SYSTEM COMPLEXITY, STABILITY AND PERFORMANCE: APPLICATION TO PREDICTION

JOEL RATSABY

Abstract. Complexity and stability of a system are interrelated. We consider a system that predicts an input binary Markov chain. The system's output is a binary sequence that indicates when prediction errors occur. System complexity is defined as the average number of information bits needed to describe the output, per input bit. System stability is the discrepancy between the average number of prediction errors and average number of margin errors, on two input sequences. A bound on the prediction error is derived and used as a system's performance guarantee. All three concepts are interdependent: As complexity increases, stability decreases and performance guarantee becomes more sensitive to changes in the input therefore less robust.

Keywords: Markov chain prediction, system complexity, stability, performance guarantee

1. Overview

How are complexity and stability of a system related? It is known, for instance, in the study of ecological systems [9] that more complex systems tend to be less stable. Is this true in general? In this paper we undertake a probabilistic analysis for deterministic systems that act in a random environment. A system acts as a deterministic switch, whose action is dictated by a binary decision rule. Based on past input from the environment, it predicts the environment's future state. Our results indicate that system complexity and stability are oppositely related. As a system becomes more complex, its stability decreases. While our analysis is for prediction systems, because the underlying action is that of a deterministic switch which selects and copies input to output based on any binary decision rule, then the analysis can be extended, in general, to other kinds of systems. We use samples for the sake of defining complexity, stability and performance guarantee. The paper is not on the subject of statistical learning. Our results apply to systems that can be derived by any means, such as by system design, through a search, or by inference from training samples.

The paper is organized as follows: In Section 2 we start with an introduction. In Section 3 we set up the notation, definitions and assumptions. In Section 4, we define and estimate system complexity. In Section 5 we define a notion of system stability based on a significance test and state a result on its critical value. In Section 6 we state a result on system's performance guarantee, followed by the conclusions in Section 7. For easier reading, the proofs of the results are deferred to the appendix.

2. Introduction

Let $\mathbb{Z}$ denote the set of all integers. Let an environment be modeled as a stationary binary Markov chain

$$X := \{X_t : t \in \mathbb{Z}\},$$

where $X_t \in \{-1, 1\}$, and consider a system which acts as a digital switch (or gate) on this chain. Its aim is to produce an output sequence Y from this random input X such that every bit of the output sequence consists only of -1. It does so by predicting at time $t-1$ the value of the next input bit at time t. If the prediction is -1 with sufficient confidence then it simply selects the input bit at time t, otherwise if the prediction is 1 with sufficient confidence then it selects the input bit at time t and inverts its value. It then places it as the output bit at time t. The decision of whether to select or not to select an input is based on a margin function $f(t)$, whose value is positive if the system predicts 1 for X_t and negative if it predicts -1 for X_t. The larger the absolute value $|f(t)|$ the higher the confidence in prediction hence we refer to $|f(t)|$ as the confidence function.

Let us describe this action in more details. Denote by $\mathrm{sgn}(a)$ the sign function which equals 1 if $a \geq 0$, and -1 if $a < 0$. Let $a > 0$ be any fixed positive real value. Let $h(t) := \mathrm{sgn}(f(t))$ be the system's binary prediction for X_t. At every time t, if the system predicts -1 for input X_t, which we write as $h(t) = -1$, with confidence level at least a, that is, $f(t) \leq -a$, then it selects X_t. It then copies X_t to be the output bit $Y_t = X_t$. If the system predicts the input X_t to be 1, that is $h(t) = 1$ where $f(t) \geq a$, then it selects X_t, inverts its value to obtain $\overline{X}_t$ and copies it to be output at time t, $Y_t = \overline{X}_t$. Otherwise, $|f(t)| < a$ and the

system does not produce an output value Y_t to avoid making a prediction error. We refer to a as confidence threshold of the system and denote a prediction system by (f, a). Denote by

$$\nu := |\{t : |f(t)| \geq a\}|$$

the number of times that the confidence function exceeds the threshold a, namely, the number of times that a system predicts. The output sequence is

$$Y := \{Y_{t_l}\}_{l=1}^{\nu}.$$

Denote by $\Psi_t := \mathbb{I}\{h(t) \neq X_t\}$ an indicator of the event that at time t the prediction is incorrect. The sequence

$$\Psi := \{\Psi_t : |f(t)| \geq a\}$$

is a prediction error sequence. The following example depicts the above behavior with $1 \leq t \leq m$, $m = 12$. We denote by $>_a$, $<_{-a}$ and $\diamondsuit_a$ the event that $f(t) \geq a$, $f(t) \leq -a$, and $|f(t)| < a$, respectively. We write $+$ for 1 and $-$ for -1,

Time:	t	1	2	3	4	5	6	7	8	9	10	11	12
Input:	X_t	$+$	$-$	$+$	$+$	$-$	$-$	$+$	$-$	$+$	$-$	$-$	$+$
Confidence:	$f(t)$	$>_a$	$>_a$	$\diamondsuit_a$	$<_{-a}$	$>_a$	$\diamondsuit_a$	$\diamondsuit_a$	$>_a$	$>_a$	$>_a$	$<_{-a}$	$>_a$
Prediction:	$h(t)$	$+$	$+$		$-$	$+$			$+$	$+$	$+$	$-$	$+$
Error:	Ψ_t	0	1		1	1			1	0	1	0	0
Output:	Y_t	$-$	$+$		$+$	$+$			$+$	$-$	$+$	$-$	$-$

As can be seen, the output Y is a subsequence of the input X or its complement $\overline{X}$ as follows: at times t when $f(t) \leq -a$ it is a subsequence of X, and at times t when $f(t) \geq a$ it is a subsequence of $\overline{X}$. Since not every bit of the output sequence is -1 it means that the system is not completely successful in meeting its aim.

Let us define a function $\phi(y) := (y+1)/2$ and apply it to the output Y to form a prediction error sequence,

$$\Xi := \{\phi(Y_t)\}_{t=1}^{\nu}.$$

The last row in the next table depicts this sequence with $\nu = 9$,

Time:	t	1	2	3	4	5	6	7	8	9	10	11	12
Input:	X_t	$+$	$-$	$+$	$+$	$-$	$-$	$+$	$-$	$+$	$-$	$-$	$+$
Confidence:	$f(t)$	$>_a$	$>_a$	$\diamondsuit_a$	$<_{-a}$	$>_a$	$\diamondsuit_a$	$\diamondsuit_a$	$>_a$	$>_a$	$>_a$	$<_{-a}$	$>_a$
Prediction:	$h(t)$	$+$	$+$		$-$	$+$			$+$	$+$	$+$	$-$	$+$
Error:	Ψ_t	0	1		1	1			1	0	1	0	0
Output:	Y_t	$-$	$+$		$+$	$+$			$+$	$-$	$+$	$-$	$-$
$\phi(Y)$:	Ξ_t	0	1		1	1			1	0	1	0	0

As can be seen above, the two sequences are equal, $\Xi = \Psi$, hence the error sequence Ψ can be obtained directly from the output sequence Y by evaluating the function $\phi(Y)$.

We can view the event $\Xi_t = 1$ as a penalty of making a false prediction. We now generalize this notion. Let $b > 0$ be a margin penalty parameter. Define a margin error event at time t as the event that $X_t f(t) < b$ (while the notion of margin is used in statistical learning theory [1], here we use it in a more general context not for learning). This event occurs if the sign of the input differs from the sign of the margin function (which means that a prediction error $h(t) \neq X_t$ occurs) or if their sign is the same but the absolute value of f is less than b. We can now generalize the notion of penalty and define an indicator variable $\Xi_t^{(b)}$ to take a value 1 if $X_t f(t) < b$, and zero otherwise. We define

$$\Xi^{(b)} := \left\{\Xi_t^{(b)}\right\}_{t=1}^{m}$$

a sequence of indicators of margin errors at every time $1 \leq t \leq m$ (in contrast to the sequence Ξ which is defined only at time instants t where $|f(t)| \geq a$). This margin error sequence has at least the number of 1 bits as the sequence Ξ since Ξ is equivalent to a subsequence of a margin error sequence with $b = 0$. $\Xi^{(b)}$ is a result of a more strict assessment of the system's prediction since we penalize the system even if it correctly predicts the value of X yet with an insufficient level b of confidence. We are free to choose the value of b independent of the system (f, a) and thereby control the level of assessment. The larger we set b the more strict the assessment. The next table depicts $\Xi^{(b)}$ where $>_b$, $<_{-b}$ and $\diamondsuit_b$ denote the event that $f(t) \geq b$, $f(t) \leq -b$, and $|f(t)| < b$, respectively.

Input:	X_t	$+$	$-$	$+$	$+$	$-$	$-$	$+$	$-$	$+$	$-$	$-$	$+$
Confidence:	$f(t)$	$>_a$	$>_a$	$\diamondsuit_a$	$<_{-a}$	$>_a$	$\diamondsuit_a$	$\diamondsuit_a$	$>_a$	$>_a$	$>_a$	$<_{-a}$	$>_a$
		$>_b$	$>_b$	$\diamondsuit_b$	$<_{-b}$	$\diamondsuit_b$	$\diamondsuit_b$	$\diamondsuit_b$	$>_b$	$\diamondsuit_b$	$>_b$	$\diamondsuit_b$	$>_b$
Prediction:	$h(t)$	$+$	$+$		$-$	$+$			$+$	$+$	$+$	$-$	$+$
Prediction error:	Ψ_t	0	1		1	1			1	0	1	0	0
Output:	Y_t	$-$	$+$		$+$	$+$			$+$	$-$	$+$	$-$	$-$
$\phi(Y)$:	Ξ_t	0	1		1	1			1	0	1	0	0
Margin error:	$\Xi_t^{(b)}$	0	1	1	1	1	1	1	1	1	1	1	0

The only time instants where the sequence $\Xi^{(b)}$ has a 0 are when the prediction is correct and the confidence level is at least b.

In this paper we define three properties of a prediction system: complexity, stability and performance guarantee. Complexity is defined as the conditional entropy of Y (conditioned on its length ν) divided by m. It represents the average number of information bits, per input bit, needed to describe the failures of a system in predicting an input sequence of length m drawn randomly from the environment. Stability is defined as follows: we consider the difference between the average of the sequence Ξ based on an input sequence $X^{(n)}$ and the average of the sequence $\Xi^{(b)}$ based on another input sequence $X^{(m)}$. If this difference deviates from its expected value by more than a critical value, then a system is unstable. Performance is defined as the prediction error probability, namely, the probability that $\Xi_t = 1$. We state an upper bound on this probability, which serves as a performance guarantee. We study how these three properties interdepend.

We now proceed with the general setup and definitions.

3. Setup

Let k^*, k, m, n, be positive integers. We denote random variables by capital letters, for instance, X, S, Θ, Y and lower case letters x, s, θ, y to represent their values. We also use capital letters, for instance A, B, R, U, to denote sets. The letters a, b denote functions.

3.1. **Environment.** In this section we present a probabilistic setup that defines the random environment. Let $k^* \geq 1$ and denote by $\{X_t : t \in \mathbb{Z}\}$ a sequence of binary random variables possessing the following Markov property,

$$\begin{aligned} &P\left(X_t = x_t \mid X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, \ldots\right) \\ &\quad = \quad P\left(X_t = x \mid X_{t-1} = x_{t-1}, \ldots, X_{t-k^*} = x_{t-k^*}\right) \end{aligned} \tag{3.1}$$

where $x_{t-k^*}, \ldots, x_{t-1}$, x_t take a binary value of -1 or 1. This sequence is known as a discrete-time Markov stochastic process, or Markov *chain*, of order k^*. Let the *environment* be a stationary homogeneous Markov chain of order k^*. We assume that k^* is unknown.

From the environment, we draw $m + \max\{k, k^*\}$ consecutive values that form a finite Markov chain

$$X^{(m)} \quad := \quad \{X_t\}_{t=-\max\{k,k^*\}+1}^{m}. \tag{3.2}$$

Denote by $\mathbb{S}_{k^*}$ a set of states $s^{*(i)}$, $i = 0, 1, \ldots, 2^{k^*} - 1$, where $s^{*(0)} := [s^{*(0)}_{k^*-1}, \ldots, s^{*(0)}_0] = [-1, \ldots, -1, -1]$, $s^{*(1)} := [s^{*(1)}_{k^*-1}, \ldots, s^{*(1)}_0] = [-1, \ldots, -1, 1]$, $\ldots$,$s^{(2^{k^*}-1)} := [1, \ldots, 1]$. Based on $\mathbb{S}_{k^*}$, the chain $X^{(m)}$ can be represented as a sequence

$$S^{*(m)} = \{S^*_t\}_{t=1}^{m} \tag{3.3}$$

of random states where

$$S^*_t \quad := \quad \left(X_{t-(k^*-1)}, X_{t-(k^*-2)}, \ldots, X_t\right) \in \mathbb{S}_{k^*} \tag{3.4}$$

defines the random state at time t. With respect to $\mathbb{S}_{k^*}$, a state transition occurs from S^*_t to S^*_{t+1} by shifting left the sequence of bits in (3.4), to obtain $S^*_{t+1} := (X_{t-(k^*-2)}, \ldots, X_t, X_{t+1})$. There are two possible transitions that can occur from S^*_t into S^*_{t+1}: a negative transition, where the bit X_{t+1} is -1 and positive transition where X_{t+1} is 1.

We denote by Q the $2^{k^*} \times 2^{k^*}$ transition probability matrix of the Markov chain $\{X_t : t \in \mathbb{Z}\}$ that satisfies (3.1) and its $(ij)^{th}$ entry is denoted by

$$Q[i, j] := q\left(s^{*(j)} \,\middle|\, s^{*(i)}\right). \tag{3.5}$$

We denote the probability of the two possible transitions from a state $s^{*(i)}$ by

$$q(1|s^{*(i)}), \quad q(-1|s^{*(i)}) \tag{3.6}$$

(the remaining transitions from $s^{*(i)}$ have probability zero) and we assume that for all $0 \leq i \leq 2^{k^*} - 1$,

$$0 < q\left(1 \left| s^{*(i)} \right.\right) < 1 \tag{3.7}$$

thus the environment's Markov chain is irreducible. (Another minor assumption about Q is made in Section 6.)

Let

$$\pi^* := \left[\pi_0^*, \ldots, \pi_{2^{k^*}-1}^*\right] \tag{3.8}$$

denote the stationary probability distribution, where π_i^* is the probability that $S_t^* = s^{*(i)}$. We also write $\pi_{s^{*(i)}}^* := \pi_i^*$ for $0 \leq i \leq 2^{k^*} - 1$. That a stationary probability distribution exists follows from the fact that the Markov chain is irreducible and the state space $\mathbb{S}_{k^*}$ is finite (Corollary 8.2, [11]).

When we write $P(A)$ we mean 'the probability of A' without explicitly mentioning the underlying probability distribution with which it is measured.

