)} \\ \sqrt{\nu_{\max}(\delta)} &\overset{\text{def}}{=} 2\sqrt{\beta_{\max}(\delta)} && \ge \sqrt{\nu_h(\delta)} \\ \sqrt{\gamma_{\max}(\delta)} &\overset{\text{def}}{=} \sqrt{2dH\nu(\delta)\log(4dT/\delta)} && \ge \sqrt{\gamma_h(\delta)}\end{aligned}$$

Now we proceed with a covering argument. For any $h \in [H]$, we define

$$Q_h^{\theta,\mathbf{\Sigma}}(s,a) \stackrel{\text{def}}{=} \begin{cases} \phi_h(s,a)^\top \theta, & \text{if } \|\phi_h(s,a)\|_{\mathbf{\Sigma}} \le \alpha_L \\ H-h+1, & \text{if } \|\phi_h(s,a)\|_{\mathbf{\Sigma}} \ge \alpha_U \\ \left(\frac{\alpha_U - \|\phi_h(s,a)\|_{\mathbf{\Sigma}}}{\alpha_U - \alpha_L}\right)\phi_h(s,a)^\top\theta + & \\ \left(1 - \frac{\alpha_U - \|\phi_h(s,a)\|_{\mathbf{\Sigma}}}{\alpha_U - \alpha_L}\right)(H-h+1) & \text{otherise} \end{cases}. \tag{4.15}$$

Let $V_h^{\theta,\mathbf{\Sigma}}(s) \stackrel{\text{def}}{=} \max_{a\in\mathcal{A}} Q_h^{\theta,\mathbf{\Sigma}}(s,a)$. Note that $\overline{V}_{h+1,k} = V_{h+1}^{\overline{\theta}_{h+1,k},\widetilde{\mathbf{\Sigma}}^{-1}_{h+1,k-1}}$. For any $h \in [H]$, we then define

$$O_{h+1} \stackrel{\text{def}}{=} \Bigg\{ \theta \in \mathbb{R}^d, \mathbf{\Sigma} \in \mathbb{R}^{d\times d} : \|\theta\| \le 2H\sqrt{\frac{(\lambda + KL_\phi^2)kd}{\lambda^2}} + \sqrt{\frac{\gamma_{\max}(\delta)}{\lambda}}, \\ \mathbf{\Sigma} = (\mathbf{H} + \lambda \mathbf{I}_d)^{-1}, \mathbf{H} \succeq 0, \text{rank}(\mathbf{H}) \le m, \|\mathbf{H}\|_F \le k, \\ \left|Q_{h+1}^{\theta,\mathbf{\Sigma}}(s,a) - Q_{h+1}^\star(s,a)\right| \le H-h \; \forall s,a \Bigg\}.$$

By the event $\mathcal{G}_{h+1,k}^{\overline{Q}}$ we condition on, (4.14) and (4.5) we have $(\overline{\theta}_{h+1,k}, \widetilde{\mathbf{\Sigma}}^{-1}_{h+1,k-1}) \in O_{h+1}$. Next, define for any $(\theta, \mathbf{\Sigma}) \in O_{h+1}$ and $i \in [k-1]$,

$$x_{hi}^{\theta,\mathbf{\Sigma}} \stackrel{\text{def}}{=} V_{h+1}^{\theta,\mathbf{\Sigma}}(s_{h+1,i}) - \mathbb{E}_{s'|s_{hi},a_{hi}}\left[V_{h+1}^{\theta,\mathbf{\Sigma}}(s')\right].$$

Clearly $\{x_{hi}^{\theta,\mathbf{\Sigma}}, \mathcal{H}_{hi}\}$ is a martingale difference sequence bounded by $2H$. By Lemma 4.7.7, Lemma 4.10.2, and Lemma 4.7.5, with probability at least $1-\delta/2T$,

$$\left\|\sum_{i=1}^k \phi_{hi} x_{hi}^{\theta,\mathbf{\Sigma}}\right\|^2_{\widetilde{\mathbf{\Sigma}}^{-1}_{h,k-1}} \\ \le (1+\varepsilon_{h,m})\left\|\sum_{i=1}^k \phi_{hi} x_{hi}^{\theta,\mathbf{\Sigma}}\right\|^2_{\mathbf{\Sigma}^{-1}_{h,k-1}}$$

$$\leq 8(1+\varepsilon_{h,m})H^2 \log\left(\frac{2T\det(\mathbf{\Sigma}_{h,k-1})^{1/2}\det(\lambda \mathbf{I}_d)^{-1/2}}{\delta}\right)$$
$$\leq 8(1+\varepsilon_{h,m})H^2\left[\frac{d}{2}\log\left(1+\varepsilon_{h,m}\right)+\frac{m}{2}\log\left(1+\frac{(k-1)L_\phi^2}{m\lambda}\right)+\log\left(\frac{2T}{\delta}\right)\right].$$

By Lemma 4.10.7, the ϵ-covering number of O_{h+1}, denoted as $\mathcal{N}_\epsilon$, is upper bounded by

$$\log \mathcal{N}_\epsilon \leq d\log\left(\frac{6H\sqrt{(\lambda+KL_\phi^2)kd}+3\sqrt{\lambda\gamma_{\max}(\delta)}}{\epsilon\lambda}\right)+(2d+1)m\cdot\log\left(9kd/\lambda^2\epsilon^2\right).$$

Taking a union bound over the at most $\mathcal{N}_\epsilon$ different elements in the ϵ-covering, we know that there exists $(\theta,\mathbf{\Sigma})\in O_{h+1}$ such that $\|\theta-\overline{\theta}_{h+1,k}\|\leq\epsilon$ and $\|\mathbf{\Sigma}-\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}\|\leq\epsilon$, and with probability at least $1-\delta/2T$,

$$\begin{aligned}
&\left\|\sum_{i=1}^{k-1}\phi_{hi}\left(\overline{V}_{h+1,k}(s_{h+1,i})-\mathbb{E}_{s'|s_{hi},a_{hi}}\left[\overline{V}_{h+1,k}(s')\right]\right)\right\|^2_{\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}}\\
&\leq 2\left\|\sum_{i=1}^{k}\phi_{hi}x_{hi}^{\theta,\mathbf{\Sigma}}\right\|^2_{\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}}+2\left\|\sum_{i=1}^{k}\phi_{hi}\left(x_{hi}^{\theta,\mathbf{\Sigma}}-x_{hi}^{\overline{\theta}_{h+1,k},\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}}\right)\right\|^2_{\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}}\\
&\leq 16(1+\varepsilon_{h,m})H^2\left[\frac{d}{2}\log\left(1+\varepsilon_{h,m}\right)+\frac{m}{2}\log\left(1+\frac{kL_\phi^2}{m\lambda}\right)+\log\left(\frac{2T\mathcal{N}_\epsilon}{\delta}\right)\right]\\
&\quad+2\left\|\sum_{i=1}^{k}\phi_{hi}\left(x_{hi}^{\theta,\mathbf{\Sigma}}-x_{hi}^{\overline{\theta}_{h+1,k},\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}}\right)\right\|^2_{\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}}\\
&\leq 16(1+\varepsilon_{h,m})H^2\left[\frac{d}{2}\log\left(1+\frac{KL_\phi^2}{m\lambda}\right)+\frac{m}{2}\log\left(1+\frac{kL_\phi^2}{m\lambda}\right)\right.\\
&\quad+d\log\left(\frac{6H\sqrt{(\lambda+KL_\phi^2)kd}+3\sqrt{\lambda\gamma_{\max}(\delta)}}{\epsilon\lambda}\right)+(2d+1)m\cdot\log\left(9kd/\lambda^2\epsilon^2\right)\\
&\quad\left.+\log\left(\frac{2T}{\delta}\right)\right]+2\left\|\sum_{i=1}^{k}\phi_{hi}\left(x_{hi}^{\theta,\mathbf{\Sigma}}-x_{hi}^{\overline{\theta}_{h+1,k},\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}}\right)\right\|^2_{\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}}. \qquad (4.16)
\end{aligned}$$

Using the fact that $\|\theta - \overline{\theta}_{h+1,k}\| \leq \epsilon$ and $\|\Sigma - \widetilde{\Sigma}_{h,k-1}^{-1}\| \leq \epsilon$, the last quantity can be bounded with Lemma 4.9.9 as follows.

$$\begin{aligned}
&\left\|\sum_{i=1}^{k} \phi_{hi}\left(x_{hi}^{\theta,\Sigma} - x_{hi}^{\overline{\theta}_{h+1,k},\widetilde{\Sigma}_{h,k-1}^{-1}}\right)\right\|_{\widetilde{\Sigma}_{h,k-1}^{-1}} \\
&\leq \frac{kL_\phi}{\sqrt{\lambda}} \sup_i \left|x_{hi}^{\theta,\Sigma} - x_{hi}^{\overline{\theta}_{h+1,k},\widetilde{\Sigma}_{h,k-1}^{-1}}\right| \\
&= \frac{kL_\phi}{\sqrt{\lambda}} \sup_i \left|V_{hi}^{\theta,\Sigma}(s_{h+1,i}) - \overline{V}_{h+1,i}(s_{h+1,i}) - \mathbb{E}_{s'|s_{hi},a_{hi}}\left[V_{hi}^{\theta,\Sigma}(s') - \overline{V}_{h+1,i}(s')\right]\right| \\
&\leq \frac{2kL_\phi}{\sqrt{\lambda}} \sup_{s,a} \left|Q_{h+1,k}^{\theta,\Sigma}(s,a) - \overline{Q}_{h+1,k}(s,a)\right| \\
&\leq \frac{8H^2kL_\phi^2\sqrt{\epsilon}}{(\alpha_U - \alpha_L)\sqrt{\lambda}}.
\end{aligned} \tag{4.17}$$

Combining (4.16) and (4.17), with probability at least $1 - \delta/2T$,

$$\begin{aligned}
&\left\|\sum_{i=1}^{k-1} \phi_{hi}\left(\overline{V}_{h+1,k}(s_{h+1,i}) - \mathbb{E}_{s'|s_{hi},a_{hi}}\left[\overline{V}_{h+1,k}(s')\right]\right)\right\|_{\widetilde{\Sigma}_{h,k-1}^{-1}}^2 \\
&\qquad \leq 16(1+\varepsilon_{h,m})H^2\left[\frac{d}{2}\log\left(1 + \frac{KL_\phi^2}{m\lambda}\right) + \frac{m}{2}\log\left(1 + \frac{kL_\phi^2}{m\lambda}\right)\right. \\
&+ d\log\left(\frac{6H\sqrt{(\lambda + KL_\phi^2)kd} + 3\sqrt{\lambda\gamma_{\max}(\delta)}}{\epsilon\lambda}\right) + (2d+1)m \cdot \log\left(9kd/\lambda^2\epsilon^2\right) \\
&\qquad \left. + \log\left(\frac{2T}{\delta}\right)\right] + \frac{128H^4k^2L_\phi^4\epsilon}{(\alpha_U - \alpha_L)^2\lambda}.
\end{aligned}$$