Let us denote by $\mathbb{P}$ the joint probability distribution of a state sequence $(S_1^*, \ldots, S_l^*)$ defined as follows: for any sequence $(s_1^*, \ldots, s_l^*) \in \mathbb{S}_{k^*}^l$,

$$\mathbb{P}\left((S_1^*, \ldots, S_l^*) = (s_1^*, \ldots, s_l^*)\right) := \pi_{s_1^*}^* \prod_{r=1}^{l-1} p\left(s_{r+1}^* \left| s_r^* \right.\right). \tag{3.9}$$

As mentioned in Section 1, in this paper we consider specific systems that predict the environment. Because k^*, and hence $\mathbb{S}_{k^*}$, are not known, a prediction system is based on a binary function (Section 3.4) which is defined on a set $\mathbb{S}_k$ of states where k may be different from k^*.

Denote by $S^{(m)}$ the sequence of states of $\mathbb{S}_k$ that corresponds to $X^{(m)}$, that is,

$$S^{(m)} = \{S_t\}_{t=1}^m \tag{3.10}$$

and

$$S_t \quad := \quad \left(X_{t-(k-1)}, X_{t-(k-2)}, \ldots, X_t\right) \in \mathbb{S}_k. \tag{3.11}$$

There is a one-to-one correspondence between $S^{(m)}$ and $S^{*(m)}$ because given one of these sequences we can obtain the uniquely corresponding sequence $X^{(m)}$ from which the second sequence is obtained.

Corresponding to the transition matrix Q there is a Markov model which is a directed labeled graph. Its vertices are the states of $\mathbb{S}_k$, the edges are the positive and negative transitions labeled with their corresponding transition probabilities. We use this graph to define a metric on $\mathbb{S}_k$ in Section 3.3.

3.2. **Sequences.** For $1 \leq q \leq r$, define a projection operator $< \cdot >_q \colon \mathbb{S}_r \to \mathbb{S}_q$ as a mapping that takes a state $s^{(i)} \in \mathbb{S}_r$ to a state

$$s =< s^{(i)} >_q = [s_{q-1}^{(i)}, \ldots, s_0^{(i)}] \in \mathbb{S}_q \tag{3.12}$$

whose q bits correspond to the q least significant (rightmost) bits of $s^{(i)}$. We define

$$r(k, k^*) := \begin{cases} 1 & \text{if } k^* \geq k+1 \\ k - k^* + 2 & \text{if } k^* \leq k. \end{cases} \tag{3.13}$$

For any state-sequence $\theta \in \mathbb{S}_{k^*}^{r(k,k^*)}$ consisting of $r(k, k^*)$ states in $\mathbb{S}_{k^*}$ (if $r(k, k^*) = 1$ then θ is a single state of k^* bits) we extend the definition (3.12) and denote by $< \theta >_q$ the q least significant bits of the binary sequence that corresponds to the state sequence θ, and let $< \theta >_i^j$ denote the binary subsequence starting at the i^{th} bit and ending at the j^{th} bit from the right where the rightmost bit has index 0. For example, suppose $k^* = 2$, $k = 5$ and $m = 4$, therefore $-\max\{k, k^*\} + 1 = -4$, and let

$$\begin{aligned} x^{(4)} &= (x_{-4}, x_{-3}, x_{-2}, x_{-1}, x_0, x_1, x_2, x_3, x_4) \\ &= 1, 1, 1, -1, -1, 1, -1, 1, -1 \end{aligned}$$

then at $t = 2$, $\theta_t \in \mathbb{S}_2^5$ (since $k - k^* + 2 = 5$) and

$$\begin{aligned} \theta_t &= (s_{t-4}^*, s_{t-3}^*, s_{t-2}^*, s_{t-1}^*, s_t^*) \\ &= ((1,1), (1,-1), (-1,-1), (-1,1), (1,-1)) \\ &= (s^{*(3)}, s^{*(2)}, s^{*(0)}, s^{*(1)}, s^{*(2)}) \end{aligned}$$

so $< \theta_t >_1^5 = 1, 1, -1, -1, 1$ and $< \theta_t >_0 = -1$. Note that the binary sequence that corresponds to a state subsequence $\theta \in \mathbb{S}_{k^*}^{r(k,k^*)}$ has a length of $r(k,k^*) - 1 + k^* = k + 1$ bits if $k^* \leq k$ or k^* bits if $k^* \geq k + 1$. In either case, θ is equivalent to a binary sequence of length at least $k + 1$.

We denote by

$$\Theta_t := (S^*_{t-r(k,k^*)+1}, \ldots, S^*_{t-2}, S^*_{t-1}, S^*_t) \tag{3.14}$$

a sequence of $r(k,k^*)$ random state variables where Θ_t takes values in $\mathbb{S}_{k^*}^{r(k,k^*)}$ according to the joint probability measure $\mathbb{P}$.

As Θ_t is of length at least $k + 1$ bits, we also use the following equivalent representation,

$$\Theta_t := [X_{t-k}, \ldots X_t]$$

and

$$\Theta_t := [S_{t-1}, X_t] \tag{3.15}$$

where $S_{t-1} \in \mathbb{S}_k$ is the state at time $t - 1$.

3.3. **Metric.** We assume that the state space $\mathbb{S}_k$ on which the class of binary functions is defined, has a metric d that depends only on the possible state transitions, that is, on pairs of states between which there may be a transition and not on the transition probabilities themselves since they depend on the environment's transition probabilities which are assumed to be unknown. One way to define a metric based on such knowledge is to start with an undirected graph $G_k = (V_k, E_k)$ where V_k and E_k represent the vertex and edge sets. The vertices correspond to the states of $\mathbb{S}_k$ and an edge exists between two distinct vertices u and v if it is possible to have either a positive or negative transition from s to s' or from s' to s, where s and s' are the states in $\mathbb{S}_k$ that correspond to u and v. This graph is known as an undirected De Bruijn graph [7] of dimension k, and is denoted by $UB(2,k)$. We refer to it as G_k.

A path $\mathcal{P}(u,v)$ between vertices u and v is a finite sequence of edges that connects u and v. When a path $\mathcal{P}$ exists its length $l(\mathcal{P})$ is the number of edges in the sequence. Let

$$\text{dist}(u,v) := \min \{l(\mathcal{P}) : \mathcal{P}(u,v) \text{ exists}\}$$

be the length of the shortest path between u and v and let $\text{dist}\,(u,v) := \infty$ if there is no path between u and v. We define $\text{dist}(u,u) = 0$ for every $u \in V_k$. It is known that an undirected De Bruijn graph of dimension k is connected, that is, there exists a path between any two vertices. The distance between the farthest pair is k hence

$$\text{diam}\,(\mathbb{S}_k) = k. \tag{3.16}$$

The length of the shortest path between two vertices, namely, $\text{dist}(u,v)$, is a metric on V_k since G_k is a connected graph.

Define the metric between states s and s' as

$$\mathsf{d}(s,s') \quad := \quad \text{dist}(u,v) \tag{3.17}$$

where u, v are the vertices that correspond to s and s'. It follows that d is a metric on $\mathbb{S}_k$. Define the diameter of $\mathbb{S}_k$ as

$$\text{diam}(\mathbb{S}_k) := \max_{s,s' \in \mathbb{S}_k} \mathsf{d}\,(s,s')\,.$$

Next we mention a few additional standard concepts.

A dominating set of a graph G_k is a set of vertices $U \subseteq V_k$ such that for every vertex $v \in V_k \setminus U$ there exists a vertex $u \in U$ such that u and v are connected by an edge. The domination number of G_k is the size of the smallest dominating set. The directed De Bruijn graph, denoted by $B(2,k)$, has an arc instead of an edge from a vertex u to v if there is a transition from u to v. From [4], the domination number of the directed De Bruijn graph $B(2,k)$ is $\lceil 2^k/3 \rceil$. This bounds the domination number of $UB(2,k)$, which we refer to as G_k.

A *γ-cover* of $\mathbb{S}_k$ with respect to the metric d is a set $C \subseteq \mathbb{S}_k$ such that for every element $s \in \mathbb{S}_k$ there exists an $s' \in C$ such $\mathsf{d}(s,s') \leq \gamma$. The size of the smallest γ-cover of $\mathbb{S}_k$ is defined as the *γ-covering number* of $\mathbb{S}_k$ with respect to d, and is denoted by N_γ. Any dominating set of G_k is also a 1-cover of $\mathbb{S}_k$ with respect to the metric (3.17). Therefore the domination number of G_k is an upper bound on the 1-covering number of $\mathbb{S}_k$. We are not aware of existing results which bound the γ-covering number of G_k for general γ but a simple greedy algorithm for set covering yields an upper bound on the covering number of G_k which is accurate to within a factor of $k + 1$ (see [2], Section 6) .

3.4. **Classifier.** In Section 1 we discussed prediction systems in general. In this section we consider classification, or decision rules, which subsequently are used to form prediction systems. We start by defining the class $\mathcal{H}$ of all binary functions $h : \mathbb{S}_k \to \{-1, 1\}$. We henceforth refer to h as a *classifier* since it classifies each state s into one of two categories. For a subset $R \subseteq \mathbb{S}_k$ let

$$\text{dist}(s, R) := \min_{s' \in R} \mathsf{d}(s, s').$$

From [2], we define the *width* of h at s as

$$w_h(s) := \text{dist}\left(s, R_{\overline{h}(s)}\right) \tag{3.18}$$

where R_+, $R_- \subseteq \mathbb{S}_k$ are regions classified as 1 and -1 by h, , respectively, and $\overline{h}(s)$ is the complement of $h(s)$. For every $s \in \mathbb{S}_k$, we have $s \in R_{h(s)}$, hence $s \notin R_{\overline{h}(s)}$ and therefore $w_h(s) > 0$. Define

$$f_h : \mathbb{S}_k \to \{-\text{diam}\,(\mathbb{S}_k), \ldots, -1, 1, \ldots, \text{diam}\,(\mathbb{S}_k)\}$$

by

$$f_h(s) := h(s) w_h(s) \tag{3.19}$$

to be a *signed width function* associated with h. We refer to this function also as *margin* function. The decision of h at s can be expressed in terms of the margin as follows, $h(s) = \text{sgn}\,(f_h(s))$. Thus f_h not only contains the binary decision information of h but its value gives a form of confidence in the decision of h. We use this further below to define a notion of confidence and assign a penalty for making a low-confidence decision. Note that we can evaluate the width and margin functions because k is known and thus the edges of the De Bruijn graph on $\mathbb{S}_k$ are known (the environment's space $\mathbb{S}_{k^*}$ and its corresponding transition matrix Q, both of which are assumed to be unknown, are not needed for this evaluation).

Using the notation of (3.15), for $s \in \mathbb{S}_k$ and $x \in \{-1, 1\}$, let

$$\theta := [s, x] \tag{3.20}$$

and for any $h \in \mathcal{H}$ define the *margin* of h at s as

$$x f_h(s). \tag{3.21}$$

Consider $\theta_t := [s_{t-1}, x_t]$ at time t. We use h to define a decision for x_t based on state s_{t-1} using the following decision rule:

DECISION RULE: If $h(s_{t-1}) = 1$ then decide 1 for x_t, else decide -1.

Note that if $h(s_{t-1}) \neq x_t$ then $x_t f_h(s_{t-1})$ is negative, and vice versa, so a *decision error event* at time t is expressed as either one of the following two equivalent events

$$h(s_{t-1}) \neq x_t \iff x_t f_h(s_{t-1}) < 0. \tag{3.22}$$

If $h(s_{t-1}) = x_t$ then $x_t f_h(s_{t-1})$ is non-negative. If $x_t f_h(s_{t-1}) < \gamma$ then either h decides the wrong value for x_t or the margin value $f_h(s_{t-1})$ is lower than γ. We use this further below to define margin error.

3.5. **Margin.** Denote by $\gamma \in (0, \text{diam}(\mathbb{S}_k)]$ a parameter and let

$$b := b(\gamma) > 0 \tag{3.23}$$

where $b(0) = 0$, be a non-decreasing function which is called the *margin penalty* whose value is used to define the margin error.

Definition 1. (*Margin error*) For $\Theta_t = (S_{t-1}, X_t)$, let the event

$$X_t f_h(S_{t-1}) < b(\gamma) \tag{3.24}$$

be defined as a *margin error* of classifier h at time t.

When $b(\gamma) = 0$, (3.24) is the same as decision error event (3.22). Note that (3.24) is true if the sign of $f_h(S_{t-1})$ differs from the sign of X_t, that is, if h decides the wrong value for X_t based on the state S_{t-1} at time $t-1$, or if it decides the correct value but the margin is lower than $b(\gamma)$. The event (3.24) may also be expressed in terms of Θ_t as follows,

$$< \Theta_t >_0 f_h\left(< \Theta_t >_1^k\right) < b(\gamma). \tag{3.25}$$

Definition 2. (*Margin error sequence*) Based on $X^{(m)}$ define the *margin error sequence* as the following sequence of margin error indicators

$$\Psi^{(m,\gamma)}(h) := \left\{\Psi_t^{(m,\gamma)}(h)\right\}_{t=1}^{m} = \left\{\mathbb{I}\left\{X_t f_h(S_{t-1}) < b(\gamma)\right\}\right\}_{t=1}^{m}. \tag{3.26}$$

Definition 3. (*Average margin error*) The average number of times that a margin error occurs on $X^{(m)}$ by classifier h is defined as the *average margin error*

$$L_m^{(b(\gamma))}(h) := \frac{1}{m}\sum_{t=1}^{m}\Psi_t^{(m,\gamma)}(h). \tag{3.27}$$

Its expected value, with respect to the probability distribution of the chain $X^{(m)}$, is denoted by

$$\begin{aligned} L^{(b(\gamma))}(h) &:= \mathbb{E}\left[L_m^{(b(\gamma))}(h)\right] \\ &= \mathbb{P}\left(X_t f_h(S_{t-1}) < b(\gamma)\right). \end{aligned} \tag{3.28}$$

3.6. **System.** In the previous sections we introduced several definitions that involve a classifier based on a binary function h. In this section we define a *prediction system* to be a system that uses classifier h to predict only when there is a sufficiently high confidence. We declare the value of the width function w_h at a state s to be the confidence of prediction at that state. Denote by $\gamma \in (0, \operatorname{diam}(\mathbb{S}_k)]$ a parameter and let

$$a := a(\gamma) \tag{3.29}$$

denote a non-decreasing function whose value represents decision confidence threshold and where $a(0) = 0$. We now use $a(\gamma)$ to define a prediction system.

Definition 4. (*Prediction system*) Given $h \in \mathcal{H}$ and $\gamma \in (0, \operatorname{diam}(\mathcal{X})]$ denote by (h, γ) a *prediction system* which is based on the following prediction rule for predicting the value of X_t: if $f_h(S_{t-1}) \geq a(\gamma)$ then decide 1, otherwise if $f_h(S_{t-1}) \leq -a(\gamma)$ decide -1, otherwise reject making a decision. We refer to $w_h := |f_h|$ as the *confidence function* of the system and to $a(\gamma)$ as the decision *confidence threshold* of the system.