By picking $\epsilon = \min\left\{\frac{(\alpha_U - \alpha_L)^2\lambda}{128H^4k^2L_\phi^4}, 1, \frac{H}{3}, \alpha_U - \alpha_L\right\}$ and recalling the definition of parameters in Section 4.9.1, we can conclude that there exists absolute constant C such that for any $c_\beta > C$,

$$\left\|\sum_{i=1}^{k-1} \phi_{hi}\left(\overline{V}_{h+1,k}(s_{h+1,i}) - \mathbb{E}_{s'|s_{hi},a_{hi}}\left[\overline{V}_{h+1,k}(s')\right]\right)\right\|_{\widetilde{\Sigma}_{h,k-1}^{-1}}$$

$$\leq c_\beta \cdot HL_d L_\lambda \sqrt{(1+\varepsilon_{h,m})\iota} \leq \sqrt{\beta_h(\delta)}.$$

Plugging into (4.13) concludes the proof. □

Lemma 4.9.9 ([Zanette et al., 2019, Lemma E.4]). *Using the same notation as in Lemma 4.9.8 but suppressing indices. Suppose we have $(\theta_1, \mathbf{\Sigma}_1) \in O$ and $(\theta_2, \mathbf{\Sigma}_2) \in O$ such that $\|\theta_1 - \theta_2\| \leq \epsilon$ and $\|\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2\| \leq \epsilon$ with $\epsilon \leq \min\{1, \frac{H}{3}, \alpha_U - \alpha_L\}$, then*

$$\sup_{s,a} \left| Q^{\theta,\mathbf{\Sigma}}(s,a) - Q^{\theta',\mathbf{\Sigma}'}(s,a) \right| \leq \frac{4H^2 L_\phi \sqrt{\epsilon}}{\alpha_U - \alpha_L}.$$

Lemma 4.9.10 (Boundedness Induction). *Condition on event $\mathcal{G}^{\overline{Q}}_{h+1,k}$ and assume that*

$$\left| \phi_h(s,a)^\top \left(\overline{\eta}_{hk} + \overline{\xi}_{hk} + \overline{\lambda}^\star_{hk} \right) \right| \leq \left(\sqrt{\nu_h(\delta)} + \sqrt{\gamma_h(\delta)} \right) \|\phi_h(s,a)\|_{\widetilde{\mathbf{\Sigma}}^{-1}_{h,k-1}}$$

where $\overline{\lambda}^\star_{hk}$ is defined in (4.12) with $\pi = \pi^\star$. Then the event $\mathcal{G}^{\overline{Q}}_{h,k}$ holds.

Proof. We break into cases depending on whether the feature is large.

Case 1 (large feature): $\|\phi_h(s,a)\|_{\widetilde{\mathbf{\Sigma}}^{-1}_{h,k-1}} \geq \alpha_U$. By definition of our Q function in (4.3.3), $0 \leq \overline{Q}_{hk}(s,a) \leq H - h + 1$. Given that $0 \leq Q^\star_h(s,a) \leq H - h + 1$,

$$\left| \overline{Q}_{hk}(s,a) - Q^\star_h(s,a) \right| \leq H - h + 1 \; \forall (s,a) \in \mathcal{S} \times \mathcal{A}.$$

Case 2 (small feature): $\|\phi_h(s,a)\|_{\widetilde{\mathbf{\Sigma}}^{-1}_{h,k-1}} \leq \alpha_L$. By definition of our Q function in (4.3.3), $\overline{Q}_{hk}(s,a) = \phi_h(s,a)^\top \overline{\theta}_{hk}$. Applying Lemma 4.9.5 and condition on the event $\mathcal{G}^{\overline{Q}}_{h+1,k}$, $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned} \left| \overline{Q}_{hk}(s,a) - Q^\star_h(s,a) \right| &= \mathbb{E}_{s'|s,a} \left[\overline{V}_{h+1,k}(s') - V^\star_{h+1}(s') \right] + \phi_h(s,a)^\top \left(\overline{\eta}_{hk} + \overline{\xi}_{hk} + \overline{\lambda}^\star_{hk} \right) \\ &\leq H - t + \left(\sqrt{\nu_h(\delta)} + \sqrt{\gamma_h(\delta)} \right) \|\phi_h(s,a)\|_{\widetilde{\mathbf{\Sigma}}^{-1}_{h,k-1}} \end{aligned}$$

$$\leq H - t + \left(\sqrt{\nu_h(\delta)} + \sqrt{\gamma_h(\delta)}\right)\alpha_L$$
$$\leq H - t + 1.$$

Case 3 (medium feature): $\alpha_L < \|\phi_h(s,a)\|_{\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}} < \alpha_U$**.** By definition of our Q function in (4.3.3), $\overline{Q}_{hk}(s,a) = qQ_1 + (1-q)Q_2$ for some $0 < q < 1$ and Q_1, Q_2 such that

$$|Q_i(s,a) - Q_h^\star(s,a)| \leq H - h + 1 \; \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \; i = 1, 2.$$

Triangle inequality implies that $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\left|\overline{Q}_{hk}(s,a) - Q_h^\star(s,a)\right| \leq q\left|Q_1(s,a) - Q_h^\star(s,a)\right| + (1-q)\left|Q_2(s,a) - Q_h^\star(s,a)\right|$$
$$\leq H - h + 1.$$

□

Lemma 4.9.11 (Good Event Probability)**.** *There exists absolute constant C such that for any $c_\beta > C$ and $K > 0, \delta \in (0,1)$, if we set the parameters as in Section 4.9.1, then with probability at least $1-\delta$ we have $\cap_{k \leq K} \overline{\mathcal{G}}_k$.*

Proof. Let $\delta' = \delta/K$ and $\delta'' = \delta'/2H = \delta/2T$, by union bound it suffices to show that for each $k \in [K]$, $\overline{\mathcal{G}}_k$ happens with probability at least $1-\delta'$. We consider backward induction over step $h \in [H]$.

As the base case, $\overline{Q}_{H+1,k} = Q_{H+1}^\star = 0$ so we get $\mathcal{G}_{H+1,k}^{\overline{Q}}$, we can invoke Lemma 4.9.6 and get $\mathcal{G}_{Hk}^{\overline{\lambda}}$. At the same time, we invoke Lemma 4.9.7 and get $\mathcal{G}_{Hk}^{\overline{\xi}}$ with probability at least $1-\delta''$. Moreover, by inspection we see that $\overline{\eta}_{Hk} = 0$ so we get $\mathcal{G}_{Hk}^{\overline{\eta}}$. Now, conditioning on $\mathcal{G}_{H+1,k}^{\overline{Q}} \cap \mathcal{G}_{Hk}^{\overline{\lambda}} \cap \mathcal{G}_{Hk}^{\overline{\xi}} \cap \mathcal{G}_{Hk}^{\overline{\eta}}$, we invoke Lemma 4.9.10 and get $\mathcal{G}_{H,k}^{\overline{Q}}$ with probability at least $1-\delta''$. Next we repeat the process starting with $\mathcal{G}_{H,k}^{\overline{Q}}$. Invoke Lemma 4.9.6 and Lemma 4.9.7 to get $\mathcal{G}_{H-1,k}^{\overline{\lambda}}$ and $\mathcal{G}_{H-1,k}^{\overline{\xi}}$ with probability at least $1-\delta''$. Condition on $\mathcal{G}_{H,k}^{\overline{Q}} \cap \mathcal{G}_{Hk}^{\overline{\xi}}$ we invoke Lemma 4.9.8 and get $\mathcal{G}_{H-1,k}^{\overline{\eta}}$ with probability at least $1-\delta''$.

Then, condition on $\mathcal{G}_{H,k}^{\overline{Q}} \cap \mathcal{G}_{H-1,k}^{\overline{\lambda}} \cap \mathcal{G}_{H-1,k}^{\overline{\xi}} \cap \mathcal{G}_{H-1,k}^{\overline{\eta}}$, we invoke Lemma 4.9.10 and get $\mathcal{G}_{H-1,k}^{\overline{Q}}$. To sum up, we get $\mathbb{P}(\overline{\mathcal{G}}_{h-1,k}|\overline{\mathcal{G}}_{h,k}) \geq (1-\delta'')^2 \ \forall h \in [H]$. As a result, $\mathbb{P}(\overline{\mathcal{G}}_k) \geq (1-\delta'')^{2H} \geq 1-\delta$. □

4.9.3 Regret Bounds

In this section, we show that the regret of S-RLSVI is bounded with high probability. We begin with two lemmas that will prove useful for later proofs. The first lemma shows that the estimated value functions are optimistic with at least constant probability, while the second lemma bounds a summation term that will appear when bounding regrets.

Lemma 4.9.12 (Optimism)**.** *For any episode k, if $0 < \delta < \frac{\Phi(-1)}{2}$,*

$$\mathbb{P}(\overline{V}_{1k}(s_{1k}) - V_1^\star(s_{1k}) \geq 0 | \mathcal{H}_k) \geq \frac{\Phi(-1)}{2}.$$

Proof. The proof is essentially the same as the proof of [Zanette et al., 2019, Lemma F.2] so is omitted. □

Lemma 4.9.13 (Warmup Bound)**.**

$$\sum_{k=1}^{K}\sum_{h=1}^{H} H \mathbb{1}\{\|\phi_{hk}\|_{\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}} > \alpha_L\} \leq \sum_{h=1}^{H} \frac{2H}{\alpha_L^2}(1+\varepsilon_{h,m})\widetilde{m}_h.$$

Proof.

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H} H \mathbb{1}\{\|\phi_{hk}\|_{\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}} > \alpha_L\} \\
&= H \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{1}\left\{\frac{\|\phi_{hk}\|_{\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}}}{\alpha_L} > 1\right\} \\
&= H \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{1}\left\{\frac{\|\phi_{hk}\|^2_{\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}}}{\alpha_L^2} > 1\right\}
\end{aligned}$$

$$
\begin{aligned}
&\leq H \sum_{k=1}^{K} \sum_{h=1}^{H} \min \left\{ 1, \frac{\|\phi_{hk}\|^2_{\widetilde{\mathbf{\Sigma}}^{-1}_{h,k-1}}}{\alpha_L^2} \right\} \\
&\overset{(a)}{\leq} \sum_{k=1}^{K} \sum_{h=1}^{H} \frac{H}{\alpha_L^2} \min\{1, \|\phi_{hk}\|^2_{\widetilde{\mathbf{\Sigma}}^{-1}_{h,k-1}}\} \\
&\overset{(b)}{\leq} \sum_{h=1}^{H} \frac{2H}{\alpha_L^2}(1+\varepsilon_{h,m}) \left[d \log\left(1+\varepsilon_{h,m}\right) + m \log\left(1 + \frac{KL_\phi^2}{m\lambda}\right) \right] \\
&= \sum_{h=1}^{H} \frac{2H}{\alpha_L^2}(1+\varepsilon_{h,m}) \widetilde{m}_h.
\end{aligned}
$$

In the derivation, step (a) uses the fact that $\alpha_L^2 \leq 1$ and step (b) uses Lemma 4.7.6. □

The following key lemma provides an upper bound on the sum of estimation error $\overline{V}_{1k}(s_{1k}) - V_1^{\pi_k}(s_{1k})$, where π_k is the policy according to S-RLSVI before episode k.