From (3.29), if $\gamma = 0$ then $a(\gamma) = 0$ and in this case the prediction system always makes a decision since the confidence function is non-negative. Henceforth, by a *system* we mean a prediction system.

Definition 5. (*System output*) The *output* $Y_t \in \{-1, 1\}$ of a prediction system (h, γ) at time t is defined as follows:

$$Y_t := \begin{cases} X_t & \text{if } f_h(S_{t-1}) \leq -a(\gamma) \\ \overline{X}_t & \text{if } f_h(S_{t-1}) \geq a(\gamma) \\ \text{null} & \text{otherwise} \end{cases}$$

where

$$\overline{X} := \begin{cases} 1 & \text{if } X = -1 \\ -1 & \text{if } X = 1 \end{cases}$$

and null means there is no output value at t.

Definition 6. (*Aim of system*) The *aim* of a prediction system is to output -1 at time t if at time $t-1$ it predicts with confidence at least $a(\gamma)$, otherwise not to output any value. That is, a system aims to output only the value -1 at all times in which it makes a prediction and no output at any other time.

So at time $t-1$, if a system is confident that X_t will be -1 then at time t it will copy the value X_t to its output. If at time $t-1$, it is confident that X_t will be 1 then at time t it will copy the complement value of $\overline{X}_t$ to the output. (This aim of outputting only -1 extends what is referred to in [12] as matching the environment.) Definition 6 implies that we can represent the action of a prediction system as copying the input (or its complement) to the output, as described in Definition 5. This fact can be used to apply the analysis and results of the paper to other systems which perform a subsequence selection operation to produce output from input [5, 12–16].

From Definition 5 it follows that whenever a system predicts correctly it outputs $Y_t = -1$, otherwise if it predicts incorrectly then it outputs $Y_t = 1$. Hence the probability that a system (h, γ) wrongly predicts X_t equals

$$\mathbb{P}\left(Y_t = 1 \middle| |f_h(S_{t-1})| \geq a(\gamma)\right) \tag{3.30}$$

which can also be expressed as

$$\mathbb{P}\left(X_t f_h\left(S_{t-1}\right)<0\middle||f_h\left(S_{t-1}\right)|\geq a(\gamma)\right). \tag{3.31}$$

We henceforth refer to (3.31) as *probability of prediction error* and denote it by

$$L(h\,|\gamma\,):=L\left(h|a(\gamma)\right). \tag{3.32}$$

In Section 3.1, we defined a random sequence $X^{(m)}$ of the environment. After obtaining a $X^{(m)}$ from the stationary environment, at a future time we draw an additional $n+\max\{k,k^*\}$ consecutive bits to obtain a second sequence

$$X^{(n)} \quad := \quad \{X_t\}_{t=-\max\{k,k^*\}+1}^{n}. \tag{3.33}$$

For input $X^{(n)}$, denote by

$$\nu:=\nu^{(a)} \tag{3.34}$$

the number of times that the sequence $S^{(n)}$ enters a state $s\in\mathbb{S}_k$ such that the width satisfies $w_h(s)\geq a$ (or equivalently, $|f_h(s)|\geq a$). The system *output sequence* is denoted by

$$Y^{(\nu)}:=\{Y_{t_l}\}_{l=1}^{\nu^{(a)}}, \tag{3.35}$$

where t_l, $1\leq l\leq\nu^{(a)}$, are time instants when the width function satisfies $w_h\left(S_{t_l-1}\right)\geq a$.

Definition 7. (*Error sequence*) Let system (h,γ) predict an input sequence $X^{(n)}$. We denote by *error sequence* the sequence of prediction error indicators when the confidence function satisfies $|f_h(s)|\geq a$,

$$\Psi^{(\nu^{(a)})}(h):=\{\Psi_{t_l}\}_{l=1}^{\nu^{(a)}}=\left\{\mathbb{I}\{X_{t_l}f_h(S_{t_l-1})<0\}\right\}_{l=1}^{\nu^{(a)}}. \tag{3.36}$$

Definition 8. (*Average prediction error*) For input $X^{(n)}$, the *average number of prediction errors* conditioned on having a confidence of at least $a(\gamma)$, is defined as

$$\mathfrak{L}_{\nu}^{(n,\gamma)}(h):=\frac{1}{\nu}\sum_{l:\left|f_h(S_{t_l-1})\right|\geq a(\gamma)}\Psi_{t_l}, \tag{3.37}$$

where ν is defined in (3.34).

For $y\in\{-1,1\}$ let us define the function

$$\varphi\left(y\right):=\frac{y+1}{2}. \tag{3.38}$$

Consider a system (h,γ) and its error sequence $\Psi^{(\nu^{(a)})}(h)$. Recall from above that a system's output equals -1 when the system predicts correctly and 1 when it predicts incorrectly. Therefore, from (3.36), for every $1\leq l\leq\nu^{(a(\gamma))}$, we have

$$\Psi_{t_l}=\varphi\left(Y_{t_l}\right). \tag{3.39}$$

Note that $\{\Psi_t\}_{t=1}^{\nu^{(a)}}$ is a zero-one representation of the output sequence (3.35), namely, the output sequence $Y^{(\nu)}$ contains all the information about a system's prediction errors.

3.7. **Discrepancy.** A notion of system stability is presented in Section 5. It is based on the following function of $X^{(m)}$ and $X^{(n)}$ which we denote as the *discrepancy* of system (h,γ),

$$\Upsilon_{m,n}(h,\gamma):=\mathfrak{L}_{\nu^{(a)}}^{(n,\gamma/2)}\left(h\right)-L_m^{(b(2\gamma))}\left(h\right). \tag{3.40}$$

Discrepancy measures the difference in performance between two systems that have the same classifier h, where one is based on a positive decision confidence threshold $a(\gamma)>0$ and whose performance is measured on a future input sequence $X^{(n)}$, and the other is based on a zero decision confidence threshold with performance based on a past input sequence $X^{(m)}$. In Section 5 we use the discrepancy to define a notion of system stability.

Define the function $\chi_t^{(h,\gamma)}:\{s\in\mathbb{S}_k:|f_h\left(s\right)|\geq a(\gamma)\}\to\{0,1\}$ as follows,

$$\chi_t^{(h,\gamma)}:=\begin{cases}1, & \text{if } X_t f_h\left(S_{t-1}\right)<0 \text{ given that } \left|f_h\left(S_{t-1}\right)\right|\geq a(\gamma)\\ 0, & \text{if } X_t f_h\left(S_{t-1}\right)\geq 0 \text{ given that } \left|f_h\left(S_{t-1}\right)\right|\geq a(\gamma).\end{cases}$$

We have

$$\mathbb{E}\left[\chi_t^{(h,\gamma)}\right] = \mathbb{P}\left(X_t f_h\left(S_{t-1}\right) < 0 \middle| \left|f_h\left(S_{t-1}\right)\right| \geq a(\gamma)\right). \tag{3.41}$$

With (3.31) and (3.32) we have

$$\mathbb{E}\left[\chi_t^{(h,\gamma)}\right] = L(h|\gamma). \tag{3.42}$$

Note that $L\left(h\,|\gamma\right)$ is constant with respect to t because the Markov chain is stationary (Section 3.1). Fix any $h \in \mathcal{H}$ and $\gamma \in (0, \mathrm{diam}(\mathbb{S}_k)]$. The expected value of the average prediction error is

$$\begin{aligned}
\mathbb{E}\left[\mathfrak{L}_\nu^{(n,\gamma)}\left(h\right)\right] &= \mathbb{E}\left[\frac{1}{\nu}\sum_{l:\left|f_h\left(S_{t_l-1}\right)\right|\geq a(\gamma)} \mathbb{I}\left\{X_{t_l} f_h\left(S_{t_l-1}\right) < 0\right\}\right] \\
&= \mathbb{E}_\nu\left[\mathbb{E}\left[\frac{1}{\nu}\sum_{l:\left|f_h\left(S_{t_l-1}\right)\right|\geq a(\gamma)} \mathbb{I}\left\{X_{t_l} f_h\left(S_{t_l-1}\right) < 0\right\}\middle|\nu\right]\right] \\
&= \mathbb{E}_\nu\left[\frac{1}{\nu}\mathbb{E}\left[\sum_{l:\left|f_h\left(S_{t_l-1}\right)\right|\geq a(\gamma)} \chi_{t_l}^{(h,\gamma)}\middle|\nu\right]\right] \\
&= \mathbb{E}_\nu\left[\frac{1}{\nu}\nu L\left(h\,|\gamma\right)\right] \\
&= L\left(h\,|\gamma\right).
\end{aligned} \tag{3.43}$$

3.8. **Admissible.** Let $\gamma = 0$ and consider a prediction system $(h, 0)$. This system has a confidence threshold $a(0) = 0$. Define the probability of *false* prediction at time t by $(h, 0)$ as

$$p_0 := P\left(h\left(S_{t-1}\right) = -1, X_t = 1\right) + P\left(h\left(S_{t-1}\right) = 1, X_t = -1\right),$$

and the probability of *correct* prediction at time t by system $(h, 0)$ as

$$q_0 := P\left(h\left(S_{t-1}\right) = -1, X_t = -1\right) + P\left(h\left(S_{t-1}\right) = 1, X_t = 1\right).$$

Consider a system (h, γ) with $\gamma > 0$ which has a confidence threshold $|f_h(S_{t-1})| \geq a(\gamma)$ where a is any non-decreasing function with $a(0) = 0$. Denote the probability of *false* prediction at time t by (h, γ) as

$$p_a := p_{a(\gamma)} = P\left(f_h\left(S_{t-1}\right) \leq -a(\gamma), X_t = 1\right) + P\left(f_h\left(S_{t-1}\right) \geq a(\gamma), X_t = -1\right)$$

and the probability of *correct* prediction at time t by system (h, γ) as

$$q_a := q_{a(\gamma)} = P\left(f_h\left(S_{t-1}\right) \leq -a(\gamma), X_t = -1\right) + P\left(f_h\left(S_{t-1}\right) \geq a(\gamma), X_t = 1\right).$$

Proposition 9. *Let $a(\gamma)$, $b(\gamma)$ be any non-decreasing functions that take the value 0 at $\gamma = 0$. For a prediction system (h, γ), with confidence threshold $a(\gamma)$, where $h \in \mathcal{H}$, $\gamma \in (0, diam(\mathbb{S}_k)]$, if the following inequality is satisfied*

$$\frac{q_a}{p_a} \geq \frac{q_0}{p_0}, \tag{3.44}$$

then

$$L\left(h\,|a(\gamma)\right) \leq L^{(b(\gamma))}\left(h\right).$$

Proof. Define the sets $A := \{\theta = [s, x] : |f_h(s)| \geq a(\gamma)\}$, $A_0 := \{\theta = [s, x] : |f_h(s)| \geq 0\}$. The set $A_0 = \Omega$, where Ω denotes the sample space which consists of all possible $\theta = [s, x]$, $s \in \mathbb{S}_k$, $x \in \{-1, 1\}$, since the inequality $|f_h(s)| \geq 0$ is true. Define the set $B_0 := \{\theta : x f_h(s) < 0\}$ and $B_\gamma := \{\theta : x f_h(s) < b(\gamma)\}$. From (3.41) and (3.42) we have,

$$\begin{aligned}
L\left(h\,|a(\gamma)\right) &= \mathbb{P}\left(X_t f_h\left(S_{t-1}\right) < 0 \middle| \left|f_h\left(S_{t-1}\right)\right| \geq a(\gamma)\right) \\
&= P\left(B_0 \middle| A\right).
\end{aligned}$$

From (3.28) we have

$$\begin{aligned} L^{(b(\gamma))}(h) &= \mathbb{P}\left(X_t f_h(S_{t-1}) < b(\gamma)\right) \\ &= P\left(B_\gamma\right) \\ &= P\left(B_\gamma \middle| A_0\right) \end{aligned} \tag{3.45}$$

where (3.45) follows from the fact that $A_0 = \Omega$. We have,

$$P\left(B_0 \middle| A_0\right) \leq P\left(B_\gamma \middle| A_0\right)$$

since $b(\gamma) \geq 0$. It suffices to show that if (3.44) is satisfied then

$$P\left(B_0 \middle| A\right) \leq P\left(B_0 \middle| A_0\right). \tag{3.46}$$

For brevity, we write $a := a(\gamma)$. We have

$$\begin{aligned} A \bigcap B_0 &= \{\theta : |f_h(s)| \geq a, x f_h(s) < 0\} \\ &= \{\theta : f_h(s) \leq -a, x = 1\} \bigcup \{\theta : f_h(s) \geq a, x = -1\}. \end{aligned} \tag{3.47}$$

Denote by $A^- := \{\theta = [s,x] : f_h(s) \leq -a\}$ and $A^+ := \{\theta = [s,x] : f_h(s) \geq a\}$. Thus,

$$P\left(A, B_0\right) = P\left(A^-, X = 1\right) + P\left(A^+, X = -1\right).$$

Denote by $A_0^- := \{\theta = [s,x] : f_h(s) < 0\}$ and $A_0^+ := \{\theta = [s,x] : f_h(s) \geq 0\}$. Substituting 0 for a in (3.47) and A_0 for A yields

$$P\left(A_0, B_0\right) = P\left(A_0^-, X = 1\right) + P\left(A_0^+, X = -1\right).$$

The inequality (3.46) can be written as

$$\frac{P\left(A, B_0\right)}{P\left(A_0, B_0\right)} \leq \frac{P(A)}{P\left(A_0\right)} \tag{3.48}$$

and after substituting, it is expressed as

$$\frac{P\left(A^-, X = 1\right) + P\left(A^+, X = -1\right)}{P\left(A_0^-, X = 1\right) + P\left(A_0^+, X = -1\right)} \leq \frac{P(A)}{P(A_0)}. \tag{3.49}$$

We have

$$P(A) = P\left(A^-, X = 1\right) + P\left(A^+, X = 1\right) + P\left(A^-, X = -1\right) + P\left(A^+, X = -1\right)$$

and similarly for $P(A_0)$. Thus, (3.49) is expressed as

$$\begin{aligned} &\frac{P\left(A^-, X = 1\right) + P\left(A^+, X = -1\right)}{P\left(A_0^-, X = 1\right) + P\left(A_0^+, X = -1\right)} \\ &\leq \frac{P\left(A^-, X = 1\right) + P\left(A^+, X = 1\right) + P\left(A^-, X = -1\right) + P\left(A^+, X = -1\right)}{P\left(A_0^-, X = 1\right) + P\left(A_0^+, X = 1\right) + P\left(A_0^-, X = -1\right) + P\left(A_0^+, X = -1\right)}. \end{aligned}$$

This is rearranged into the inequality,

$$1 \leq \frac{1 + \frac{P\left(A^-, X=-1\right) + P\left(A^+, X=1\right)}{P(A^-, X=1) + P(A^+, X=-1)}}{1 + \frac{P\left(A_0^-, X=-1\right) + P\left(A_0^+, X=1\right)}{P\left(A_0^-, X=1\right) + P\left(A_0^+, X=-1\right)}}$$

and then expressed as the inequality,

$$\frac{P\left(A^-, X = -1\right) + P\left(A^+, X = 1\right)}{P\left(A^-, X = 1\right) + P\left(A^+, X = -1\right)} \geq \frac{P\left(A_0^-, X = -1\right) + P\left(A_0^+, X = 1\right)}{P\left(A_0^-, X = 1\right) + P\left(A_0^+, X = -1\right)}.$$

The latter is precisely the inequality

$$\frac{q_a}{p_a} \geq \frac{q_0}{p_0},$$

which is satisfied by the premise of the Proposition. Therefore it follows that (3.49), (3.48) and therefore (3.46) holds. Therefore the statement of the proposition holds. □

The Proposition holds for any functions $a(\gamma)$ and $b(\gamma)$, where $\gamma \in (0, \mathrm{diam}(\mathbb{S}_k)]$, as defined in (3.29) and (3.23). In particular, it holds for $a(\gamma/2)$ and $b(2\gamma)$. Therefore from (3.28), (3.43) and Proposition 9, if

$$\frac{q_{a(\gamma/2)}}{p_{a(\gamma/2)}} \geq \frac{q_0}{p_0}$$

then

$$L\left(h\,|\gamma/2\right) := L\left(h\,|a(\gamma/2)\right) \leq L^{(b(2\gamma))}\left(h\right).$$

This means that the expected discrepancy is non-positive since

$$\begin{aligned} \mathbb{E}\left[\Upsilon_{m,n}(h,\gamma)\right] &:= \mathbb{E}\left[\mathfrak{L}_{\nu}^{(n,\gamma/2)}\left(h\right)\right] - \mathbb{E}\left[L_{m}^{(b(2\gamma))}\left(h\right)\right] \\ &= L(h\,|\gamma/2) - L^{(b(2\gamma))}(h) \\ &\leq 0. \end{aligned} \tag{3.50}$$

This fact is used in the definition of system's stability (Section 5). We denote the class of *admissible* prediction systems by

$$\mathcal{A} \subset \mathcal{H} \times (0, \mathrm{diam}\left(\mathbb{S}_k\right)]$$

and define it as the following collection

$$\mathcal{A} := \left\{(h,\gamma) : (h,\gamma) \text{ satisfies (3.44)}\right\}.$$

3.9. **Assumption.** We need to state an assumption on the Markov environment. In Section A.1 we show that there exists a finite integer $\mathfrak{l}_0$, such that for $l \geq \mathfrak{l}_0$, the transition matrix Q in (3.5) satisfies $Q^l > 0$, that is, every entry of Q^l, denoted by $p^{(l)}(s^{(j)}|s^{(i)})$, is positive. We henceforth choose

$$\mathfrak{l}_0 := \min\{l : Q^l > 0\} \tag{3.51}$$

and in theory, if Q was known then $\mathfrak{l}_0$ can be evaluated by computing Q^l for a sequence of $l \geq 1$ until the first l is found such that $Q^l > 0$. Denote by μ_0 the minimum entry of $Q^{\mathfrak{l}_0}$,

$$\mu_0 := \min_{i,j} p^{(\mathfrak{l}_0)}(j|i) \tag{3.52}$$

then the fact that $Q^{\mathfrak{l}_0} > 0$ implies that $\mu_0 > 0$.

We henceforth make the following assumption:

Assumption 1. *The environment's transition matrix Q satisfies one of the following conditions: (i) the minimum entry of $Q^{\mathfrak{l}_0}$ is $\mu_0 \neq 2^{-k^*}$ or (ii) $\mu_0 = 2^{-k^*}$ and for all $0 \leq i \leq 2^{k^*} - 1$, the transitions probabilities (3.6) are $\frac{1}{2}$.*

Remark 10. In both parts (i) and (ii) of the above assumption, Q may have a uniform stationary distribution $\pi^T = \left[2^{-k^*}, \ldots, 2^{-k^*}\right]$, which means Q is doubly stochastic and $\lim_{l\to\infty} Q^l$ is a matrix U, of the same size as Q, with all its entries identical to 2^{-k^*}. Part (ii) treats the special case where this limit U is reached exactly at time $\mathfrak{l}_0$, that is, $Q^{\mathfrak{l}_0} = U$.

We use $\mathfrak{l}_0$ and μ_0 in the following definition. According to the cases of Assumption 1, define

$$\rho(k^*, \mathfrak{l}_0) := \begin{cases} \frac{1-2^{k^*}\mu_0}{2\mu_0} & \text{if case } (i) \text{ holds and } \mathfrak{l}_0 = 1 \\ \frac{2^{k^*}-1}{\left(1-2^{k^*}\mu_0\right)^{(\mathfrak{l}_0-1)/\mathfrak{l}_0}\left(1-\left(1-2^{k^*}\mu_0\right)^{1/\mathfrak{l}_0}\right)} & \text{if case } (i) \text{ holds and } \mathfrak{l}_0 \geq 2 \\ 2^{k^*-1} & \text{if case } (ii) \text{ holds.} \end{cases} \tag{3.53}$$

In the first condition of (3.53), (3.7) implies that $Q > 0$ and thus μ_0 is the minimum entry of Q.

3.10. **Cover.** In this section we provide some estimates on the covering number of the class $\mathcal{F}$ of signed width functions. This is used in an upper bound on the performance of a prediction system (Section 6).

Denote the l_∞ -norm of f_h by

$$\|f_h\| := \max_{s\in\mathbb{S}_k} |f_h(s)|. \tag{3.54}$$

Denote by

$$\mathcal{F} := \left\{f_h : h \in \mathcal{H}\right\}. \tag{3.55}$$

An α-cover of $\mathcal{F}$ with respect to the l_∞ norm on $\mathbb{S}_k$ is a set $\hat{F}_\alpha := \left\{f_j^{(\alpha)}\right\}_{j=1}^{r}$ such that for every element $f \in \mathcal{F}$ there exists an $f_j^{(\alpha)} \in \hat{F}_\alpha$ such

$$\left\|f - f_j^{(\alpha)}\right\| \leq \alpha. \tag{3.56}$$

We denote by

$$h_j := \operatorname{sgn}\left(f_j^{(\alpha)}\right) \tag{3.57}$$

the binary function that corresponds to $f_j^{(\alpha)}$ (note that $j := j(\alpha)$ and we omit the dependence on α for brevity). The size r of the smallest α-cover of $\mathcal{F}$ is defined as the α-covering number of $\mathcal{F}$ with respect to l_∞ norm on $\mathbb{S}_k$ and is denoted by $\mathcal{N}_\alpha$.

From [2], Section 3.2, it follows that

$$\mathcal{N}_\alpha \leq \left(2\left\lceil\frac{3\operatorname{diam}(\mathbb{S}_k)}{\alpha}\right\rceil + 1\right)^{N_{\alpha/3}} \tag{3.58}$$

where N_α is the α-covering number of $\mathbb{S}_k$ with respect to the metric d defined in (3.17). In the next section, we introduce a notion of system complexity.

4. Complexity

In [17], system complexity is defined as the uncertainty that a system meets its functional requirements. We take a similar approach and define complexity of a system to be the uncertainty that it fails to predict. We define this as the average expected description length of the sequence of indicators for the errors made by a prediction system. This definition of system complexity not only depends on the system but also on the probabilistic properties of its environment (via the expected value) since the environment is the source of uncertainty in its ability to meet the functional requirement of predicting the random input.

Recall that by Definition 4, a prediction system makes a decision only when its confidence exceeds a threshold $a(\gamma)$. Because of the relationship (3.39), the output sequence $Y^{(\nu)}$ defined in (3.35) indicates when a system makes a prediction error. Essentially, it is a binary description of a system's failure to predict a random input sequence. Denote by $H(Y^{(\nu)}|\nu)$ the conditional entropy of the sequence $Y^{(\nu)}$ given its length ν. By the property of entropy [6], $H(Y^{(\nu)}|\nu)$ is the minimal expected length of a codeword needed to describe the output of a system given an input sequence of length n, conditioned on knowing the number ν of times that the system produces an output (namely, makes a prediction). Thus it is the number of information bits needed to describe the sequence of errors.

Definition 11. *(Complexity)* Let (h, γ) be a prediction system with confidence function $a(\gamma)$. Let $X^{(n)}$ be an input sequence from the environment and $Y^{(\nu)}$ the corresponding system's output sequence. Let length ν denote the random length of $Y^{(\nu)}$. Define the complexity of (h, γ) as

$$\mathcal{C}(h, \gamma) := \frac{1}{n} H\left(Y^{(\nu)} | \nu\right). \tag{4.1}$$

The complexity of a system is the average number of information bits (minimal expected description length) per input bit, which is needed to describe failures of a system in predicting an input sequence of length n from the environment.

Let us obtain an upper bound on the complexity $\mathcal{C}(h, \gamma)$. For a binary random variable that takes one of its two values with probability $p \in [0, 1]$ and the other with probability $1 - p$, denote its entropy by $H(p) := -p \log p - (1 - p) \log(1 - p)$, where all logs are to base 2. We have,

$$\begin{aligned} H(Y^{(\nu)} | \nu) &= \sum_{l=1}^{n} P(\nu = l)\, H\left(Y^{(\nu)} | \nu = l\right) \\ &\leq \sum_{l=1}^{n} P(\nu = l) \sum_{r=1}^{l} H(Y_{t_r}). \end{aligned} \tag{4.2}$$

From (3.30), (3.31) and (3.32) the binary variable Y_{t_r} has entropy $H(L(h|\gamma))$. Continuing from (4.2), we have

$$\sum_{l=1}^{n} P(\nu = l) \sum_{i=1}^{l} H(Y_{t_r}) = H(L(h|\gamma)) \sum_{l=1}^{n} l P(\nu = l) = H(L(h|\gamma))\, \mathbb{E}[\nu].$$

We have

$$\begin{aligned}\mathbb{E}[\nu] &= \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{I}\{|f_h(S_{t-1})| \geq a(\gamma)\}\right] \\ &= \sum_{t=1}^{n} \mathbb{E}[\mathbb{I}\{|f_h(S_{t-1})| \geq a(\gamma)\}] \\ &= \sum_{t=1}^{n} \mathbb{P}(|f_h(S_{t-1})| \geq a(\gamma)) \\ &= nP_a\end{aligned}$$

where

$$P_a := P_a^{(h,\gamma)} := \mathbb{P}(|f_h(S_{t-1})| \geq a(\gamma)).$$

Hence the complexity of system (h,γ) is bounded as follows,

$$0 \leq \mathcal{C}(h,\gamma) \leq H(L(h|\gamma))\, P_a \leq 1 \tag{4.3}$$

since the entropy of a binary random variable has a maximum value of 1.

As γ increases, P_a decreases (since $a(\gamma)$ is non-decreasing) and the upper bound on $\mathcal{C}(h,\gamma)$ decreases. As the prediction by (h,γ) becomes more accurate, $L(h|\gamma)$ decreases away from $1/2$ and the entropy $H(L(h|\gamma))$ decreases away from 1, therefore the upper bound on $\mathcal{C}(h,\gamma)$ decreases. If the prediction error $L(h|\gamma)$ decreases with increasing confidence threshold $a(\gamma)$ (this is typical behavior for a good system), then as γ increases, $a(\gamma)$ increases, and therefore both P_a and $H(L(h|\gamma))$ decrease which results in decreasing complexity $\mathcal{C}(h,\gamma)$.

A particular case is when the confidence threshold $a = 0$ and h is the Bayes' classifier. The system predicts at every state of $\mathbb{S}_k$ (has full coverage) with optimal accuracy. In this case, $P_a = 1$ and $H(L(h|\gamma))$ takes a lowest value over all systems with $a = 0$, hence the system minimizes the upper bound on the complexity over all such systems.

The next section studies a second property, stability of a system.

5. Stability

According to Definition 5, for input $X^{(n)}$, a system's output $Y^{(\nu)}$ corresponds to the error sequence (3.36) via the function (3.39). In the current section we define a notion of stability of an admissible system (h,γ). It measures how the average of this sequence changes when we present two different random input sequences. We evaluate the discrepancy $\Upsilon_{m,n}(h,\gamma)$ and declare a system (h,γ) as unstable if the discrepancy deviates from the expected value by a statistically significant amount. We start by constructing a test of significance.

Define the following null hypothesis:

Null Hypothesis: The expected discrepancy for an admissible system (h,γ) is non-positive, that is, $\mathbb{E}\Upsilon_{m,n}(h,\gamma) \leq 0$.

From (3.50) it follows that the null hypothesis is true.

Theorem 13 which is stated below, shows that the event $\Upsilon_{m,n}(h,\gamma) > \epsilon$ has a probability less than δ. We use this result for constructing a significance test, as follows: let $x^{(m)}$, $x^{(n)}$ be the realizations of the random variables $X^{(m)}$ and $X^{(n)}$, respectively. Denote by β_n the realization (based on $x^{(n)}$) of the random variable $\mathfrak{L}_\nu^{(n,\gamma/2)}(h)$ and define by α_m the realization (based on $x^{(m)}$) of the random variable $L_m^{(b(2\gamma))}(h)$. We define the following significance test of level δ and critical value ϵ (ϵ is stated in Theorem 13):

Significance Test: if $\beta_n > \alpha_m + \epsilon$ then reject the null hypothesis.

We use this test to decide if system (h,γ) is stable as follows: Calculate the critical value ϵ using Theorem 13 and apply the above significance test. If the difference $\beta_n - \alpha_m$ is larger than ϵ then reject the null hypothesis. This means that the discrepancy value deviates by a statistically significant amount from its expected value and we declare that the system is *unstable*. This is formalized in the next definition.

Definition 12. *(Stability)* Let (h,γ) be an admissible system. Let m, n, and $\ell \leq n$ be positive integers. Let $X^{(m)}$, $X^{(n)}$ be drawn from a Markov chain (environment). Evaluate the discrepancy $\Upsilon_{m,n}(h,\gamma)$. For any $0 < \delta \leq 1$ and $\epsilon(\delta) > 0$ we say that system (h,γ) is $\epsilon(\delta)$-*stable* if

$$P(\exists \omega \in (\ell/n, 1] : \nu \geq \omega n,\ \Upsilon_{m,n}(h,\gamma) > \epsilon(\delta)) \leq \delta \tag{5.1}$$

where ν is defined in (3.34). Alternatively stated, a system (h,γ) is $\epsilon(\delta)$-*stable* if with a confidence of at least $1-\delta$, for all $\omega \in (\ell/n,1]$ such that $\nu \geq \omega n$, the discrepancy $\Upsilon_{m,n}(h,\gamma) \leq \epsilon(\delta)$.

According to this definition, a larger value of ϵ means that the range of possible discrepancy values that are permitted while still declaring the system as ϵ-stable, is larger. A smaller ϵ means that the system is more stable. In (5.1), ω is the ratio of the output sequence length ν to n. It needs to be larger than ℓ/n to ensure that with large probability, the average (in Definition 8) converges to its mean.