Lemma 4.9.14 (Bound on Estimation)**.** *For any $0 < \delta < 1$, it holds with probability at least $1 - \delta/2$ that*

$$
\sum_{k=1}^{K} \left(\overline{V}_{1k}(s_{1k}) - V_1^{\pi_k}(s_{1k}) \right) = \widetilde{\mathcal{O}}\Bigg(\sum_{h=1}^{H} \left(\sqrt{\nu_h(\delta)} + \sqrt{\gamma_h(\delta)} \right) \sqrt{(1+\varepsilon_{h,m}) K \widetilde{m}_h} + \sum_{h=1}^{H} \frac{H}{\alpha_L^2}(1+\varepsilon_{h,m}) \widetilde{m}_h \Bigg).
$$

Proof. The proof proceeds by induction over step $h \in [H]$. Let $G_k = \cap_{l \leq k} \overline{\mathcal{G}}_l$ measurable with respect to $\overline{\mathcal{H}}_k$. Consider a generic time step h, we will have either $\|\phi_{hk}\|_{\widetilde{\mathbf{\Sigma}}^{-1}_{h,k-1}} \leq \alpha_L$, denoted as $\mathcal{S}_{hk}$, or $\|\phi_{hk}\|_{\widetilde{\mathbf{\Sigma}}^{-1}_{h,k-1}} > \alpha_L$, denoted as $\mathcal{S}_{hk}^c$. Under $\mathcal{S}_{hk}$, the estimated value function is linear. Under $\mathcal{S}_{hk}^c$, we can still have an upper bound H on the difference in estimated and true value function by conditioning on G_k. Therefore,

$$
\begin{aligned}
&\mathbb{1}\{G_k\} \left(\overline{V}_{hk}(s_{hk}) - V_h^{\pi_k}(s_{hk}) \right) \\
&= \mathbb{1}\{G_k\} \left(\left[\overline{V}_{hk}(s_{hk}) - V_h^{\pi_k}(s_{hk}) \right] \mathbb{1}\{\mathcal{S}_{hk}\} + \left[\overline{V}_{hk}(s_{hk}) - V_h^{\pi_k}(s_{hk}) \right] \mathbb{1}\{\mathcal{S}_{hk}^c\} \right) \\
&= \mathbb{1}\{G_k\} \left(\left[\phi_{hk}^\top \overline{\theta}_{hk} - Q_h^{\pi_k}(s_{hk}, a_{hk}) \right] \mathbb{1}\{\mathcal{S}_{hk}\} + \left[\overline{V}_{hk}(s_{hk}) - V_h^{\pi_k}(s_{hk}^c) \right] \mathbb{1}\{\mathcal{S}_{hk}^c\} \right)
\end{aligned}
$$

$$\leq \mathbb{1}\{G_k\}\left(\left[\phi_{hk}^\top \overline{\theta}_{hk} - Q_h^{\pi_k}(s_{hk}, a_{hk})\right]\mathbb{1}\{\mathcal{S}_{hk}\} + H\mathbb{1}\{\mathcal{S}_{hk}^c\}\right).$$

The first term we bound by applying Lemma 4.9.5 and condition on $\overline{\mathcal{G}}_k$.

$$\begin{aligned}
&\phi_{hk}^\top \overline{\theta}_{hk} - Q_h^{\pi_k}(s_{hk}, a_{hk}) \\
&\leq \mathbb{E}_{s'|s_{hk},a_{hk}}\left[\overline{V}_{h+1,k}(s') - V_{h+1}^{\pi_k}(s')\right] + \phi_{hk}^\top\left(\overline{\eta}_{hk} + \overline{\xi}_{hk} + \overline{\lambda}_{hk}^{\pi_k}\right) \\
&\leq \mathbb{E}_{s'|s_{hk},a_{hk}}\left[\overline{V}_{h+1,k}(s') - V_{h+1}^{\pi_k}(s')\right] + \left(\sqrt{\nu_h(\delta)} + \sqrt{\gamma_h(\delta)}\right)\|\phi_{hk}\|_{\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}}. \qquad (4.18)
\end{aligned}$$

Now define

$$\begin{aligned}
\dot{\zeta}_{hk} \stackrel{\text{def}}{=} \mathbb{1}\{G_k\}\mathbb{1}\{\mathcal{S}_{hk}\}\Bigg(&\mathbb{E}_{s'|s_{hk},a_{hk}}\left[\overline{V}_{h+1,k}(s') - V_{h+1}^{\pi_k}(s')\right] \\
&- \left[\overline{V}_{h+1,k}(s_{h+1,k}) - V_{h+1}^{\pi_k}(s_{h+1,k})\right]\Bigg) \qquad (4.19)
\end{aligned}$$

Observe that $\{\dot{\zeta}_{hk}\}$ is a martingale difference sequence bounded by $2H$. Applying Azuma-Hoeffding we have with probability at least $1-\delta/4$, $\sum_{k=1}^K\sum_{h=1}^H\{\dot{\zeta}_{hk}\} = \widetilde{\mathcal{O}}(H\sqrt{T})$. Now, combining (4.18) and (4.19), we have

$$\begin{aligned}
\mathbb{1}\{G_k\}\left(\overline{V}_{hk}(s_{hk}) - V_h^{\pi_k}(s_{hk})\right) \leq \mathbb{1}\{G_k\}\Bigg(&\left[\overline{V}_{h+1,k}(s_{h+1,k}) - V_{h+1}^{\pi_k}(s_{h+1,k})\right]\mathbb{1}\{\mathcal{S}_{hk}\} \\
&+ \left(\sqrt{\nu_h(\delta)} + \sqrt{\gamma_h(\delta)}\right)\|\phi_{hk}\|_{\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}}\mathbb{1}\{\mathcal{S}_{hk}\} + H\mathbb{1}\{\mathcal{S}_{hk}^c\}\Bigg) + \dot{\zeta}_{hk}.
\end{aligned}$$

Induction gives

$$\begin{aligned}
&\mathbb{1}\{G_k\}\left(\overline{V}_{hk}(s_{hk}) - V_h^{\pi_k}(s_{hk})\right) \\
&\leq \mathbb{1}\{G_k\}\sum_{h=1}^H\left(\left(\sqrt{\nu_h(\delta)} + \sqrt{\gamma_h(\delta)}\right)\|\phi_{hk}\|_{\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}}\left(\prod_{h'=1}^h \mathbb{1}\{\mathcal{S}_{h'k}\}\right)\right.
\end{aligned}$$

$$
\begin{aligned}
&+ H\mathbb{1}\{\mathcal{S}_{hk}^c\}\left(\prod_{h'=1}^{h-1}\mathbb{1}\{\mathcal{S}_{h'k}\}\right)\Bigg) + \sum_{h=1}^{H}\dot{\zeta}_{hk} \\
&\leq \sum_{h=1}^{H}\left(\sqrt{\nu_h(\delta)}+\sqrt{\gamma_h(\delta)}\right)\min\{\alpha_L, \|\phi_{hk}\|_{\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}}\} + \sum_{h=1}^{H} H\mathbb{1}\{\mathcal{S}_{hk}^c\} + \sum_{h=1}^{H}\dot{\zeta}_{hk} \\
&\leq \sum_{h=1}^{H}\left(\sqrt{\nu_h(\delta)}+\sqrt{\gamma_h(\delta)}\right)\min\{1, \|\phi_{hk}\|_{\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}}\} + \sum_{h=1}^{H} H\mathbb{1}\{\mathcal{S}_{hk}^c\} + \sum_{h=1}^{H}\dot{\zeta}_{hk}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\sum_{k=1}^{K}\mathbb{1}\{G_k\}\left(\overline{V}_{hk}(s_{hk}) - V_h^{\pi_k}(s_{hk})\right) \\
&\leq \sum_{h=1}^{H}\left(\sqrt{\nu_h(\delta)}+\sqrt{\gamma_h(\delta)}\right)\sum_{k=1}^{K}\min\{1, \|\phi_{hk}\|_{\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}}\} + \sum_{k=1}^{K}\sum_{h=1}^{H} H\mathbb{1}\{\mathcal{S}_{hk}^c\} + \sum_{k=1}^{K}\sum_{h=1}^{H}\dot{\zeta}_{hk} \\
&\overset{(a)}{\leq} \sum_{h=1}^{H}\left(\sqrt{\nu_h(\delta)}+\sqrt{\gamma_h(\delta)}\right)\sqrt{K}\left(\sum_{k=1}^{K}\min\left\{1, \|\phi_{hk}\|^2_{\widetilde{\mathbf{\Sigma}}_{h,k-1}^{-1}}\right\}\right)^{1/2} \\
&\quad + \sum_{h=1}^{H}\frac{2H}{\alpha_L^2}(1+\varepsilon_{h,m})\widetilde{m}_h + \sum_{k=1}^{K}\sum_{h=1}^{H}\dot{\zeta}_{hk} \\
&\overset{(b)}{\leq} \sum_{h=1}^{H}\left(\sqrt{\nu_h(\delta)}+\sqrt{\gamma_h(\delta)}\right)\sqrt{2(1+\varepsilon_{h,m})K\widetilde{m}_h} + \sum_{h=1}^{H}\frac{2H}{\alpha_L^2}(1+\varepsilon_{h,m})\widetilde{m}_h + \sum_{k=1}^{K}\sum_{h=1}^{H}\dot{\zeta}_{hk}.
\end{aligned}
$$

In the derivation step (a) uses Lemma 4.9.13 and step (b) uses Lemma 4.7.6. The proof is finished by conditioning on $G_K = \cap_{k\leq K}\overline{\mathcal{G}}_k$ which happens with probability at least $1-\delta/4$ from Lemma 4.9.11 and the event that $\sum_{k=1}^{K}\sum_{h=1}^{H}\dot{\zeta}_{hk} = \widetilde{\mathcal{O}}(H\sqrt{T})$ with probability at least $1-\delta/4$. □

The next key lemma provides an upper bound on the sum of pessimism $\left(V_1^\star(s_{1k}) - \overline{V}_{1k}(s_{1k}))\right)$.

Lemma 4.9.15 (Bound on Pessimism)**.** *For any* $0 < \delta < \frac{\Phi(-1)}{2}$*, it holds with probability at least* $1-\delta/2$ *that*

$$\sum_{k=1}^{K}\left(V_1^\star(s_{1k})-\overline{V}_{1k}(s_{1k})\right)\Big)=\widetilde{\mathcal{O}}\Bigg(\sum_{h=1}^{H}\left(\sqrt{\nu_h(\delta)}+\sqrt{\gamma_h(\delta)}\right)\sqrt{(1+\varepsilon_{h,m})K\widetilde{m}_h}$$
$$+\sum_{h=1}^{H}\frac{H}{\alpha_L^2}(1+\varepsilon_{h,m})\widetilde{m}_h\Bigg).$$

Proof. We prove this Lemma by finding an upper bound for $V_1^\star(s_{1k})$ and a lower bound for $\overline{V}_{1k}(s_{1k})$. In this proof, we will have not only random variables $\overline{\xi}$ but also $\widetilde{\xi}$ and $\underline{\xi}$ which are i.i.d. copies of $\overline{\xi}$. Random variables $\widetilde{\xi}$ and $\underline{\xi}$ have associated good events $\widetilde{\mathcal{G}}_k$ and $\underline{\mathcal{G}}_k$ defined analogously to $\overline{\mathcal{G}}_k$. And accordingly we define $G_k=\cap_{l\le k}\left(\overline{\mathcal{G}}_l\cap\widetilde{\mathcal{G}}_l\cap\underline{\mathcal{G}}_l\right)$. Union bound over three set of random variables and Lemma 4.9.11 implies that G_k occurs with probability $1-\delta'$ for any $\delta'>0$.