We need a function to serve as a critical value ϵ such that the significance test holds for any admissible system (h,γ). Formally, we seek a function

$$\epsilon := \epsilon(m,n,\ell,\gamma,\omega,\delta),$$

such that for any $0<\delta\leq 1$,

$$\mathbb{P}\bigg(\Big(x^{(m)},x^{(n)}\Big) : \exists (h,\gamma)\in\mathcal{A},\, \exists \omega\in(\ell/n,1],\, \nu\geq\omega n,\, \Upsilon_{m,n}(h,\gamma)>\epsilon\bigg)\leq\delta. \tag{5.2}$$

Note that the bound (5.2) holds uniformly over all admissible prediction systems $(h,\gamma)\in\mathcal{A}$. Therefore after drawing $X^{(m)}$ and $X^{(n)}$, the function ϵ serves as a critical value for any choice of admissible system (h,γ) thus one can search over all $\mathcal{A}$ for a more stable system. This choice can be made even after the two random sequences are drawn. Since the discrepancy depends on the length ν of the output sequence, which is random and hence not known in advance (in contrast to m and n), we ensure that the bound also holds uniformly over the range $\ell<\nu\leq n$. This way Theorem 13 holds for any admissible system (h,γ) and any random output sequence of length at least ℓ that may result.

Theorem 13 holds with the following choice of confidence threshold (3.29) and margin penalty (3.23),

$$a(\gamma)=42\gamma,\;\; b(\gamma)=84\gamma \tag{5.3}$$

which are used in the definition of discrepancy in (3.40). Define

$$\begin{aligned}\epsilon(m,n,\ell,\gamma,\omega,\delta) :=\;& \frac{4r(k,k^*)\rho(k^*,\mathfrak{l}_0)}{\omega}\sqrt{\frac{2}{n}\left(N_{\gamma/6}\ln\left(2\left\lceil\frac{6k}{\gamma}\right\rceil+1\right)+\ln\left(\frac{4(3-2\ell/n)k}{(\omega-\ell/n)\gamma\delta}\right)\right)}\\ +\;& 2r(k,k^*)\rho(k^*,\mathfrak{l}_0)\sqrt{\frac{2}{m}\left(N_{\gamma/3}\ln\left(2\left\lceil\frac{3k}{\gamma}\right\rceil+1\right)+\ln\left(\frac{2(3-2\ell/n)k}{\gamma\delta}\right)\right)}.\end{aligned} \tag{5.4}$$

Theorem 13. *(Critical value) Let a and b be the confidence threshold and margin penalty as defined in (5.3). For positive integers m, n, k, k^* and $\ell\leq n$, let N_γ denote the γ-covering number of the state space $\mathbb{S}_k$ with respect to distance function d. Let $X^{(m)}$, $X^{(n)}$ be drawn from a Markov chain of order k^*. For any $0<\delta\leq 1$, the probability that there exists an admissible system $(h,\gamma)\in\mathcal{A}$ and there exists $\omega\in(\ell/n,1]$ such that the length of the output sequence ν is at least ωn and the discrepancy $\Upsilon_{m,n}(h,\gamma)>\epsilon(m,n,\ell,\gamma,\omega,\delta)$, is no more than δ.*

The proof of the theorem is in Section A.2.

To use Theorem 13, one first chooses positive integers m, n, $\ell\leq n$, k, k^* where k^* represents the true unknown order of the environment's Markov chain (since k^* is assume unknown, then the choice of k^* is based on a theoretical scenario for a Markov chain environment). Then one draws $X^{(m)}$ followed by $X^{(n)}$ at a future time, then picks any admissible system $(h,\gamma)\in\mathcal{A}$, evaluates $L_m^{(b(2\gamma))}(h)$, obtains the system's error sequence $\Psi^{(\nu^{(a(\gamma))})}(h)$ (defined in (3.36)), evaluates $\mathfrak{L}_{\nu^{(a(\gamma))}}^{(n,\gamma/2)}$ and obtains the discrepancy value $\Upsilon_{m,n}(h,\gamma)$. One then chooses for ω any value in the interval $(\ell/n,1]$ such that the random length $\nu^{(a(\gamma))}$ (which is known since the sequence $X^{(n)}$ has already been drawn and $\Psi^{(\nu^{(a(\gamma))})}(h)$ has been already obtained) is no smaller than ωn. Then any value of $0<\delta\leq 1$ is chosen, and the values for m, n, ℓ, k, k^*, γ, ω are plugged into the expression of $\epsilon(m,n,\ell,\gamma,\omega,\delta)$ to obtain the critical value ϵ that is specified in the theorem. One is guaranteed by the theorem with probability at least $1-\delta$ that if the measured discrepancy value is larger than this value of ϵ then the system is unstable, otherwise it is ϵ-stable. Hence, with a single pair of random sequences $X^{(m)}$, $X^{(n)}$, we can determine if any admissible system is ϵ-stable or not for a given environment.

Remark 14. The value ℓ is the required minimum length of the error sequence $\Psi^{(\nu)}$. The theorem holds only if the value of ω is greater than ℓ/n. If after drawing $X^{(n)}$ and evaluating the error sequence $\Psi^{(\nu)}$, the length $\nu<\ell$, then the theorem cannot be applied, that is, ϵ cannot be used as a critical value.

So far we studied two properties of a system, complexity and stability. We saw that the an upper bound on the complexity of a system (h,γ) decreases as γ increases and, from Theorem 13, we see that the critical value ϵ decreases as γ increases. Therefore, a less complex system (h,γ) has a smaller value of ϵ. Hence, a *less* complex system is *more* stable.

The next section studies a third property which is referred to as system performance guarantee.

6. Performance

In this section we state an upper bound on the probability of prediction error $L(h\,|\gamma)$ (defined in (3.32)). In the context of Section 2 this bound represents a system's performance guarantee. Similar to the average prediction error in Definition 8, next we define the average prediction margin error of a system.

Definition 15. (*Average prediction margin error*) For input $X^{(m)}$, the *average number of margin errors* when making a prediction by system (h,γ), is defined as

$$\mathcal{L}^{(m,\gamma/6)}_{\nu^{(a)}}(h) := \frac{1}{\nu^{(a)}} \sum_{l:\left|f_h(S_{t_l-1})\right|\geq a(\gamma/6)} \Psi^{(m,\gamma)}_{t_l}, \tag{6.1}$$

where $\nu := \nu^{(a)}$ is the number of times that the sequence $S^{(m)}$ enters a state $s \in \mathbb{S}_k$ such that the confidence function satisfies $|f_h(s)| \geq a(\gamma/6)$ and $\Psi^{(m,\gamma)}_{t_l}$ is defined in Definition 2.

The average prediction margin error is used in the next theorem to obtain an upper bound on the prediction error (3.32) of a system. The theorem holds with the following choice of confidence threshold (3.29) and margin penalty (3.23),

$$a(\gamma) = 21\gamma,\ b(\gamma) = 44\gamma. \tag{6.2}$$

Let

$$\xi(m,\gamma,\omega,\delta) := \frac{4r(k,k^*)\rho(k^*,\mathfrak{l}_0)}{\omega}\sqrt{\frac{2}{m}\left(N_{\gamma/6}\ln\left(2\left\lceil\frac{6k}{\gamma}\right\rceil+1\right)+\ln\left(\frac{8(1-\ell/m)k)}{(\omega-\ell/m)\gamma\delta}\right)\right)} \tag{6.3}$$

where $r(k,k^*)$ is defined in (3.13).

Theorem 16. *(Performance guarantee) Let a and b be the confidence threshold and margin penalty as defined in (6.2). For $\gamma > 0$, let N_γ be the γ-covering number of $\mathbb{S}_k$ with respect to the metric d. Let m and $\ell \leq m$ be positive integers. Let $X^{(m)}$ be a Markov chain drawn randomly from the environment. For any $0 < \delta \leq 1$, the following bound holds,*

$$\mathbb{P}\left(\exists h\in\mathcal{H},\ \exists 0<\gamma\leq \mathit{diam}(\mathbb{S}_k), \exists\omega\in(\ell/m,1]:\ L(h|\gamma) > \mathcal{L}^{(m,\gamma/6)}_{\nu^{(a)}}(h) + \xi(m,\gamma,\omega,\delta), \nu^{(a)}\geq\omega m\right)\leq\delta. \tag{6.4}$$

The proof is in Section A.3.

We refer to the sum

$$\hat{\mathcal{L}}^{(\gamma,\delta)}_m(h) := \mathcal{L}^{(m,\gamma/6)}_{\nu^{(a)}}(h) + \xi(m,\gamma,\omega,\delta)$$

as the *performance guarantee function* since the theorem guarantees with a probability of at least $1-\delta$ that the prediction error probability $L(h\,|\gamma)$ (considered as the performance of system (h,γ)) is no larger than $\hat{\mathcal{L}}^{(\gamma,\delta)}_m(h)$. Note that while the bound (6.4) holds uniformly over all systems (h,γ) with probability at least $1-\delta$, it is not loose, to the extent that it still depends on the specific system (h,γ) that is considered. The dependence is via the first term $\mathcal{L}^{(m,\gamma/6)}_{\nu^{(a)}}(h)$ which depends on h and γ, and also via the second term which depends on γ.

The parameter γ influences the sensitivity of the performance guarantee function $\hat{\mathcal{L}}^{(\gamma,\delta)}_m(h)$ to an input $x^{(m)}$. Two input sequences whose states are close with respect to d, may yield very different values for $\hat{\mathcal{L}}^{(\gamma,\delta)}_m(h)$ if γ is small. Yet if γ is large, then their corresponding values $\hat{\mathcal{L}}^{(\gamma,\delta)}_m(h)$ will be close. A higher value of γ means the performance guarantee function is less sensitive to a change in the input, and we say that the performance guarantee function is more robust.

7. Conclusions

This paper studies a problem of predicting a binary Markov chain by a prediction system. A system is a pair of structural and behavioral components. The structural component is a classifier h, a binary function defined over a finite state space. The behavioral component is based on a parameter γ that controls the confidence threshold used for prediction with h. The output of a system consists of a binary sequence that indicates errors made by the system in predicting an input sequence. We define system complexity, system stability and system performance guarantee function. Complexity of a system is the average number of information bits (minimal expected description length) needed to describe the output, per input bit. Stability is defined as the difference between average prediction error and the average prediction margin error on two input sequences. Performance guarantee is stated as an upper bound on the prediction error probability. These three quantities are interrelated via their dependence on the behavioral parameter γ. This parameter influences the performance guarantee function in such a way that a larger value of γ makes it more robust and less sensitive to fluctuations in random input. Our results show that, with a larger value of γ, a system becomes less complex, more stable and its performance guarantee is less sensitive to fluctuations in input.

Appendix

Starting in Section A.3 and onwards the proofs of the statements are provided. Section A.1 states a lemma which is used in the proofs.

A.1. **Concentration.** The total variation distance between two discrete probability distributions p and q on a domain $\mathcal{X}$ is defined as (see for instance, Definition 13.2, [3])

$$\|p-q\|_{TV} := \sup_{A\subset\mathcal{X}} |p(A)-q(A)|$$

and is related to the l_1-distance as follows (Lemma 13.3, [3]),

$$\|p-q\|_{TV} = \frac{1}{2}\|p-q\|_1 . \tag{A.1}$$

A normalized Hamming metric for sequences $s^{*(n)} = \{s_t^*\}_{t=1}^n \in \mathbb{S}_{k^*}^n$ is defined as follows: for any $s^{*(n)}, q^{*(n)} \in \mathbb{S}_{k^*}^n$,

$$d_H\left(s^{*(n)}, q^{*(n)}\right) := \sum_{t=1}^n \mathbb{I}\{s_t^* \neq q_t^*\} . \tag{A.2}$$

For a discrete time Markov chain with transition matrix Q defined in (3.5), denote its i^{th} row by $q(\cdot|i)$, that is, the conditional probability distribution given that the current state is $s^{*(i)}$. For discrete time l, we denote the entries of the matrix Q^l by $q^{(l)}(j|i)$, and $q^{(l)}(\cdot|i)$ denotes the i^{th} row of Q^l. Define

$$\tau_l := \max_{0\le i\le 2^{k^*}-1} \left\| q^{(l)}(\cdot|i) - \pi^* \right\|_{TV}$$

where π^* is defined in (3.8). The next lemma is Theorem 1.1 of [8] applied to our finite Markov chain. It establishes a concentration bound for $S^{*(n)}$ and functions that are Lipschitz with respect to the Hamming norm. (That we can apply this theorem follows from the fact that a Markov chain can be regarded as a hidden Markov chain by letting the emission alphabet be identical to the state space and emission probabilities to be delta-functions.)

Lemma 17. *([8] Theorem 1.1) For $1 \le \alpha < \infty$ and $0 \le \beta < 1$, if $\tau_l \le \alpha\beta^{l-1}$ for $l = 1, 2, \ldots$, then for any $\varphi : \mathbb{S}_{k^*}^n \to \mathbb{R}$ with Lipschitz constant 1 with respect to the Hamming metric, the following holds:*

$$\mathbb{P}\left(\varphi\left(S^{*(n)}\right) - \mathbb{E}\varphi\left(S^{*(n)}\right) > n\kappa\right) \le \exp\left(-\frac{n(1-\beta)^2\kappa^2}{2\alpha^2}\right) \tag{A.3}$$

and

$$\mathbb{P}\left(\mathbb{E}\varphi\left(S^{*(n)}\right) > \varphi\left(S^{*(n)}\right) + n\kappa\right) \le \exp\left(-\frac{n(1-\beta)^2\kappa^2}{2\alpha^2}\right) . \tag{A.4}$$

For this to be useful we need to ensure that we can apply this lemma to the random sequences $X^{(m)}$ and $X^{(n)}$ drawn from the environment Markov chain. Let us investigate this for $X^{(n)}$ (the case for $X^{(m)}$ would then follow directly).

From Section 3.1, it follows that the sequence $S^{*(n)}$ is a sample of a homogeneous Markov chain with two types of transitions, a negative and a positive one and, from Section 3.1, we denote their probabilities by

$p(-1|i) := p(-1|s^{*(i)})$ and $p(1|i) := p(1|s^{*(i)})$, respectively, $0 \leq i \leq 2^{k^*} - 1$. The transition matrix Q of (3.5) has the following form,

$$Q := \begin{pmatrix} q(-1|0) & q(1|0) & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & q(-1|1) & q(1|1) & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & 0 & 0 & q(-1|2) & q(1|2) & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & 0 & q(-1|2^{k^*-1}-1) & q(1|2^{k^*-1}-1) \\ q(-1|2^{k^*-1}) & q(1|2^{k^*-1}) & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & q(-1|2^{k^*-1}+1) & q(1|2^{k^*-1}+1) & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & 0 & 0 & q(-1|2^{k^*-1}+2) & q(1|2^{k^*-1}+2) & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & 0 & q(-1|2^{k^*}-1) & q(1|2^{k^*}-1) \end{pmatrix}. \tag{A.5}$$

It is clear that for all $k^* \geq 2$, Q is not a (strictly) positive matrix. Thus we need some additional work to show that the condition of Lemma 17 holds for our Markov chain under all the cases of Assumption 1 with ρ as defined in (3.53).

A non-negative irreducible matrix is regular (or primitive) if it has a single eigenvalue on the unit circle [10]. As mentioned in Section 3.3, the Markov chain is irreducible with a non-negative transition matrix. By assumption (3.7), Q has at least one positive entry on the diagonal, then it follows that Q is primitive ([10] Example 8.3.3). From [10] p.693, $\lim_{l\to\infty} Q^l$ exists and is a matrix all of whose rows are identical to the stationary probability distribution of the Markov chain. More relevant for us, by the Frobenius's test for primitivity ([10], (8.3.16)), it follows that because Q is primitive then there exists a finite integer $\mathfrak{l}_0$, such that for $l \geq \mathfrak{l}_0$, $Q^l > 0$, that is, all the elements of Q^l are positive ([10], Example 8.3.4 shows that $\mathfrak{l}_0 \leq 2^{2k^*} - 2^{k^*+1} + 2$). And Q^l is row-stochastic, that is for every row, the sum of the entries is 1 (this can be shown by induction on l). We choose $\mathfrak{l}_0$ as in (3.51). That $Q^{\mathfrak{l}_0} > 0$ implies the minimum value μ_0 of $Q^{\mathfrak{l}_0}$ satisfies

$$\mu_0 > 0. \tag{A.6}$$

Recall from (3.8) that $\pi^* := \left[\pi^*_0, \ldots, \pi^*_{2^{k^*}-1}\right]$ is the stationary probability distribution of the chain. Note that every row of $Q^{\mathfrak{l}_0}$ (which is a $2^{k^*} \times 2^{k^*}$ matrix) has an entry whose value is no larger than $1/2^{k^*}$ because $Q^{\mathfrak{l}_0}$ is row-stochastic and hence the sum of the entries in every row is 1. Hence

$$\mu_0 \leq \frac{1}{2^{k^*}}. \tag{A.7}$$

Define the constant

$$c_0 = c_0(k^*, \mathfrak{l}_0) := 1 - 2^{k^*}\mu_0 \tag{A.8}$$

then, by (A.6) and (A.7), it follows that

$$0 \leq c_0 < 1.$$

We need to consider the cases of Assumption 1. We start with case (i) where the environment has a Q such that $\mu_0 \neq 2^{-k^*}$ so that $c_0 > 0$. In this case, by Proposition 10.5(ii) [3], if $\mathfrak{l}_0 \geq 2$, we have for every $0 \leq i$, $j \leq 2^{k^*} - 1$, and every $l \geq \mathfrak{l}_0$,

$$\left|q^{(l)}(j|i) - \pi^*_j\right| \leq \left(\frac{1}{c_0}\right) c_0^{l/\mathfrak{l}_0}. \tag{A.9}$$

This means that the distance between the i^{th} row of Q^l and the stationary distribution is

$$\left\|q^{(l)}(\cdot|i) - \pi^*\right\|_1 = \sum_{j=0}^{2^{k^*}-1} \left|q^{(l)}(j|i) - \pi^*_j\right| \leq \left(\frac{2^{k^*}}{c_0}\right) c_0^{l/\mathfrak{l}_0} \tag{A.10}$$

where $\|\cdot\|_1$ denotes the l_1-norm. Therefore, from (A.1) and (A.10),

$$\begin{aligned} \left\|q^{(l)}(\cdot|i) - \pi^*\right\|_{TV} &\leq \left(\frac{2^{k^*}}{2c_0}\right) c_0^{l/\mathfrak{l}_0} \\ &= \left(\frac{2^{k^*} c_0^{1/\mathfrak{l}_0}}{2c_0}\right) \left(c_0^{1/\mathfrak{l}_0}\right)^{l-1}. \end{aligned} \tag{A.11}$$

Letting

$$\alpha = 2^{k^*-1} c_0^{-(\mathfrak{l}_0-1)/\mathfrak{l}_0} \tag{A.12}$$

and

$$\beta = c_0^{1/\mathfrak{l}_0} \tag{A.13}$$

means that we may use Lemma 17 for $S^{*(n)}$ (whose transition matrix is (3.5)) together with any function φ that is Lipschitz with constant 1. Substituting (A.8) for c_0 and plugging (A.12) and (A.13) for α and β in the bound of Lemma 17, then the concentration bound (A.3) becomes

$$\mathbb{P}\left(\varphi\left(S^{*(n)}\right) - \mathbb{E}\varphi\left(S^{*(n)}\right) > n\kappa\right) \leq \exp\left\{-\frac{n}{2}\left(\frac{\kappa}{\rho}\right)^2\right\}, \tag{A.14}$$

and (A.4) becomes

$$\mathbb{P}\left(\mathbb{E}\varphi\left(S^{*(n)}\right) - \varphi\left(S^{*(n)}\right) > n\kappa\right) \leq \exp\left\{-\frac{n}{2}\left(\frac{\kappa}{\rho}\right)^2\right\}, \tag{A.15}$$

with

$$\rho := \frac{\alpha}{1-\beta} = \frac{2^{k^*-1}}{\left(1-2^{k^*}\mu_0\right)^{(\mathfrak{l}_0-1)/\mathfrak{l}_0}\left(1-\left(1-2^{k^*}\mu_0\right)^{1/\mathfrak{l}_0}\right)}. \tag{A.16}$$

If $\mathfrak{l}_0 = 1$ (and still under case (i) of Assumption 1) then $Q > 0$ and by Proposition 10.5(i) [3], $\left|q^{(l)}(j|i) - \pi_j^*\right| \leq c_0^l$. Following the above steps it suffices to choose $\alpha = 2^{k^*-1}c_0$ and $\beta = c_0$ to obtain

$$\rho := \frac{\alpha}{1-\beta} = \frac{1-2^{k^*}\mu_0}{2\mu_0}.$$

We now consider case (ii) of Assumption 1 where the environment's Q has $\mu_0 = 2^{-k^*}$ and therefore (A.16) cannot be used. The stationary distribution in this case is uniform so Q is doubly stochastic and $\lim_{l\to\infty} Q^l = U$ is reached exactly at time $\mathfrak{l}_0$, that is, $Q^{\mathfrak{l}_0} = U$. The matrix Q is as in (A.5) with $q(1|i) = q(-1|i) = \frac{1}{2}$, for all $0 \leq i \leq 2^{k^*} - 1$. We have $\mathfrak{l}_0 = k^*$ and the limit matrix U has all entries equal to 2^{-k^*} (if $k^* = 1$ then $Q = U$). For $l \geq \mathfrak{l}_0$ the left side of (A.9) equals zero because the limit is reached at $\mathfrak{l}_0$. But for $1 \leq l < \mathfrak{l}_0$ the left side of (A.9) is bounded from above by 2^{-l}. Thus for all $l \geq 1$, the left side of (A.11) is bounded from above by $2^{k^*-1}(1/2)^l$. We let $\alpha = 2^{k^*-2}$ and $\beta = 1/2$ to yield $\rho = 2^{k^*-1}$.

All the above holds for the sequence $S^{*(m)}$, with m replacing n.

In summary, we showed that for every case of Assumption 1, the necessary condition of Lemma 17 that $\tau_l \leq \alpha\beta^{l-1}$ holds and therefore the lemma can be used as a concentration inequality for both sequences $X^{(m)}$ and $X^{(n)}$. In the following sections, we use this fact in proving Lemma 16 and Theorem 13 which deal with $S^{*(m)}$ and $S^{*(n)}$, respectively.

A.2. **Proof of Theorem 13.** Recall the definition of discrepancy,

$$\Upsilon_{m,n}(h,\gamma) := \mathfrak{L}_\nu^{(n,\gamma/2)}(h) - L_m^{(b(2\gamma))}(h)$$

where $\nu := \nu(a(\gamma))$ is the random length of the output sequence and error sequence based on which the average $\mathfrak{L}$ is defined. We need to show that for arbitrary and fixed $0 < \delta < 1$, the probability is no more than δ that for input sequences $X^{(m)}$ and $X^{(n)}$ which are randomly drawn from the environment, there exists $\omega \in (\ell/n, 1]$ and an admissible system $(h,\gamma) \in \mathcal{A}$ such that the length ν of its output sequence satisfies $\nu \geq \omega n$ and its discrepancy satisfies $\Upsilon_{m,n}(h,\gamma) > \epsilon(m,n,\ell,\gamma,\omega,\delta)$, where $\ell \leq n$ is a positive integer fixed in advance.

The probability of this event is bounded from above by

$$P\left(\left\{\left(x^{(m)}, x^{(n)}\right) : \exists h, \exists\gamma, \exists\omega,\ \mathfrak{L}_\nu^{(n,\gamma/2)}(h) - L_m^{(b(2\gamma))}(h) > \epsilon,\ \nu > \omega n\right\}\right) \tag{A.17}$$

where $\nu := \nu^{(a(\gamma/2))}$ is the length of the output sequence and for brevity we write $\exists\gamma$, $\exists h$, $\exists\omega$, instead of $\exists\gamma \in (0, \operatorname{diam}(\mathbb{S}_k)]$, $\exists h \in \mathcal{H}$, $\exists\omega \in (\ell/n, 1]$, respectively.

A.2.1. *Bounding (A.17).* In addition to $a(\gamma)=42\gamma,\ b(\gamma)=84\gamma$ (defined in (5.3)) let us define

$$\tilde{a}(\gamma):=4\gamma,\ \dot{a}(\gamma):=8\gamma,\ \hat{a}(\gamma):=10\gamma,$$

and

$$\hat{b}(\gamma):=43\gamma,\ \tilde{b}(\gamma):=44\gamma,\ \dot{b}(\gamma):=45\gamma,$$

all of which are positive for $\gamma>0$ and satisfy

$$\begin{aligned}
&-b(2\gamma)<-b(\gamma)<-b(\gamma)+\gamma\\
&\leq\ -\dot{b}(\gamma)<-\dot{b}(\gamma)+\gamma\\
&\leq\ -\tilde{b}(\gamma)<-\tilde{b}(\gamma)+\gamma \qquad (A.18)\\
&\leq\ -\hat{b}(\gamma)<-\hat{b}(\gamma)+\gamma\leq-b(\gamma/2)\\
&\leq\ -a(\gamma)\\
&<\ -a(\gamma/2)<-a(\gamma/2)+\gamma\\
&\leq\ -\hat{a}(2\gamma)<-\hat{a}(\gamma)<-\hat{a}(\gamma)+2\gamma \qquad (A.19)\\
&\leq\ -\dot{a}(\gamma)<-\dot{a}(\gamma)+3\gamma\\
&\leq\ -\hat{a}(\gamma/2)<-\hat{a}(\gamma/2)+\gamma\\
&\leq\ -\tilde{a}(\gamma)<-\tilde{a}(\gamma)+\gamma\\
&<\ -\hat{a}(\gamma/4)<-\hat{a}(\gamma/8)\\
&<\ 0. \qquad (A.20)
\end{aligned}$$

For $\theta\in\mathbb{S}_{k+1}$ we write $\theta=[s,x]$ and define the sets $A_{h,\gamma},\ B_{h,\gamma}\subseteq\mathbb{S}_{k+1}$ by

$$A_{h,\gamma}:=\{\theta:\ |f_h(s)|\geq a(\gamma)\},\ B_{h,\gamma}:=\{\theta:\ xf_h(s)\geq b(\gamma)\}. \qquad (A.21)$$

Define the counterpart of $\mathfrak{L}_\nu^{(n,\gamma/2)}(h)$ as follows

$$\overline{\mathfrak{L}}_\nu^{(n,\gamma/2)}(h):=\frac{1}{\nu}\sum_{l:\left|f_h(S_{t_l-1})\right|\geq a(\gamma/2)}\mathbb{I}\{X_{t_l}f_h(S_{t_l-1})\geq 0\}. \qquad (A.22)$$

Denote by

$$\tilde{L}^{(\gamma)}(h|\gamma):=\mathbb{P}\left(X_tf_h(S_{t-1})<\tilde{b}(\gamma)\middle||f_h(S_{t-1})|\geq\tilde{a}(\gamma)\right)$$

and

$$\overline{\tilde{L}^{(\gamma)}}(h|\gamma):=1-\tilde{L}^{(\gamma)}(h|\gamma).$$

Define the sets

$$\tilde{A}_{h,\gamma}:=\{\theta:\ |f_h(s)|\geq\tilde{a}(\gamma)\},\ \tilde{B}_{h,\gamma}:=\left\{\theta:\ xf_h(s)\geq\tilde{b}(\gamma)\right\}.$$

Let us fix γ, h and ω in the expression whose probability is (A.17). We consider the event $\mathfrak{L}_\nu^{(n,\gamma/2)}(h)-L_m^{(b(2\gamma))}(h)>\epsilon,\ \nu>\omega n$, and bound its probability from above by a sum of two terms. The first is a probability of an event that involves just the sequence $X^{(n)}$ and the second term is a probability that involves just the sequence $X^{(m)}$, as follows,

$$\begin{aligned}
&\mathbb{P}\left(\mathfrak{L}_\nu^{(n,\gamma/2)}(h)-L_m^{(b(2\gamma))}(h)>\epsilon,\ \nu>\omega n\right)\\
&\leq\ \mathbb{P}\left(\mathfrak{L}_\nu^{(n,\gamma/2)}(h)-\tilde{L}^{(\gamma)}(h|\gamma)>\epsilon_1,\ \nu>\omega n\right)+\mathbb{P}\left(\tilde{L}^{(\gamma)}(h|\gamma)-L_m^{(b(2\gamma))}(h)>\epsilon_2\right) \qquad (A.23)
\end{aligned}$$

where

$$\epsilon=\epsilon_1+\epsilon_2 \qquad (A.24)$$

and the second term in (A.23) does not have the condition that involves ν because ν does not depend on $X^{(m)}$.