The lower bound for $\overline{V}_{1k}(s_{1k})$ is constructed as follows. Let $\xi_{hk}\in\mathbb{R}^d\ \forall h\in[H]$ and let V_{hk}^{ξ} be the estimated value function obtained by running Algorithm 5 with non-random ξ_{hk} in place of $\overline{\xi}_{hk}$. Consider the minimization problem

$$\min_{\{\xi_{hk}\}_{h=1}^H} V_{1k}^{\xi}(s_{1k})$$
$$\|\xi_{hk}\|_{\widetilde{\mathbf{\Sigma}}_{h,k-1}}\le\sqrt{\gamma_h(\delta)},\forall h\in[H]. \tag{4.20}$$

The minimum exists because $V_{1k}^{\xi}(s_{1k})$ is a continuous function on ξ_{hk}'s which are defined on a compact set. Let $\underline{V}_{1k}(s_{1k})$ denote the minimum value and $\{\underline{\xi}_{hk}\}_{h=1}^H$ denote a minimizer of the optimization problem (4.20). Observe that under G_k , $\underline{V}_{1k}(s_{1k})\le\overline{V}_{1k}(s_{1k})$ because $\{\overline{\xi}_{hk}\}_{h=1}^H$ is a feasible solution of problem (4.20).

Next we construct an upper bound for $V_1^\star(s_{1k})$. Consider drawing i.i.d. copy $\widetilde{\xi}_{hk}$ of $\overline{\xi}_{hk}$'s and run Algorithm 5 again with $\widetilde{\xi}_{hk}$ in place of $\overline{\xi}_{hk}$ to get new estimated value functions $\widetilde{V}_{hk}$. Denote as $\widetilde{O}_k$ the event that $\widetilde{V}_{1k}(s_{1k})$ is optimistic. Applying Lemma 4.9.12,

$$\mathbb{P}(\widetilde{O}_k)=\mathbb{P}\left(\widetilde{V}_{1k}(s_{1k})-V_1^\star(s_{1k})\ge0|\mathcal{H}_k\right)\ge\frac{\Phi(-1)}{2}.$$

As a result

$$\begin{aligned}\left(V_1^\star(s_{1k}) - \overline{V}_{1k}(s_{1k})\right) \mathbb{1}\{G_k\} &\leq \mathbb{E}_{\widetilde{\xi}|\widetilde{O}_k}\left[\widetilde{V}_{1k}(s_{1k}) - \overline{V}_{1k}(s_{1k})\right] \mathbb{1}\{G_k\} \\ &\leq \mathbb{E}_{\widetilde{\xi}|\widetilde{O}_k}\left[\widetilde{V}_{1k}(s_{1k}) - \underline{V}_{1k}(s_{1k})\right] \mathbb{1}\{G_k\}.\end{aligned}$$

Now we use law of total expectation under $\widetilde{\mathcal{G}}_k$:

$$\begin{aligned}&\mathbb{E}_{\widetilde{\xi}}\left[\widetilde{V}_{1k}(s_{1k}) - \underline{V}_{1k}(s_{1k})\right] \\ &= \mathbb{E}_{\widetilde{\xi}|\widetilde{O}_k}\left[\widetilde{V}_{1k}(s_{1k}) - \underline{V}_{1k}(s_{1k})\right] \mathbb{P}(\widetilde{O}_k) + \mathbb{E}_{\widetilde{\xi}|\widetilde{O}_k^c}\left[\widetilde{V}_{1k}(s_{1k}) - \underline{V}_{1k}(s_{1k})\right] \mathbb{P}(\widetilde{O}_k^c) \\ &\geq \mathbb{E}_{\widetilde{\xi}|\widetilde{O}_k}\left[\widetilde{V}_{1k}(s_{1k}) - \underline{V}_{1k}(s_{1k})\right] \mathbb{P}(\widetilde{O}_k).\end{aligned}$$

The inequality is because $\{\widetilde{\xi}_{hk}\}_{h=1}^H$ is a feasible solution of problem (4.20). Hence,

$$\begin{aligned}&\left(V_{1k}^\star(s_{1k}) - \overline{V}_{1k}(s_{1k})\right) \mathbb{1}\{G_k\} \\ &\leq \frac{2}{\Phi(-1)} \mathbb{E}_{\widetilde{\xi}}\left[\widetilde{V}_{1k}(s_{1k}) - \underline{V}_{1k}(s_{1k})\right] \mathbb{1}\{G_k\} \\ &= \frac{2}{\Phi(-1)}\left[\overline{V}_{1k}(s_{1k}) - \underline{V}_{1k}(s_{1k})\right] \mathbb{1}\{G_k\} + \ddot{\zeta}_k \\ &= \frac{2}{\Phi(-1)}\left[\overline{V}_{1k}(s_{1k}) - V_1^{\pi_k}(s_{1k}) + V_1^{\pi_k}(s_{1k}) - \underline{V}_{1k}(s_{1k})\right] \mathbb{1}\{G_k\} + \ddot{\zeta}_k\end{aligned}$$

where $\ddot{\zeta}_k$ is defined as

$$\ddot{\zeta}_k \stackrel{\text{def}}{=} \frac{2}{\Phi(-1)}\left(\mathbb{E}_{\widetilde{\xi}}\left[\widetilde{V}_{1k}(s_{1k})\right] - \overline{V}_{1k}(s_{1k})\right) \mathbb{1}\{G_k\}.$$

Observe that $\widetilde{V}_{1k}$ and $\overline{V}_{1k}$ are i.i.d. because $\overline{\xi}_{hk}$ and $\widetilde{\xi}_{hk}$ are i.i.d.. Therefore, $\ddot{\zeta}_k$ is a martingale difference sequence bounded by $2H$ so with probability at least $1-\delta'$ we have $\sum_{k=1}^K \ddot{\zeta}_k = \widetilde{\mathcal{O}}(H\sqrt{K})$.

Next, we decompose

$$\frac{2}{\Phi(-1)}\left[\overline{V}_{1k}(s_{1k}) - V_1^{\pi_k}(s_{1k}) + V_1^{\pi_k}(s_{1k}) - \underline{V}_{1k}(s_{1k})\right]\mathbb{1}\{G_k\}$$
$$= \frac{2}{\Phi(-1)}\left[\overline{V}_{1k}(s_{1k}) - V_1^{\pi_k}(s_{1k})\right]\mathbb{1}\{G_k\} + \frac{2}{\Phi(-1)}\left[V_1^{\pi_k}(s_{1k}) - \underline{V}_{1k}(s_{1k})\right]\mathbb{1}\{G_k\}.$$

The first term can be bounded with Lemma 4.9.14, while the second term can be bounded following the same reasoning. Thus,

$$\sum_{k=1}^{K}\left(V_{1k}^{\star}(s_{1k}) - \overline{V}_{1k}(s_{1k})\right)\mathbb{1}\{G_k\}$$
$$\leq \frac{4}{\Phi(-1)}\cdot\widetilde{\mathcal{O}}\left(\sum_{h=1}^{H}\left(\sqrt{\nu_h(\delta)} + \sqrt{\gamma_h(\delta)}\right)\sqrt{(1+\varepsilon_{h,m})K\widetilde{m}_h}\right.$$
$$\left. + \sum_{h=1}^{H}\frac{H}{\alpha_L^2}(1+\varepsilon_{h,m})\widetilde{m}_h\right) + \widetilde{\mathcal{O}}(H\sqrt{K}).$$

To conlcude we pick $\delta' = \delta/6$ and take a union bound over two applications of Azuma-Hoeffding and G_K. □

Equipped with Lemma 4.9.14 and 4.9.15, we can prove the main theorem on the regret bound for S-RLSVI. We first restate the theorem.

Theorem 4.5.2. *Under Assumption 4.2.1, if we set* $\sigma = \widetilde{\mathcal{O}}\left(H^{3/2}L_dL_\lambda(1+\varepsilon_{h,m})\right)$ *where* $L_d = \max\left\{\sqrt{md}, L_\psi, \frac{L_r}{H}\right\}$, $L_\lambda = \max\{1, \sqrt{\lambda}\}$ $\alpha_U = 1/\widetilde{\mathcal{O}}(\sigma\sqrt{d})$ *and* $\alpha_L = \alpha_U/2$ *in Algorithm 5, then for any* $0 < \delta < \Phi(-1)/2$*, with probability at least* $1-\delta$*, the total regret of S-RLSVI is at most* $\widetilde{\mathcal{O}}\left(\sum_{h=1}^{H}\sigma\sqrt{(1+\varepsilon_{h,m})d\widetilde{m}_hK}\right)$*. If we further assume that* $\lambda = 1$ *and* $L_r, L_\psi = \widetilde{\mathcal{O}}(\sqrt{md})$*, and denote* $\varepsilon_m = \max_h\{\varepsilon_{h,m}\}$, $\widetilde{m} = \max_h\{\widetilde{m}_h\}$*. Then the regret bound can be simplified as* $\widetilde{\mathcal{O}}\left(\sqrt{(1+\varepsilon_m)^3md^2\widetilde{m}\cdot H^4T}\right)$.

Proof.

$$\text{Regret}(K) = \sum_{k=1}^{K}\left(V_1^{\star}(s_{1k}) - V_1^{\pi_k}(s_{1k})\right)$$

$$= \sum_{k=1}^{K} \left(V_1^\star(s_{1k}) - \overline{V}_{1k}(s_{1k})\right) + \sum_{k=1}^{K} \left(\overline{V}_{1k}(s_{1k}) - V_1^{\pi_k}(s_{1k})\right).$$

By Lemma 4.9.11, there exists absolute constant C such that for any $c_\beta > C$, if we set the parameters as described in Section 4.9.1, i.e.,

$$\begin{aligned}
\sqrt{\beta_h(\delta)} &\stackrel{\text{def}}{=} c_\beta \cdot HL_dL_\lambda(1+\varepsilon_{h,m})\sqrt{\iota} = \widetilde{\mathcal{O}}\left(HL_dL_\lambda(1+\varepsilon_{h,m})\right) \\
\sqrt{\nu_h(\delta)} &\stackrel{\text{def}}{=} 2\sqrt{\beta_h(\delta)} = \widetilde{\mathcal{O}}\left(HL_dL_\lambda(1+\varepsilon_{h,m})\right) \\
\sigma &\stackrel{\text{def}}{=} \sqrt{H\nu_h(\delta)} = \widetilde{\mathcal{O}}\left(H^{3/2}L_dL_\lambda(1+\varepsilon_{h,m})\right) \\
\sqrt{\gamma_h(\delta)} &\stackrel{\text{def}}{=} \sqrt{2dH\nu_h(\delta)\log(4dT/\delta)} = \widetilde{\mathcal{O}}\left(\sigma\sqrt{d}\right) \\
\alpha_U &\stackrel{\text{def}}{=} \frac{1}{4\sqrt{\gamma_h(\delta)}} \le \frac{1}{2\left(\sqrt{\nu_h(\delta)} + \sqrt{\gamma_h(\delta)}\right)} = 1/\widetilde{\mathcal{O}}(\sigma\sqrt{d}) \\
\alpha_L &\stackrel{\text{def}}{=} \alpha_U/2
\end{aligned}$$

then Lemma 4.9.14 and 4.9.15 imply that with probability at least $1-\delta$,

$$\begin{aligned}
\text{Regret}(K) &= \widetilde{\mathcal{O}}\left(\sum_{h=1}^{H}\left(\sqrt{\nu_h(\delta)} + \sqrt{\gamma_h(\delta)}\right)\sqrt{(1+\varepsilon_{h,m})K\widetilde{m}_h} + \sum_{h=1}^{H}\frac{H}{\alpha_L^2}(1+\varepsilon_{h,m})\widetilde{m}_h\right) \\
&= \widetilde{\mathcal{O}}\left(\sum_{h=1}^{H}\sqrt{(1+\varepsilon_{h,m})\gamma_h(\delta)K\widetilde{m}_h} + \sum_{h=1}^{H}\frac{H}{\alpha_L^2}(1+\varepsilon_{h,m})\widetilde{m}_h\right) \\
&= \widetilde{\mathcal{O}}\left(\sum_{h=1}^{H}\sigma\sqrt{(1+\varepsilon_{h,m})dK\widetilde{m}_h} + \sum_{h=1}^{H}\frac{H}{\alpha_L^2}(1+\varepsilon_{h,m})\widetilde{m}_h\right) \\
&= \widetilde{\mathcal{O}}\left(\sum_{h=1}^{H}\sigma\sqrt{(1+\varepsilon_{h,m})d\widetilde{m}_hK} + \sum_{h=1}^{H}(1+\varepsilon_{h,m})^3H^4L_d^2L_\lambda^2\widetilde{m}_h\right).
\end{aligned}$$

Observe that the second term does not scale with K so is a low order term, we get the desired bound. □

4.10 Auxilliary Lemmas and Proofs

This section states some auxilliary lemmas and gives their proofs.

Lemma 4.10.1 (Linear Weight and Norm Bound)**.** *For a linear MDP, for any policy π, there exists weights $\{\theta_h^\pi\}_{h\in[H]}$ such that for any $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, we have $Q_h^\pi(s,a) = \phi_h(s,a)^\top\theta_h^\pi$. Moreover, $\forall h\in[H], \|\theta_h^\pi\| \le L_r + HL_\psi$.*

Proof. By the Bellman equation $Q_h^\pi(s,a) = r_h(s,a) + \mathbb{E}_{s'|s,a}\left[V_{h+1}^\pi(s,a)\right]$ we can derive that

$$\theta_h^\pi = \theta_h^r + \int V_{h+1}^\pi(s)\psi_h(s).$$

Since $V_{h+1}^\pi(s) \le H$,

$$\|\theta_h^\pi\| \le \|\theta_h^r\| + H\int \|\psi_h(s)\| = L_r + HL_\psi.$$

□

Lemma 4.10.2 (Concentration of Self-Normalized Process [Abbasi-yadkori et al., 2011])**.** *Let $\{x_i\}_{i=1}^\infty$ be a real valued stochastic process with filtration $\{\mathcal{F}_i\}_{i=1}^\infty$. Let x_i be conditionally B-subgaussian given $\mathcal{F}_{i-1}$. Let $\{\phi_i\}_{i=1}^\infty$ with $\phi_i \in \mathcal{F}_{i-1}$ be a $\mathbb{R}^d$-valued stochastic process with $\|\phi_i\| \le L_\phi$. Let $\mathbf{\Sigma}_k = \lambda\mathbf{I}_d + \sum_{i=1}^k \phi_i\phi_i^\top$. Then for any $\delta > 0$, with probability at least $1-\delta$, for all $k \ge 0$, we have*

$$\left\|\sum_{i=1}^k \phi_i x_i\right\|_{\mathbf{\Sigma}_k^{-1}}^2 \le 2B^2\log\left(\frac{\det(\mathbf{\Sigma}_k)^{1/2}\det(\lambda\mathbf{I}_d)^{-1/2}}{\delta}\right).$$

Lemma 4.10.3 ([Zanette et al., 2019, Lemma I.3])**.** *With the above notation,*

$$\sum_{i=1}^k \|\phi_i\|_{\mathbf{\Sigma}_k^{-1}}^2 \le d.$$

Lemma 4.10.4 (Covering Number of Euclidean Ball, [Pollard, 1990, Section 4])**.** *A Euclidean ball of radius B in $\mathbb{R}^d$ has ε-covering number at most $(3B/\varepsilon)^d$.*

Lemma 4.10.5 (Covering Number of Low Rank Matrices)**.** *Let $S_r = \{\mathbf{X} \in \mathbb{R}^{d\times d} : \mathbf{X} \succeq 0, \text{rank}(\mathbf{X}) \leq r, \|X\|_F \leq B\}$. Then the ε-covering number of S_r with respect to the Frobenius norm is no more than $(9B/\varepsilon)^{(2d+1)r}$.*

Proof. [Candès and Plan, 2011, Lemma 3.1] states that the covering number of $\{\mathbf{X} \in \mathbb{R}^{d\times d} : \text{rank}(\mathbf{X}) \leq r, \|\mathbf{X}\|_F = 1\}$ is no more than $(9/\varepsilon)^{(2d+1)r}$. The extension to this lemma is trivial and is omitted. □

Lemma 4.10.6. *Let $\mathcal{V}$ denote the following class of function from $\mathcal{S}$ to $\mathbb{R}$*

$$V(\cdot) \stackrel{\text{def}}{=} \min\{\max_a \phi(\cdot, a)^\top \theta + \beta\sqrt{\phi(\cdot, a)^\top \mathbf{\Sigma}^{-1}\phi(\cdot, a)}, H\},$$

where $\theta \in \mathbb{R}^d$ satisfies $\|\theta\| \leq L$, $\beta \in [0, B]$ and $\mathbf{\Sigma} = \mathbf{H} + \lambda\mathbf{I_d}$ where $\lambda > 0, \mathbf{H} \in \mathbb{R}^{d\times d}$ is positive semi-definite, has rank at most m, and the Frobenius norm of $\mathbf{H}$ satisfies $\|\mathbf{H}\|_F \leq K$. Moreover, assume that $\|\phi(\cdot, \cdot)\| \leq L_\phi$. Then the ϵ-covering number $\mathcal{N}_\epsilon$ of $\mathcal{V}$ with respect to the distance $\text{dist}(V_1, V_2) \stackrel{\text{def}}{=} \sup_x |V_1(x) - V_2(x)|$ is upper bounded by

$$\log \mathcal{N}_\epsilon \leq d \log(6L/\epsilon) + (2d+1)m \cdot \log\left(36KB^2L_\phi^2 d/\lambda^2\epsilon^2\right).$$

Proof. Let $V_1(\cdot)$ be parametrized by $(\theta_1, \mathbf{\Sigma}_1^{-1}) = (\theta_1, (\mathbf{H}_1 + \lambda\mathbf{I}_d)^{-1})$ and $V_2(\cdot)$ be parametrized by $(\theta_2, \mathbf{\Sigma}_2^{-1}) = (\theta_1, (\mathbf{H}_2 + \lambda\mathbf{I}_d)^{-1}) \in O$. We write

$$\begin{aligned}
\text{dist}(V_1, V_2) =& \sup_x |V_1(x) - V_2(x)| \\
=& \sup_x \Bigg| \left(\min\{\max_a \phi(x, a)^\top \theta_1 + \beta\sqrt{\phi(x, a)^\top \mathbf{\Sigma}_1^{-1}\phi(x, a)}, H\}\right) \\
&- \left(\min\{\max_a \phi(x, a)^\top \theta_2 + \beta\sqrt{\phi(x, a)^\top \mathbf{\Sigma}_2^{-1}\phi(x, a)}, H\}\right) \Bigg|
\end{aligned}$$

$$
\begin{aligned}
&\leq \sup_x \left| \left(\max_a \phi(x,a)^\top \theta_1 + \beta \sqrt{\phi(x,a)^\top \boldsymbol{\Sigma}_1^{-1} \phi(x,a)} \right) \right. \\
&\quad \left. - \left(\max_a \phi(x,a)^\top \theta_2 + \beta \sqrt{\phi(x,a)^\top \boldsymbol{\Sigma}_2^{-1} \phi(x,a)} \right) \right| \\
&\leq \sup_{x,a} \left| \left(\phi(x,a)^\top \theta_1 + \beta \sqrt{\phi(x,a)^\top \boldsymbol{\Sigma}_1^{-1} \phi(x,a)} \right) \right. \\
&\quad \left. - \left(\phi(x,a)^\top \theta_2 + \beta \sqrt{\phi(x,a)^\top \boldsymbol{\Sigma}_2^{-1} \phi(x,a)} \right) \right| \\
&\leq \sup_{\|\phi\| \leq L_\phi} \left| \left(\phi^\top \theta_1 + \beta \sqrt{\phi^\top \boldsymbol{\Sigma}_1^{-1} \phi} \right) - \left(\phi^\top \theta_2 + \beta \sqrt{\phi^\top \boldsymbol{\Sigma}_2^{-1} \phi} \right) \right| \\
&\leq L_\phi \left(\sup_{\|\phi\| \leq 1} |\phi^\top (\theta_1 - \theta_2)| + \beta \sup_{\|\phi\| \leq 1} \sqrt{\phi^\top (\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}) \phi} \right) \\
&= L_\phi \|\theta_1 - \theta_2\| + B L_\phi \sup_{\|\phi\| \leq 1} \sqrt{\phi^\top (\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}) \phi}. \qquad (4.21)
\end{aligned}
$$

Let $f(\mathbf{H}) \stackrel{\text{def}}{=} \phi^\top \left(\mathbf{H} + \lambda \mathbf{I}_d\right)^{-1} \phi$ where $\|\phi\| \leq 1$, then $\frac{\partial f(\mathbf{H})}{\partial \mathbf{H}_{ij}} = -\phi^\top \left(\mathbf{H} + \lambda \mathbf{I}_d\right)^{-1} \mathbf{E}_{ij} \left(\mathbf{H} + \lambda \mathbf{I}_d\right)^{-1} \phi$ where $\mathbf{E}_{ij}$ is a $d \times d$ matrix whose value is all 0 except the entry on the ith row and jth column which is 1. Hence,

$$
\left| \frac{\partial f(\mathbf{H})}{\partial \mathbf{H}_{ij}} \right| \leq \|\phi^\top \left(\mathbf{H} + \lambda \mathbf{I}_d\right)^{-1} \|^2 \leq \frac{1}{\lambda^2}.
$$

Therefore,

$$
\begin{aligned}
\phi^\top (\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}) \phi &= f(\mathbf{H}_1) - f(\mathbf{H}_2) \\
&\leq \frac{1}{\lambda^2} \|\text{vec}(\mathbf{H}_2 - \mathbf{H}_1)\|_1 \leq \frac{d}{\lambda^2} \|\mathbf{H}_1 - \mathbf{H}_2\|_F. \qquad (4.22)
\end{aligned}
$$

Combining (4.21) and (4.22), we have

$$
\text{dist}(V_1, V_2) \leq L_\phi \|\theta_1 - \theta_2\| + \sqrt{\frac{B^2 L_\phi^2 d \|\mathbf{H}_1 - \mathbf{H}_2\|_F}{\lambda^2}}. \qquad (4.23)
$$

Now, let $\mathcal{C}_\theta$ be a $\epsilon/(2L\phi)$-cover of $\{\theta \in \mathbb{R}^d : \|\theta\| \le L\}$ with respect to the 2-norm, and $\mathcal{C}_{\mathbf{H}}$ be a $\lambda^2\epsilon^2/(4B^2L_\phi^2 d)$-cover of $\{\mathbf{H} \in \mathbb{R}^{d\times d} : \mathbf{H} \succeq 0, \text{rank}(\mathbf{H}) \le r, \|\mathbf{H}\|_F \le K\}$ with respect to the Frobineus norm. Lemma 4.10.4 and Lemma 4.10.5 yield that

$$|\mathcal{C}_\theta| \le (6LL_\phi/\epsilon)^d, \quad |\mathcal{C}_{\mathbf{H}}| \le \left(36KB^2L_\phi^2 d/\lambda^2\epsilon^2\right)^{(2d+1)m}.$$