We have

$$\overline{\tilde{L}^{(\gamma)}}(h|\gamma)=P\left(\tilde{B}_{h,\gamma}\middle|\tilde{A}_{h,\gamma}\right) \qquad (A.25)$$

thus the first term in (A.23) equals

$$\mathbb{P}\left(\overline{\tilde{L}^{(\gamma)}}\left(h|\gamma\right)-\overline{\mathfrak{L}}_{\nu}^{(n,\gamma/2)}(h)>\epsilon_1,\,\nu>\omega n\right)=\mathbb{P}\left(P\left(\tilde{B}_{h,\gamma}\left|\tilde{A}_{h,\gamma}\right.\right)-\overline{\mathfrak{L}}_{\nu}^{(n,\gamma/2)}(h)>\epsilon_1,\,\nu>\omega n\right). \tag{A.26}$$

Denote by

$$\begin{aligned}\hat{P}_n(A_{h,\gamma/2}) &:= \frac{1}{n}\sum_{t=1}^{n}\mathbb{I}\left\{\Theta_t\in A_{h,\gamma/2}\right\}\\ &= \frac{\nu}{n}\end{aligned} \tag{A.27}$$

where the equality follows from (3.34). We have,

$$\begin{aligned}P\left(\tilde{B}_{h,\gamma}\left|\tilde{A}_{h,\gamma}\right.\right)-\overline{\mathfrak{L}}_{\nu}^{(n,\gamma/2)}(h) &= \frac{1}{\hat{P}_n\left(A_{h,\gamma/2}\right)}\left(P\left(\tilde{B}_{h,\gamma},\tilde{A}_{h,\gamma}\right)-\overline{\mathfrak{L}}_{\nu}^{(n,\gamma/2)}(h)\hat{P}_n\left(A_{h,\gamma/2}\right)\right)\\ &\quad+P\left(\tilde{B}_{h,\gamma},\tilde{A}_{h,\gamma}\right)\left(\frac{1}{P\left(\tilde{A}_{h,\gamma}\right)}-\frac{1}{\hat{P}_n\left(A_{h,\gamma/2}\right)}\right)\end{aligned} \tag{A.28}$$

We have,

$$\begin{aligned}&P\left(\tilde{B}_{h,\gamma},\tilde{A}_{h,\gamma}\right)\left(\frac{1}{P\left(\tilde{A}_{h,\gamma}\right)}-\frac{1}{\hat{P}_n\left(A_{h,\gamma/2}\right)}\right)\\ &=P\left(\tilde{B}_{h,\gamma}\left|\tilde{A}_{h,\gamma}\right.\right)P\left(\tilde{A}_{h,\gamma}\right)\left(\frac{\hat{P}_n\left(A_{h,\gamma/2}\right)-P\left(\tilde{A}_{h,\gamma}\right)}{\hat{P}_n\left(A_{h,\gamma/2}\right)P\left(\tilde{A}_{h,\gamma}\right)}\right)\\ &=\frac{P\left(\tilde{B}_{h,\gamma}\left|\tilde{A}_{h,\gamma}\right.\right)}{\hat{P}_n\left(A_{h,\gamma/2}\right)}\left(\hat{P}_n\left(A_{h,\gamma/2}\right)-P\left(\tilde{A}_{h,\gamma}\right)\right).\end{aligned}$$

Thus the first term of (A.23) is bounded from above by

$$\begin{aligned}&\mathbb{P}\left(P\left(\tilde{B}_{h,\gamma},\tilde{A}_{h,\gamma}\right)-\overline{\mathfrak{L}}_{\nu}^{(n,\gamma/2)}(h)\hat{P}_n\left(A_{h,\gamma/2}\right)>\epsilon_1\hat{P}_n(A_{h,\gamma/2})/2,\,\nu>\omega n\right)\\ &\quad+\quad\mathbb{P}\left(\hat{P}_n(A_{h,\gamma/2})-P(\tilde{A}_{h,\gamma})>\epsilon_1\hat{P}_n(A_{h,\gamma/2})/2P\left(\tilde{B}_{h,\gamma}\left|\tilde{A}_{h,\gamma}\right.\right),\,\nu>\omega n\right).\end{aligned} \tag{A.29}$$

For convenience, we denote by

$$\hat{P}_n\left(B_{h,0},A_{h,\gamma/2}\right):=\overline{\mathfrak{L}}_{\nu}^{(n,\gamma/2)}(h)\hat{P}_n\left(A_{h,\gamma/2}\right)$$

since $\overline{\mathfrak{L}}_{\nu}^{(n,\gamma/2)}(h)$ is the empirical probability of the set $B_{h,0}$ conditioned on the state being in $A_{h,\gamma/2}$.

Therefore (A.17) is bounded from above by

$$\mathbb{P}\left(\exists h,\exists\gamma,\exists\omega:\,P\left(\tilde{B}_{h,\gamma},\tilde{A}_{h,\gamma}\right)-\hat{P}_n\left(B_{h,0},A_{h,\gamma/2}\right)>\epsilon_1\hat{P}_n(A_{h,\gamma/2})/2,\,\nu>\omega n\right) \tag{A.30}$$

$$+\quad\mathbb{P}\left(\exists h,\exists\gamma,\exists\omega:\hat{P}_n\left(A_{h,\gamma/2}\right)-P\left(\tilde{A}_{h,\gamma}\right)>\epsilon_1\hat{P}_n(A_{h,\gamma/2})/2P\left(\tilde{B}_{h,\gamma}\left|\tilde{A}_{h,\gamma}\right.\right),\,\nu>\omega n\right) \tag{A.31}$$

$$+\quad\mathbb{P}\left(\exists h,\exists\gamma:\tilde{L}^{(\gamma)}\left(h|\gamma\right)-L_m^{(b(2\gamma))}(h)>\epsilon_2\right). \tag{A.32}$$

Denote by

$$A_{j,\gamma'}^{(\gamma)}:=\left\{\theta=[s,x]:\left|f_j^{(\gamma)}(s)\right|\geq\hat{a}(\gamma')\right\}, \tag{A.33}$$

and

$$B_{j,\gamma'}^{(\gamma)}:=\left\{\theta:\,xf_j^{(\gamma)}(s)\geq\hat{b}(\gamma')\right\}.$$

We have the following.

Claim 18. For h and h_j such that f_h and f_j satisfy (3.56) with $\alpha=\gamma$, we have

$$P\left(\tilde{B}_{h,\gamma},\tilde{A}_{h,\gamma}\right)\leq P\left(B_{j,\gamma}^{(\gamma)},A_{j,\gamma/4}^{(\gamma)}\right). \tag{A.34}$$

Proof. Consider $\theta = [s, x] \in \tilde{B}_{h,\gamma} \bigcap \tilde{A}_{h,\gamma}$, then $x f_h(s) \geq \tilde{b}(\gamma)$ and $|f_h(s)| \geq \tilde{a}(\gamma)$. If $x = -1$ then

$$f_h(s) \leq -\tilde{b}(\gamma)$$

and we have,

$$\begin{aligned} f_j^{(\gamma)}(s) &\leq f_h(s) + \gamma \\ &\leq -\tilde{b}(\gamma) + \gamma \\ &\leq -\hat{b}(\gamma) \end{aligned}$$

where the last inequality follows from (A.20). Therefore $x f_j^{(\gamma)}(s) \geq \hat{b}(\gamma)$ and hence $\theta \in B_{j,\gamma}^{(\gamma)}$. If $x = 1$ then $f_h(s) \geq \tilde{b}(\gamma)$ and we have,

$$\begin{aligned} f_j^{(\gamma)}(s) &\geq f_h(s) - \gamma \\ &\geq \tilde{b}(\gamma) - \gamma \\ &\geq \hat{b}(\gamma) \end{aligned}$$

therefore $x f_j^{(\gamma)}(s) \geq \hat{b}(\gamma)$ and hence $\theta \in B_{j,\gamma}^{(\gamma)}$.

We conclude that $\tilde{B}_{h,\gamma} \subseteq B_{j,\gamma}^{(\gamma)}$. Next, consider $\theta = [s, x] \in \tilde{A}_{h,\gamma}$. If $f_h(s) \leq -\tilde{a}(\gamma)$, then from (A.20) we have

$$\begin{aligned} f_j^{(\gamma)}(s) &\leq f_h(s) + \gamma \\ &< -\tilde{a}(\gamma) + \gamma \\ &\leq -\hat{a}(\gamma/4). \end{aligned} \tag{A.35}$$

If $f_h(s) \geq \tilde{a}(\gamma)$ then

$$\begin{aligned} f_j^{(\gamma)}(s) &\geq f_h(s) - \gamma \\ &\geq \tilde{a}(\gamma) - \gamma \\ &\geq \hat{a}(\gamma/4). \end{aligned}$$

Therefore, we conclude that $\tilde{A}_{h,\gamma} \subseteq A_{j,\gamma/4}^{(\gamma)}$. And with the above, $\tilde{B}_{h,\gamma} \bigcap \tilde{A}_{h,\gamma} \subseteq B_{j,\gamma}^{(\gamma)} \bigcap A_{j,\gamma/4}^{(\gamma)}$ from which (A.34) follows. □

Next, we have the following:

Claim 19. For h and h_j such that f_h and $f_j^{(\gamma)}$ satisfy (3.56) with $\alpha = \gamma$, we have

$$\hat{P}_n \left(B_{j,\gamma}^{(\gamma)}, A_{j,\gamma/4}^{(\gamma)} \right) \leq \hat{P}_n \left(B_{h,0}, A_{h,\gamma} \right). \tag{A.36}$$

Proof. Consider any $\theta = [s, x] \in A_{j,\gamma/4}^{(\gamma)} \bigcap B_{j,\gamma}^{(\gamma)}$. If $x = -1$ then $f_j^{(\gamma)}(s) \leq -\hat{b}(\gamma)$. From (3.56) we have

$$\begin{aligned} f_h(s) &\leq f_j^{(\gamma)}(s) + \gamma \\ &\leq -\hat{b}(\gamma) + \gamma \\ &\overset{(i)}{<} -a(\gamma) \\ &\overset{(ii)}{<} 0 \end{aligned}$$

where (i) and (ii) follow from (A.20). Hence for this θ, $f_h(s) < -a(\gamma)$ and $x f_h(s) > 0$ (since $x = -1$) and therefore $\theta \in A_{h,\gamma} \bigcap B_{h,0}$. If $x = 1$ then $f_j^{(\gamma)}(s) \geq \hat{b}(\gamma)$ and we have

$$\begin{aligned} f_h(s) &\geq f_j^{(\gamma)}(s) - \gamma \\ &\geq \hat{b}(\gamma) - \gamma \\ &> a(\gamma) \\ &> 0 \end{aligned}$$

so we conclude that

$$A_{j,\gamma/4}^{(\gamma)} \bigcap B_{j,\gamma}^{(\gamma)} \subseteq A_{h,\gamma} \bigcap B_{h,0}. \tag{A.37}$$

Hence (A.36) follows. □

We have the following claim.

Claim 20. For h and h_j such that f_h and $f_j^{(\gamma)}$ satisfy (3.56) with $\alpha = \gamma$,

$$\hat{P}_n(A_{j,2\gamma}^{(\gamma)}) \geq \hat{P}_n(A_{h,\gamma/2}). \tag{A.38}$$

Proof. Consider $\theta \in A_{h,\gamma/2}$. If $f_h(s) \leq -a(\gamma/2)$ then we have

$$\begin{aligned} f_j^{(\gamma)}(s) &\leq f_h(s) + \gamma \\ &\leq -a(\gamma/2) + \gamma \\ &\leq -\hat{a}(2\gamma) \end{aligned}$$

where the last inequality follows from (A.20). It follows that $\theta \in A_{j,2\gamma}^{(\gamma)}$. If $f_h(s) \geq a(\gamma/2)$ then we have

$$\begin{aligned} f_j^{(\gamma)}(s) &\geq f_h(s) - \gamma \\ &\geq a(\gamma/2) - \gamma \\ &\geq \hat{a}(2\gamma) \end{aligned}$$

so $\theta \in A_{j,2\gamma}^{(\gamma)}$. This proves (A.38). □

Claim 21. For h and h_j such that f_h and $f_j^{(\gamma)}$ satisfy (3.56) with $\alpha = \gamma$,

$$P\left(\tilde{A}_{h,\gamma}\right) \geq P\left(A_{j,\gamma/2}^{(\gamma)}\right). \tag{A.39}$$

Proof. Consider any $\theta = [s, x] \in A_{j,\gamma/2}^{(\gamma)}$. If $f_j^{(\gamma)}(s) \leq -\hat{a}(\gamma/2)$ then

$$\begin{aligned} f_h(s) &\leq f_j^{(\gamma)}(s) + \gamma \\ &< -\hat{a}(\gamma/2) + \gamma \\ &\leq -\tilde{a}(\gamma) \end{aligned}$$

which follows from (A.20) and hence $\theta \in \tilde{A}_{h,\gamma}$. If $f_j^{(\gamma)}(s) \geq \hat{a}(\gamma/2)$ then

$$\begin{aligned} f_h(s) &\geq f_j^{(\gamma)}(s) - \gamma \\ &\geq \hat{a}(\gamma/2) - \gamma \\ &\geq \tilde{a}(\gamma) \end{aligned}$$

so $\theta \in \tilde{A}_{h,\gamma}$. Therefore $A_{j,\gamma/2}^{(\gamma)} \subseteq \tilde{A}_{h,\gamma}$ and (A.39) is proved. □

A.2.2. *Bounding (A.30).* Denote by

$$\lambda := \frac{\nu}{n} = \hat{P}_n(A_{h,\gamma/2}).$$

The probability in (A.30) is expressed as follows,

$$\mathbb{P}\left(\exists h, \exists \gamma, \exists \omega : P\left(\tilde{B}_{h,\gamma}, \tilde{A}_{h,\gamma}\right) - \hat{P}_n\left(B_{h,0}, A_{h,\gamma/2}\right) > \epsilon_1 \lambda/2,\, \lambda > \omega\right)$$

which is bounded from above by

$$\mathbb{P}\left(\exists h, \exists \gamma, \exists \omega : P\left(\tilde{B}_{h,\gamma}, \tilde{A}_{h,\gamma}\right) - \hat{P}_n\left(B_{h,0}, A_{h,\gamma/2}\right) > \epsilon_1 \omega/2\right). \tag{A.40}$$

Let

$$\kappa(\gamma, \omega, \eta) := \frac{2r(k, k^*)\rho(k^*, \mathfrak{l}_0)}{\omega} \sqrt{\frac{2}{n} \ln\left(\frac{\mathcal{N}_\gamma}{\eta}\right)} \tag{A.41}$$

where ρ and r are defined in (3.53) and (3.13), and denote by

$$\omega_0 := \frac{\ell}{n}.$$

Henceforth, for conciseness, we write $\exists \omega$ for $\exists \omega \in (\omega_0, 1]$. Define

$$J(\gamma_1, \gamma_2, \eta) := \left\{ x^{(n)} : \exists h\, \exists \omega, P\left(\tilde{B}_{h,\gamma_2}, \tilde{A}_{h,\gamma_2}\right) > \hat{P}_n\left(B_{h,0}, A_{h,\gamma_1}\right) + \frac{\omega\kappa\left(\gamma_1, \frac{\omega}{2}, \frac{(\omega-\omega_0)\eta}{2}\right)}{2} \right\}. \tag{A.42}$$