By (4.23), for any $V_1 \in V$ parametrized by θ_1 and $\mathbf{H}_1$, there exists $\theta_2 \in \mathcal{C}_\theta$ and $\mathbf{H}_2 \in \mathcal{C}_{\mathbf{H}}$ such that V_2 parametrized by θ_2 and $\mathbf{H}_2$ is within ϵ distance to V_1. Hence,

$$\log \mathcal{N}_\epsilon \le \log|\mathcal{C}_\theta| + \log|\mathcal{C}_{\mathbf{H}}| = d\log(6L/\epsilon) + (2d+1)m \cdot \log\left(36KB^2L_\phi^2 d/\lambda^2\epsilon^2\right).$$

□

Lemma 4.10.7. *Let O denote the following set*

$$O \stackrel{\text{def}}{=} \{\theta \in \mathbb{R}^d, \mathbf{\Sigma} \in \mathbb{R}^{d\times d} : \|\theta\| \le L, \mathbf{\Sigma} = (\mathbf{H} + \lambda\mathbf{I}_d)^{-1}, \mathbf{H} \succeq 0, \text{rank}(\mathbf{H}) \le m, \|\mathbf{H}\|_F \le K\}.$$

For any $O_1 = (\theta_1, \mathbf{\Sigma}_1) = (\theta_1, (\mathbf{H}_1 + \lambda\mathbf{I}_d)^{-1}) \in O$ and $O_2 = (\theta_2, \mathbf{\Sigma}_2) = (\theta_1, (\mathbf{H}_2 + \lambda\mathbf{I}_d)^{-1}) \in O$ define the distance $\text{dist}(O_1, O_2) \stackrel{\text{def}}{=} \max\{\|\theta_1 - \theta_2\|, \|\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2\|\}$. The ϵ-covering number $\mathcal{N}_\epsilon$ of O with respect to the distance is upper bounded by

$$\log \mathcal{N}_\epsilon \le d\log(3L/\epsilon) + (2d+1)m \cdot \log\left(9Kd/\lambda^2\epsilon^2\right).$$

Proof. By (4.22),

$$\|\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2\|^2 = \max_{\phi:\|\phi\|\le 1} \left\{\phi^\top(\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2)\phi\right\} \le \frac{d}{\lambda^2}\|\mathbf{H}_1 - \mathbf{H}_2\|_F. \tag{4.24}$$

Now, let $\mathcal{C}_\theta$ be a ϵ-cover of $\{\theta \in \mathbb{R}^d : \|\theta\| \le L\}$ with respect to the 2-norm, and $\mathcal{C}_{\mathbf{H}}$ be

a $\lambda^2\epsilon^2/d$-cover of $\{\mathbf{H} \in \mathbb{R}^{d\times d} : \mathbf{H} \succeq 0, \text{rank}(\mathbf{H}) \leq r, \|\mathbf{H}\|_F \leq K\}$ with respect to the Frobineus norm. Lemma 4.10.4 and Lemma 4.10.5 yield that

$$|\mathcal{C}_\theta| \leq (3L/\epsilon)^d, \quad |\mathcal{C}_{\mathbf{H}}| \leq \left(9Kd/\lambda^2\epsilon^2\right)^{(2d+1)m}.$$

By (4.24), for any $O_1 \in O$ parametrized by θ_1 and $\mathbf{H}_1$, there exists $\theta_2 \in \mathcal{C}_\theta$ and $\mathbf{H}_2 \in \mathcal{C}_{\mathbf{H}}$ such that O_2 parametrized by θ_2 and $\mathbf{H}_2$ is within ϵ distance to V_1. Hence,

$$\log \mathcal{N}_\epsilon \leq \log|\mathcal{C}_\theta| + \log|\mathcal{C}_{\mathbf{H}}| = d\log(3L/\epsilon) + (2d+1)m \cdot \log\left(9Kd/\lambda^2\epsilon^2\right).$$

□

Lemma 4.10.8 ([Abeille and Lazaric, 2017, Appendix A]). *Let $\xi \sim \mathcal{N}(0, \mathbf{\Sigma})$ for some $\mathbf{\Sigma}$ that is positive definite. For any $\delta > 0$, with probability at least $1-\delta$,*

$$\|\xi\|_{\mathbf{\Sigma}^{-1}} \leq \sqrt{2d\log(2d/\delta)}.$$

4.11 Experiment Details

We first describe how we encode the linear structure in the RiverSwim environment (illustrated in Figure 4.1) in a compact fashion; then we describe the infinite-state Labyrinth environment in more details and how we encode its linear structure using a finite dimensional feature space. Note the transition probability, rewards and feature map are the same within each episode in our design, so we may omit the subscript h in notations for brevity.

4.11.1 RiverSwim Environment

This environment is a tabular environment, because it has finite number of states and actions. Therefore, it can be encoded in a canonical way by setting $d = |\mathcal{S}| \times |\mathcal{A}|$ and letting

ϕ indicates each state and action pair. However, we describe in the following how we compactly encode the linear structure of a n-state RiverSwim environment using feature dimension $d = n + 1$ instead of $d = 2n$.

Figure 4.4 illustrates how we define ψ, ϕ, θ^r to construct the compact representation of the RiverSwim environment. Let the first n entries of a vector in $\mathbb{R}^d$ be indexed by $s \in \mathcal{S} = \{s_1, \ldots, s_n\}$, and let e_i be the standard basis of $\mathbb{R}^d$. Observe that there is one special state-action pair that will incur reward—swim right at s_n. For all other state-action pairs that do not incur reward, we can set $\phi(s,a)_{s'} = \mathbb{P}(s'|s,a)$ and $\psi(s') = e_{s'}$ to recover the state transition probabilities, and set θ^r be zero on the leading n entries to recover the reward function. We can then employ the last dimension of the feature space to capture the state transition reward in the special state-action pair. Namely, for swimming right at s_n, let $\phi(s_n, \text{right}) = e_d$, $\psi(s')_d = \mathbb{P}(s'|s_n, \text{right})$ and θ^r's last entry be $r = 1$ to recover the state transition probability and the reward under this state-action pair.

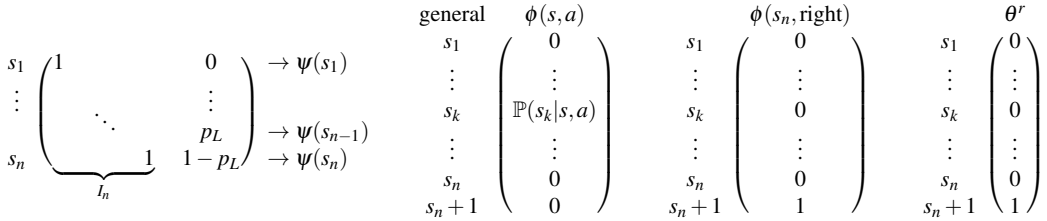

Figure 4.4: ψ, ϕ, θ^r for RiverSwim Environment.

4.11.2 Labyrinth Environment

This environment has infinite number of states ($\mathbb{N}_0$) and thus the canonical encoding of tabular environments does not apply. We explain in the following how to encode the Labyrinth environment as a linear MDP problem with d-dimensional feature space.

Given episode length H and action space $\mathcal{A}$, starting from state 0, the agent will randomly jump among the non-negative integers. Only if the agent is currently at $s_{\text{goal}} =$

$(H-1)d \in \mathbb{N}_0$ and takes an pre-defined action a_{goal} will it incur a substantial reward $r = 1$, otherwise the agent will optionally receive a diminutive auxiliary reward r_a. In general, upon selecting an action, the agent will be transferred to the next state randomly, and the state transition probability measures are designed (explained below) so that the probability of arriving at s_{goal} is very small. However, there is a shortcut from state 0 to state s_{goal}. Namely, if the current state is $s^i = (i-1)d$ for $i \in [H]$, selecting a pre-defined specific action $a^i \in \mathcal{A}$ will deterministically send the agent to state $s^i + d$. Since it is very unlikely to arrive at s_{goal} without taking the shortcut and all other actions will only incur diminutive reward, the optimal strategy is to select these a^i at s^i for $i \in [H]$. Note that $(s^H, a^H) = (s_{\text{goal}}, a_{\text{goal}})$.

Our explicit encoding of this environment is as follows:

1. We leave the last H dimensions of the feature space to encode the H shortcut state-action pairs $\{(s^i, a^i)\}_{i=1}^H$ and design ψ, ϕ, θ^r as follows:

   ϕ : Each shortcut state-action pair occupies one dimension of the feature space, i.e., $\phi(s^i, a^i) = e_{d-H+i}$ for $i \in [H]$.

   ψ : For each $i \in [H]$, $e_{d-H+i}^\top \psi$ is a probability measure over $\mathbb{N}_0$, satisfying the rule that the agent taking a^i at state s^i will be deterministically transferred to state $s^i + d$. In other words, for each $i \in [H]$, $e_{d-H+i}^\top \psi(x) = 1$ for $x = id$ and $e_{d-H+i}^\top \psi(x) = 0$ for all other x.

   θ^r : Set the last entry of θ^r to 1. If there is auxiliary reward, set the last but one $H-1$ entries to be r_a, else set to zero.

2. We now explain how to construct ψ, ϕ, θ for all other state-action pairs.

   ϕ : We use a surjective function f that maps $\mathbb{N}_0 \times \mathcal{A}$ to $[d-H]$ to define the feature map $\phi : (s, a) \mapsto e_{f(s,a)}$.

ψ : $\mathbb{N}_0 \to \mathbb{R}^d$: It suffices to illustrate the measure $e_i^\top \psi$ on $\mathbb{N}_0$ for each $i \in [d-H]$. Consider a decaying factor $\gamma \geq 2$, Set a measure such that

$$e_i^\top \psi(k) = \begin{cases} (\gamma - 1)/\gamma^{(k/i)+1} & \text{if } k \bmod i = 0, \\ 0 & \text{otherwise.} \end{cases}$$

θ^r : Set the first $d - H$ entries of θ^r to be r_a if there is auxiliary reward; else set to zero.

One can check that it satisfies the norm condition of Assumption 4.2.1. The transition probability and the reward function of this environment are defined using the linearity condition introduced in Assumption 4.2.1 that associates ϕ, ψ, θ^r and $\mathbb{P}, r$. Therefore, the linearity assumptions in Assumption 4.2.1 are trivially satisfied.

The two configurations of this environment presented in this chapter are:

feature dimension d	decay factor γ	auxiliary reward r_a	# actions	# steps per episode
100	4	0	3	3
200	2	3×10^{-4}	3	2

Table 4.1: Labyrinth environment configurations

Bibliography

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pages 265–283, 2016.

Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24*, pages 2312–2320. 2011.

Marc Abeille and Alessandro Lazaric. Linear thompson sampling revisited. *Electron. J. Statist.*, 11(2):5165–5197, 2017.

Jan M. Ache, Jason Polsky, Shada Alghailani, Ruchi Parekh, Patrick Breads, Martin Y. Peek, Davi D. Bock, Catherine R. von Reyn, and Gwyneth M. Card. Neural basis for looming size and velocity encoding in the *Drosophila* giant fiber escape pathway. *Current Biology*, 29(6):1073–1081, 2019/05/18 2019a. doi: 10.1016/j.cub.2019.01.079. URL https://doi.org/10.1016/j.cub.2019.01.079.

Jan M Ache, Jason Polsky, Shada Alghailani, Ruchi Parekh, Patrick Breads, Martin Y Peek, Davi D Bock, Catherine R von Reyn, and Gwyneth M Card. Neural basis for looming size and velocity encoding in the drosophila giant fiber escape pathway. *Current Biology*, 29(6):1073–1081, 2019b.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 127–135, 2013.

Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems 30*, pages 1184–1194. 2017.

Ahmed El Alaoui and Michael W. Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *NIPS*, 2015.

E. Arias-Castro, D. L. Donoho, and Xiaoming Huo. Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Transactions on Information Theory*, 51(7): 2402–2425, July 2005.

Ery Arias-Castro, Emmanuel J. Candès, and Arnaud Durand. Detection of an anomalous cluster in a network. *Ann. Statist.*, 39(1):278–304, 02 2011.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *ICML*, 2017.

Bara A Badwan, Matthew S Creamer, Jacob A Zavatone-Veth, and Damon A Clark. Dynamic nonlinearities enable direction opponency in drosophila elementary motion detectors. *Nature neuroscience*, 22(8):1318–1326, 2019.

William Ball and Edward Tronick. Infant responses to impending collision: Optical and real. *Science*, 171(3973):818–820, 1971.

Kiran Bhattacharyya, David L McLean, and Malcolm A MacIver. Visual threat assessment and reticulospinal encoding of calibrated responses in larval zebrafish. *Current Biology*, 27(18):2751–2762, 2017.

Steven J. Bradtke and Andrew G. Barto. Linear least-squares algorithms for temporal difference learning. *Mach. Learn.*, 22:33–57, 1996.

Timothy F. Brady, Talia Konkle, and George A. Alvarez. A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of Vision*, 11(5):4–4, May 2011. ISSN 1534-7362. doi: 10.1167/11.5.4. URL http://jov.arvojournals.org/article.aspx?articleid=2191865.

Lawrence D. Brown, T. Tony Cai, and Harrison H. Zhou. Robust nonparametric estimation via wavelet median regression. *Ann. Statist.*, 36(5):2055–2084, 10 2008.

Jon Cafaro, Joel Zylberberg, and Greg D Field. Global motion processing by populations of direction-selective retinal ganglion cells. *Journal of Neuroscience*, 40(30):5807–5819, 2020.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. *ArXiv*, abs/1912.05830, 2019.

Daniele Calandriello, Alessandro Lazaric, and Michal Valko. Second-order kernel online convex optimization with adaptive sketching. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 645–653, 2017.

Daniele Calandriello, Luigi Carratino, Alessandro Lazaric, Michal Valko, and Lorenzo Rosasco. Gaussian process optimization with adaptive sketching: Scalable and no regret. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 533–557, 2019.

Emmanuel J. Candès and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57:2342–2359, 2011.

Gwyneth Card and Michael H Dickinson. Visually mediated motor planning in the escape response of drosophila. *Current Biology*, 18(17):1300–1307, 2008.

Hock Peng Chan and Guenther Walther. Detection with the scan and the average likelihood ratio. *Statistica Sinica*, 23(1):409–428, 2013.

Juyue Chen, Holly B Mandel, James E Fitzgerald, and Damon A Clark. Asymmetric on-off processing of visual motion cancels variability induced by the structure of natural scenes. *Elife*, 8:e47579, 2019.

Michael B. Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Madalina Persu. Dimensionality reduction for k-means clustering and low rank approximation. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '15, page 163–172. Association for Computing Machinery, 2015.

Matthew S Creamer, Omer Mano, and Damon A Clark. Visual control of walking speed in drosophila. *Neuron*, 100(6):1460–1473, 2018.

Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 1507–1516, 2019.

Pratyay Datta and Bodhisattva Sen. Optimal inference with a multidimensional multiscale statistic, 2018. arXiv preprint arXiv:1806.02194.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

Saskia EJ De Vries and Thomas R Clandinin. Loom-sensitive neurons link computation to action in the drosophila visual system. *Current Biology*, 22(5):353–362, 2012.

Brian D DeAngelis, Jacob A Zavatone-Veth, and Damon A Clark. The manifold structure of limb coordination in walking drosophila. *Elife*, 8:e46409, 2019.

Michael S Drews, Aljoscha Leonhardt, Nadezhda Pirogova, Florian G Richter, Anna Schuetzenberger, Lukas Braun, Etienne Serbe, and Alexander Borst. Dynamic signal compression for robust motion vision in flies. *Current Biology*, 30(2):209–221, 2020.

Simon S. Du, Yuping Luo, Ruosong Wang, and Hanrui Zhang. Provably efficient q-learning with function approximation via distribution shift error checking oracle. In *NeurIPS*, 2019.

Lutz Dümbgen and Vladimir G. Spokoiny. Multiscale testing of qualitative hypotheses. *Ann. Statist.*, 29:124–152, 2001.

Lutz Dümbgen and Günther Walther. Multiscale inference about a density. *Ann. Statist.*, 36:1758–1785, 2008.

Timothy W Dunn, Christoph Gebhardt, Eva A Naumann, Clemens Riegler, Misha B Ahrens, Florian Engert, and Filippo Del Bene. Neural circuits underlying visually evoked escapes in larval zebrafish. *Neuron*, 89(3):613–628, 2016.

Yonathan Efroni, Nadav Merlis, Mohammad Ghavamzadeh, and Shie Mannor. Tight regret bounds for model-based reinforcement learning with greedy policies. In *NeurIPS*, 2019.

David J Field. What is the goal of sensory coding? *Neural computation*, 6(4):559–601, 1994.

James E Fitzgerald and Damon A Clark. Nonlinear circuits for naturalistic visual motion estimation. *Elife*, 4:e09123, 2015.

Felix Franke, Michele Fiscella, Maksim Sevelev, Botond Roska, Andreas Hierlemann, and Rava Azeredo da Silveira. Structures of neural correlation and how they favor coding. *Neuron*, 89(2):409–422, 2016.

Fabrizio Gabbiani, Holger G Krapp, and Gilles Laurent. Computation of object approach by a wide-field, motion-sensitive neuron. *Journal of Neuroscience*, 19(3):1122–1141, 1999.

Apostolos P Georgopoulos, Andrew B Schwartz, and Ronald E Kettner. Neuronal population coding of movement direction. *Science*, 233(4771):1416–1419, 1986.

Mina Ghashami, Edo Liberty, Jeff M. Phillips, and David P. Woodruff. Frequent directions: Simple and deterministic matrix sketching. *SIAM Journal on Computing*, 45(5): 1762–1792, 2016.

Mohammad Ghavamzadeh, Alessandro Lazaric, Odalric Maillard, and Rémi Munos. Lstd with random projections. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 721–729. Curran Associates, Inc., 2010.

James J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979.

Joseph Glaz, Vladimir Pozdnyakov, and Sylvan Wallenstein. *Scan Statistics: Methods and Applications*. 2009.

Eyal Gruntman, Sandro Romani, and Michael B Reiser. Simple integration of fast excitation and offset, delayed inhibition computes directional selectivity in drosophila. *Nature neuroscience*, 21(2):250–257, 2018.

Eyal Gruntman, Sandro Romani, and Michael B Reiser. The computation of directional selectivity in the drosophila off motion pathway. *Elife*, 8:e50706, 2019.

James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

Bernhard Hassenstein and Werner Reichardt. Systemtheoretische analyse der zeit-, reihenfolgen-und vorzeichenauswertung bei der bewegungsperzeption des rüsselkäfers chlorophanus. *Zeitschrift für Naturforschung B*, 11(9-10):513–524, 1956.

Alexis Hervais-Adelman, Lore B Legrand, Minye Zhan, Marco Tamietto, Beatrice de Gelder, and Alan J Pegna. Looming sensitive cortical regions without v1 input: evidence from a patient with bilateral cortical blindness. *Frontiers in integrative neuroscience*, 9:51, 2015.

M. N. Huxley. Exponential sums and lattice points iii. *Proceedings of the London Mathematical Society*, 87(3):591–609, 2003.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11:1563–1600, 2010.

X. Jessie Jeng, T. Tony Cai, and Hongzhe Li. Optimal sparse segment identification with application in copy number variation analysis. *Journal of the American Statistical Association*, 105(491):1156–1166, 2010.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. *ArXiv*, abs/1907.05388, 2019.

Robert E. Kass, S. Amari, Kensuke Arai, Emery N. Brown, Casey O. Diekman, Markus Diesmann, Brent Doiron, Uri T. Eden, Adrienne L. Fairhall, Grant M Fiddyment, Tomoki Fukai, Sonja Grün, Matthew T. Harrison, Moritz Helias, Hiroyuki Nakahara, Jun nosuke Teramae, Peter J. Thomas, Mark A Reimers, Jordan Rodu, Horacio G. Rotstein, Eric Shea-Brown, Hideaki Shimazaki, Shigeru Shinomoto, Byron M Yu, and Mark A. Kramer. Computational neuroscience: Mathematical and statistical perspectives. *Annual review of statistics and its application*, 5:183–214, 2018.

Akiko Kikuchi, Shumpei Ohashi, Naoyuki Fuse, Toshiaki Ohta, Marina Suzuki, Yoshinori Suzuki, Tomoyo Fujita, Takuya Miyamoto, Toru Aonishi, Hiroyoshi Miyakawa, et al. Experience-dependent plasticity of the optomotor response in drosophila melanogaster. *Developmental neuroscience*, 34(6):533–542, 2012.

Sheila M King, Caroline Dykeman, Peter Redgrave, and Paul Dean. Use of a distracting task to obtain defensive head movements to looming visual stimuli by human adults in a laboratory setting. *Perception*, 21(2):245–259, 1992.

Nathan C Klapoetke, Aljoscha Nern, Martin Y Peek, Edward M Rogers, Patrick Breads, Gerald M Rubin, Michael B Reiser, and Gwyneth M Card. Ultra-selective looming detection from radial motion opponency. *Nature*, 551(7679):237, 2017.

Jens Kober, J. Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32:1238 – 1274, 2012.

Claudia König, Axel Munk, and Frank Werner. Multidimensional multiscale scanning in exponential families: Limit theory and statistical consequences, 2018. arXiv preprint arXiv:1802.07995.

Jiyao Kou. Identifying the support of rectangular signals in gaussian noise, 2017. arXiv preprint arXiv:1703.06226.

Martin Kulldorff. A spatial scan statistic. *Communications in Statistics - Theory and Methods*, 26(6):1481–1496, 1997.

Martin Kulldorff, Richard Heffernan, Jessica Hartman, Renato Assunçõ, and Farzad Mostashari. A space-time permutation scan statistic for disease outbreak detection. *PLOS Medicine*, 2(3), 02 2005.

Ilja Kuzborskij, Leonardo Cella, and Nicolò Cesa-Bianchi. Efficient linear bandits through matrix sketching. In *Proceedings of Machine Learning Research*, volume 89, pages 177–185, 2019.

David N Lee. A theory of visual control of braking based on information about time-to-collision. *Perception*, 5(4):437–459, 1976.

Aljoscha Leonhardt, Georg Ammer, Matthias Meier, Etienne Serbe, Armin Bahl, and Alexander Borst. Asymmetry of drosophila on and off motion detectors enhances real-world velocity estimation. *Nature neuroscience*, 19(5):706–715, 2016.

Edo Liberty. Simple and deterministic matrix sketching. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, page 581–588. Association for Computing Machinery, 2013.

Yong-Jun Liu, Qian Wang, and Bing Li. Neuronal responses to looming objects in the superior colliculus of the cat. *Brain, Behavior and Evolution*, 77(3):193–205, 2011.

Haipeng Luo, Alekh Agarwal, Nicolò Cesa-Bianchi, and John Langford. Efficient second order online learning by sketching. *ArXiv*, abs/1602.02202, 2016.

Luo Luo, Cheng Chen, Zhihua Zhang, Wu-Jun Li, and Tong Zhang. Robust frequent directions with application in online learning. *J. Mach. Learn. Res.*, 20:45:1–45:41, 2019.

Matthew S Maisak, Juergen Haag, Georg Ammer, Etienne Serbe, Matthias Meier, Aljoscha Leonhardt, Tabea Schilling, Armin Bahl, Gerald M Rubin, Aljoscha Nern, et al. A directional tuning map of drosophila elementary motion detectors. *Nature*, 500 (7461):212, 2013.

Catherine A Matulis, Juyue Chen, Aneysis D Gonzalez-Suarez, Rudy Behnia, and Damon A Clark. Heterogeneous temporal contrast adaptation in drosophila direction-selective circuits. *Current Biology*, 30(2):222–236, 2020.

Alex S Mauss, Katarina Pankova, Alexander Arenz, Aljoscha Nern, Gerald M Rubin, and Alexander Borst. Neural circuit to integrate opposing motions in the visual field. *Cell*, 162(2):351–362, 2015.

Ronald W McLeod and Helen E Ross. Optic-flow and cognitive factors in time-to-collision estimates. *Perception*, 12(4):417–423, 1983.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen. King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

Millett Granger Morgan, Max Henrion, and Mitchell Small. *Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis*. Cambridge university press, 1990.

Mai M Morimoto, Aljoscha Nern, Arthur Zhao, Edward M Rogers, Allan M Wong, Mathew D Isaacson, Davi D Bock, Gerald M Rubin, and Michael B Reiser. Spatial readout of visual looming in the central brain of drosophila. *Elife*, 9:e57685, 2020.

Florian T Muijres, Michael J Elzinga, Johan M Melis, and Michael H Dickinson. Flies evade looming targets by executing rapid visually directed banked turns. *Science*, 344 (6180):172–177, 2014.

Thomas A Münch, Rava Azeredo Da Silveira, Sandra Siegert, Tim James Viney, Gautam B Awatramani, and Botond Roska. Approach sensitivity in the retina processed by a multifunctional neural circuit. *Nature neuroscience*, 12(10):1308–1316, 2009.

Daniel B. Neill. An empirical comparison of spatial scan statistics for outbreak detection. *International Journal of Health Geographics*, 8(1):20, Apr 2009.

Damián Oliva and Daniel Tomsic. Computation of object approach by a system of visual motion-sensitive neurons in the crab neohelice. *Journal of neurophysiology*, 112(6): 1477–1490, 2014.

Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

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.

Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2701–2710. JMLR. org, 2017.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems 26*, pages 3003–3011. 2013.

Ian Osband, Benjamin Van Roy, Daniel J. Russo, and Zheng Wen. Deep exploration via randomized value functions. *J. Mach. Learn. Res.*, 20:1–62, 2019.

Yangchen Pan, Erfan Sadeqi Azer, and Martha White. Effective sketching methods for value function approximation. *ArXiv*, abs/1708.01298, 2017.

Anitha Pasupathy and Charles E Connor. Population coding of shape in area v4. *Nature neuroscience*, 5(12):1332–1338, 2002.

Martin Y Peek and Gwyneth M Card. Comparative approaches to escape. *Current opinion in neurobiology*, 41:167–173, 2016.

David Pollard. Empirical processes: Theory and applications. *NSF-CBMS Regional Conference Series in Probability and Statistics*, 2:i–86, 1990.

Marc Potters and William Bialek. Statistical mechanics and visual signal processing. *Journal de Physique I*, 4(11):1755–1775, 1994.

Martin L. Puterman. Markov decision processes: Discrete stochastic dynamic programming. In *Wiley Series in Probability and Statistics*, 1994.

D Regan and KI Beverley. Looming detectors in the human visual pathway. *Vision research*, 18(4):415–421, 1978.

D. Regan and S.J. Hamstra. Dissociation of discrimination thresholds for time to contact and for rate of angular expansion. *Vision Research*, 33(4):447 – 462, 1993.

F Claire Rind and DI Bramwell. Neural network based on the input organization of an identified neuron signaling impending collision. *Journal of neurophysiology*, 75(3): 967–985, 1996.

Camilo Rivera and Guenther Walther. Optimal detection of a jump in the intensity of a poisson process or in a density with likelihood ratio statistics. *Scand. J. Stat.*, 40(4): 752–769, 2013.

Daniel L Ruderman and William Bialek. Statistics of natural images: Scaling in the woods. *Physical review letters*, 73(6):814, 1994.

Lindsey D Salay, Nao Ishiko, and Andrew D Huberman. A midline thalamic circuit determines reactions to visual threat. *Nature*, 557(7704):183–189, 2018.

Emilio Salazar-Gatzimas, Juyue Chen, Matthew S Creamer, Omer Mano, Holly B Mandel, Catherine A Matulis, Joseph Pottackal, and Damon A Clark. Direct measurement of correlation responses in drosophila elementary motion detectors reveals fast timescale tuning. *Neuron*, 92(1):227–239, 2016.

Roger D Santer, Peter J Simmons, and F Claire Rind. Gliding behaviour elicited by lateral looming stimuli in flying locusts. *Journal of Comparative Physiology A*, 191(1):61–73, 2005.

Keiichiro Sato and Yoshifumi Yamawaki. Role of a looming-sensitive neuron in triggering the defense behavior of the praying mantis tenodera aridifolia. *Journal of neurophysiology*, 112(3):671–682, 2014.

William Schiff and Mary Lou Detwiler. Information used in judging impending collision. *Perception*, 8(6):647–658, 1979.

Congping Shang, Zhihui Liu, Zijun Chen, Yingchao Shi, Qian Wang, Su Liu, Dapeng Li, and Peng Cao. A parvalbumin-positive excitatory visual pathway to trigger fear responses in mice. *Science*, 348(6242):1472–1477, 2015.

James Sharpnack and Ery Arias-Castro. Exact asymptotics for the scan statistic and fast alternatives. *Electron. J. Statist.*, 10:2641–2684, 2016.

Kazunori Shinomiya, Gary Huang, Zhiyuan Lu, Toufiq Parag, C Shan Xu, Roxanne Aniceto, Namra Ansari, Natasha Cheatham, Shirley Lauchie, Erika Neace, et al. Comparisons between the on-and off-edge motion pathways in the drosophila brain. *Elife*, 8: e40025, 2019.

David Siegmund and Benjamin Yakir. Tail probabilities for the null distribution of scanning statistics. *Bernoulli*, 6(2):191–213, 04 2000.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, L Robert Baker, Matthew Lai, Adrian Bolton, Yutian chen, Timothy P. Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. *Nature*, 550:354–359, 2017.

Kenneth O Stanley, Jeff Clune, Joel Lehman, and Risto Miikkulainen. Designing neural networks through neuroevolution. *Nature Machine Intelligence*, 1(1):24–35, 2019.

DG Stavenga. Angular and spectral sensitivity of fly photoreceptors. ii. dependence on facet lens f-number and rhabdomere type in drosophila. *Journal of Comparative Physiology A*, 189(3):189–202, 2003.

Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

Hongjin Sun and Barrie J Frost. Computation of different optical variables of looming objects in pigeon nucleus rotundus neurons. *Nature neuroscience*, 1(4):296–303, 1998.

Shin-ya Takemura, Aljoscha Nern, Dmitri B Chklovskii, Louis K Scheffer, Gerald M Rubin, and Ian A Meinertzhagen. The comprehensive connectome of a neural substrate for 'on'motion detection in drosophila. *Elife*, 6:e24394, 2017.

MARK A Tanouye and ROBERT J Wyman. Motor outputs of giant nerve fiber in drosophila. *Journal of neurophysiology*, 44(2):405–421, 1980.

Incinur Temizer, Joseph C Donovan, Herwig Baier, and Julia L Semmelhack. A visual pathway for looming-evoked escape in larval zebrafish. *Current Biology*, 25(14):1823–1834, 2015.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Rufin Vogels. Population coding of stimulus orientation by striate cortical cells. *Biological cybernetics*, 64(1):25–31, 1990.

Catherine R Von Reyn, Patrick Breads, Martin Y Peek, Grace Zhiyu Zheng, W Ryan Williamson, Alyson L Yee, Anthony Leonardo, and Gwyneth M Card. A spike-timing mechanism for action selection. *Nature neuroscience*, 17(7):962, 2014.

Catherine R Von Reyn, Aljoscha Nern, W Ryan Williamson, Patrick Breads, Ming Wu, Shigehiro Namiki, and Gwyneth M Card. Feature integration drives probabilistic behavior in the drosophila escape response. *Neuron*, 94(6):1190–1204, 2017.

Guenther Walther. Optimal and fast detection of spatial clusters with scan statistics. *Ann. Statist.*, 38(2):1010–1033, 04 2010.

Yining Wang, Ruosong Wang, Simon Shaolei Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. *ArXiv*, abs/1912.04136, 2019.

John P. Wann, Damian R. Poulter, and Catherine Purcell. Reduced sensitivity to visual looming inflates the risk posed by speeding vehicles when children try to cross the road. *Psychological science*, 22 4:429–34, 2011.

David P. Woodruff. Sketching as a tool for numerical linear algebra. *Found. Trends Theor. Comput. Sci.*, 10(1–2):1–157, 2014.

Le-Qing Wu, Yu-Qiong Niu, Jin Yang, and Shu-Rong Wang. Tectal neurons signal impending collision of looming objects in the pigeon. *European Journal of Neuroscience*, 22(9):2325–2331, 2005.

Ming Wu, Aljoscha Nern, W Ryan Williamson, Mai M Morimoto, Michael B Reiser, Gwyneth M Card, and Gerald M Rubin. Visual projection neurons in the drosophila lobula link feature detection to distinct behavioral programs. *Elife*, 5:e21022, 2016.

Jing-Jiang Yan, Bailey Lorv, Hong Li, and Hong-Jin Sun. Visual processing of the impending collision of a looming object: Time to collision revisited. *Journal of Vision*, 11 (12), 2011.

Lin F. Yang and Mengdi Wang. Reinforcement leaning in feature space: Matrix bandit, kernels, and regret bound. *ArXiv*, abs/1905.10389, 2019a.

Lin F. Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *ICML*, 2019b.

Yun Yang, Mert Pilanci, and Martin J. Wainwright. Randomized sketches for kernels: Fast and optimal nonparametric regression. *Ann. Statist.*, 45(3):991–1023, 2017.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7304–7312, 2019.

Andrea Zanette, David Brandfonbrener, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. *ArXiv*, abs/1911.00567, 2019.

Jacob A Zavatone-Veth, Bara A Badwan, and Damon A Clark. A minimal synaptic model for direction selective neurons in drosophila. *Journal of vision*, 20(2):2–2, 2020.

Joel Zylberberg, Jon Cafaro, Maxwell H Turner, Eric Shea-Brown, and Fred Rieke. Direction-selective circuits shape noise to ensure a precise population code. *Neuron*, 89(2):369–383, 2016.