For $\gamma_1 \leq \gamma \leq \gamma_2$, we claim $P\left(\tilde{B}_{h,\gamma_2}, \tilde{A}_{h,\gamma_2}\right) \leq P\left(\tilde{B}_{h,\gamma}, \tilde{A}_{h,\gamma}\right)$: first, $\tilde{B}_{h,\gamma_2} \subseteq \tilde{B}_{h,\gamma}$ as $\tilde{b}(\gamma_2) \geq \tilde{b}(\gamma)$ since $\tilde{b}$ is non-decreasing with γ. Secondly, $\tilde{A}_{h,\gamma_2} \subseteq \tilde{A}_{h,\gamma}$ because $\tilde{a}(\gamma_2) \geq \tilde{a}(\gamma)$ as $\tilde{a}$ is non-decreasing with γ. Hence, $\tilde{B}_{h,\gamma_2} \bigcap \tilde{A}_{h,\gamma_2} \subseteq \tilde{B}_{h,\gamma} \bigcap \tilde{A}_{h,\gamma}$ from which the claim follows. Next we claim $\hat{P}_n\left(B_{h,0}, A_{h,\gamma_1}\right) \geq \hat{P}_n\left(B_{h,0}, A_{h,\gamma}\right)$ as we next show: we have $a(\gamma) \geq a(\gamma_1)$ therefore $|f_h(s)| \geq a(\gamma)$ implies that $|f_h(s)| \geq a(\gamma_1)$ hence $A_{h,\gamma} \subseteq A_{h,\gamma_1}$. So we have $B_{h,0} \bigcap A_{h,\gamma} \subseteq B_{h,0} \bigcap A_{h,\gamma_1}$ and the claim follows. Also,

$$\kappa\left(\gamma, \frac{\omega}{2}, \frac{(\omega-\omega_0)\eta}{2}\right) \leq \kappa\left(\gamma_1, \frac{\omega}{2}, \frac{(\omega-\omega_0)\eta}{2}\right) \tag{A.43}$$

because $\mathcal{N}_\alpha$ is non-decreasing with decreasing α. It follows that

$$J\left(\gamma_1, \gamma_2, \eta\right) \subseteq J(\gamma, \gamma, \eta). \tag{A.44}$$

Let

$$M_{h,\gamma}\left(\omega_1, \omega_2, \eta\right) := \left\{x^{(n)} : P\left(\tilde{B}_{h,\gamma}, \tilde{A}_{h,\gamma}\right) > \hat{P}_n\left(B_{h,0}, A_{h,\gamma}\right) + \frac{\omega_2 \kappa\left(\gamma, \omega_1, \eta\right)}{2}\right\} \tag{A.45}$$

then for $\omega_1 \leq \omega \leq \omega_2$ we have

$$M_{h,\gamma}\left(\omega_1, \omega_2, \eta\right) \subseteq M_{h,\gamma}\left(\omega, \omega, \eta\right) \tag{A.46}$$

and for $\eta_a \leq \eta_b$,

$$M_{h,\gamma}(\omega, \omega, \eta_a) \subseteq M_{h,\gamma}(\omega, \omega, \eta_b) \tag{A.47}$$

which follows from the fact that $\kappa\left(\gamma, \omega_1, \eta\right) \geq \kappa\left(\gamma, \omega, \eta\right)$ and $\omega_2 \geq \omega$.

We have,

$$\mathbb{P}\left(J\left(\gamma, \gamma, \eta\right)\right) = \mathbb{P}\left(\left\{x^{(n)} : \exists h,\, x^{(n)} \in \bigcup_{\omega_0 \leq \omega \leq 1} M_{h,\gamma}\left(\frac{\omega}{2}, \omega, \frac{(\omega-\omega_0)\eta}{2}\right)\right\}\right). \tag{A.48}$$

Define the set $\Delta_l \subset [0,1]$ as follows,

$$\Delta_l = \left[\omega_0 + (1-\omega_0)\left(\frac{1}{2}\right)^{l+1}, \omega_0 + (1-\omega_0)\left(\frac{1}{2}\right)^{l}\right]. \tag{A.49}$$

We have $[\omega_0, 1] \subseteq \bigcup_{l=0}^{\infty} \Delta_l$. The right side of (A.48) is no larger than

$$\mathbb{P}\left(\exists h,\, X^{(n)} \in \bigcup_{l=0}^{\infty} \bigcup_{\omega \in \Delta_l} M_{h,\gamma}\left(\frac{\omega}{2}, \omega, \frac{(\omega-\omega_0)\eta}{2}\right)\right). \tag{A.50}$$

Now, for any $h \in \mathcal{H}$ there exists a $j \in \{1, \ldots, \mathcal{N}_\gamma\}$ such that (3.56) is satisfied with $\alpha = \gamma$. Define

$$M_{j,\gamma}\left(\omega_1, \omega_2, \eta\right) := \left\{x^{(n)} : P\left(B^{(\gamma)}_{j,\gamma} A^{(\gamma)}_{j,\gamma/4}\right) > \hat{P}_n\left(B^{(\gamma)}_{j,\gamma} A^{(\gamma)}_{j,\gamma/4}\right) + \frac{\omega_2 \kappa\left(\gamma, \omega_1, \eta\right)}{2}\right\}. \tag{A.51}$$

For the same reason that (A.46) and (A.47) hold, we have for $\omega_1 \leq \omega \leq \omega_2$, $\eta_a \leq \eta_b$,

$$M_{j,\gamma}\left(\omega_1, \omega_2, \eta\right) \subseteq M_{j,\gamma}\left(\omega, \omega, \eta\right) \tag{A.52}$$

and

$$M_{j,\gamma}\left(\omega, \omega, \eta_a\right) \subseteq M_{j,\gamma}\left(\omega, \omega, \eta_b\right).$$

Then from Claim 18 and Claim 19 it follows that if there exists an h such that $X^{(n)} \in M_{h,\gamma}\left(\omega_1, \omega_2, \eta\right)$ then there exists a $j \in \{1, \ldots, \mathcal{N}_\gamma\}$ such that $X^{(n)} \in M_{j,\gamma}\left(\omega_1, \omega_2, \eta\right)$. Hence if there exists $\omega \in [\omega_0, 1]$ such that there is an h with $X^{(n)} \in M_{h,\gamma}\left(\frac{\omega}{2}, \omega, \eta\right)$ then there exists a $j \in \{1, \ldots, \mathcal{N}_\gamma\}$ such that $X^{(n)} \in M_{j,\gamma}\left(\frac{\omega}{2}, \omega, \eta\right)$.

Therefore from (A.50) we have,

$$\begin{aligned}
\mathbb{P}\left(J\left(\gamma, \gamma, \eta\right)\right) &\leq \mathbb{P}\left(\exists 1 \leq j \leq \mathcal{N}_\gamma,\, X^{(n)} \in \bigcup_{l=0}^{\infty} \bigcup_{\omega \in \Delta_l} M_{j,\gamma}\left(\frac{\omega}{2}, \omega, \frac{(\omega-\omega_0)\eta}{2}\right)\right) \\
&\leq \sum_{j=1}^{\mathcal{N}_\gamma} \sum_{l=0}^{\infty} \mathbb{P}\left(M_{j,\gamma}\left(\omega_0 + (1-\omega_0)\left(\frac{1}{2}\right)^{l+1}, \omega_0 + (1-\omega_0)\left(\frac{1}{2}\right)^{l+1}, \frac{\eta}{2}(1-\omega_0)\left(\frac{1}{2}\right)^{l}\right)\right)
\end{aligned} \tag{A.53}$$

where the inequality follows from (A.52) as for every $\omega \in \Delta_l$ we have $\omega/2 \leq \omega_0 + (1-\omega_0)\left(\frac{1}{2}\right)^{l+1} \leq \omega$.

Let

$$g_{j,\gamma}^{(\gamma)}(\theta) := \mathbb{I}\left\{\left|f_j^{(\gamma)}(s)\right| \geq \hat{a}(\gamma/4),\ x f_j^{(\gamma)}(s) \geq \hat{b}(\gamma)\right\} \tag{A.54}$$

and

$$G_{j,\gamma}^{(\gamma)}(\Theta^{(n)}) := \sum_{t=1}^{n} g_{j,\gamma}^{(\gamma)}(\Theta_t).$$

Then

$$\begin{aligned} G_{j,\gamma}^{(\gamma)}(\Theta^{(n)}) &= \sum_{t=1}^{n} \mathbb{I}\left\{\left|f_j^{(\gamma)}\left(\langle\Theta_t\rangle_1^k\right)\right| \geq \hat{a}(\gamma/4),\ \langle\Theta_t\rangle_0 f_j^{(\gamma)}\left(\langle\Theta_t\rangle_1^k\right) \geq \hat{b}(\gamma)\right\} \\ &= n\hat{P}_n\left(A_{j,\gamma/4}^{(\gamma)}, B_{j,\gamma}^{(\gamma)}\right) \end{aligned} \tag{A.55}$$

and

$$\begin{aligned} \mathbb{E}\left[G_{j,\gamma}^{(\gamma)}\left(\Theta^{(n)}\right)\right] &= \mathbb{E}\left[n\hat{P}_n\left(A_{j,\gamma/4}^{(\gamma)}, B_{j,\gamma}^{(\gamma)}\right)\right] \\ &= \mathbb{E}\left[\sum_{l=1}^{n} \mathbb{I}\left\{\Theta_{t_l} \in A_{j,\gamma/4}^{(\gamma)} \bigcap B_{j,\gamma}^{(\gamma)}\right\}\right] \\ &= nP\left(A_{j,\gamma/4}^{(\gamma)} B_{j,\gamma}^{(\gamma)}\right) \end{aligned} \tag{A.56}$$

and therefore,

$$\mathbb{P}\left(M_{j,\gamma}\left(\omega,\omega,\eta\right)\right) = \mathbb{P}\left(\mathbb{E}\left[G_{j,\gamma}^{(\gamma)}\left(S^{*(n)}\right)\right] > G_{j,\gamma}^{(\gamma)}(S^{*(n)}) + \frac{n\omega\kappa\left(\gamma,\omega,\eta\right)}{2}\right). \tag{A.57}$$

We now show that the functions $G_{j,\gamma}^{(\gamma)}$ are Lipschitz. For two state sequences $s^{*(n)}$ and $u^{*(n)} \in \mathbb{S}_{k^*}^n$ we use the Hamming metric (A.2). Consider $s^{*(n)} = (s_1^*, \ldots, s_n^*)$ and $u^{*(n)} = (u_1^*, \ldots, u_n^*) \in \mathbb{S}_{k^*}^n$ such that

$$d_H(s^{*(n)}, u^{*(n)}) \leq \Delta$$

and define

$$\theta_t = (s_{t-r(k,k^*)+1}^*, \ldots, s_{t-1}^*, s_t^*)$$

and

$$\psi_t = (u_{t-r(k,k^*)+1}^*, \ldots, u_{t-1}^*, u_t^*).$$

Define the following subsets of $\{1, \ldots, n\}$,

$$\begin{aligned} I_1 &:= \left\{l : \left|f_j^{(\gamma)}(<\theta_{t_l}>_1^k)\right| < \hat{a}(\gamma/4), \left|f_j^{(\gamma)}(<\psi_{t_l}>_1^k)\right| < \hat{a}(\gamma/4)\right\} \\ I_2 &:= \left\{l : \left|f_j^{(\gamma)}(<\theta_{t_l}>_1^k)\right| < \hat{a}(\gamma/4), \left|f_j^{(\gamma)}(<\psi_{t_l}>_1^k)\right| \geq \hat{a}(\gamma/4)\right\} \\ I_3 &:= \left\{l : \left|f_j^{(\gamma)}(<\theta_{t_l}>_1^k)\right| \geq \hat{a}(\gamma/4), \left|f_j^{(\gamma)}(<\psi_{t_l}>_1^k)\right| < \hat{a}(\gamma/4)\right\} \\ I_4 &:= \left\{l : \left|f_j^{(\gamma)}(<\theta_{t_l}>_1^k)\right| \geq \hat{a}(\gamma/4), \left|f_j^{(\gamma)}(<\psi_{t_l}>_1^k)\right| \geq \hat{a}(\gamma/4)\right\}. \end{aligned} \tag{A.58}$$

Fix any $\gamma \in (0, \mathrm{diam}(\mathbb{S}_k)]$, $j \in \{1, \ldots, \mathcal{N}_\gamma\}$ we have

$$\begin{aligned}
\left|G_{j,\gamma}^{(\gamma)}(s^{*(n)}) - G_{j,\gamma}^{(\gamma)}(u^{*(n)})\right| &\leq \sum_{t=1}^{n} \left|g_{j,\gamma}^{(\gamma)}(\theta_t) - g_{j,\gamma}^{(\gamma)}(\psi_t)\right| \\
&= \sum_{l\in I_1} \left|g_{j,\gamma}^{(\gamma)}(\theta_{t_l}) - g_{j,\gamma}^{(\gamma)}(\psi_{t_l})\right| + \sum_{l\in I_2} \left|g_{j,\gamma}^{(\gamma)}(\theta_{t_l}) - g_{j,\gamma}^{(\gamma)}(\psi_{t_l})\right| + \sum_{l\in I_3} \left|g_{j,\gamma}^{(\gamma)}(\theta_{t_l}) - g_{j,\gamma}^{(\gamma)}(\psi_{t_l})\right| \\
&\quad + \sum_{l\in I_4} \left|g_{j,\gamma}^{(\gamma)}(\theta_{t_l}) - g_{j,\gamma}^{(\gamma)}(\psi_{t_l})\right| \\
&= \sum_{l\in I_2} \left|0 - \mathbb{I}\left\{< \psi_{t_l} >_0 f_j^{(\gamma)}(< \psi_{t_l} >_1^k) \geq \hat{b}(\gamma)\right\}\right| + \sum_{l\in I_3} \left|\mathbb{I}\left\{< \theta_{t_l} >_0 f_j^{(\gamma)}(< \theta_{t_l} >_1^k) \geq \hat{b}(\gamma)\right\} - 0\right| \\
&\quad + \sum_{l\in I_4} \left|\mathbb{I}\left\{< \theta_{t_l} >_0 f_j^{(\gamma)}(< \theta_{t_l} >_1^k) \geq \hat{b}(\gamma)\right\} - \mathbb{I}\left\{< \psi_{t_l} >_0 f_j^{(\gamma)}(< \psi_{t_l} >_1^k) \geq \hat{b}(\gamma)\right\}\right| \qquad \text{(A.59)} \\
&\leq \sum_{l\in I_2} \mathbb{I}\{\theta_{t_l} \neq \psi_{t_l}\} + \sum_{l\in I_3} \mathbb{I}\{\theta_{t_l} \neq \psi_{t_l}\} + \sum_{l\in I_4} \mathbb{I}\{\theta_{t_l} \neq \psi_{t_l}\} \\
&\leq \sum_{t=1}^{n} \mathbb{I}\{\theta_t \neq \psi_t\} \\
&\leq r(k, k^*) \sum_{t=1}^{n} \mathbb{I}\{s_t^* \neq q_t^*\} \qquad \text{(A.60)} \\
&= r(k, k^*) d_H(s^{*(n)}, q^{*(n)}) \\
&\leq r(k, k^*)\Delta \qquad \text{(A.61)}
\end{aligned}$$

where (A.60) holds since for every t such that $s_t^* \neq u_t^*$ (where recall, s_t^*, $u_t^* \in \mathbb{S}_{k^*}$) there are at most $r(k, k^*)$ time instants τ such that θ_τ contains state s_t^*, ψ_τ contains state u_t^* and $\theta_\tau \neq \psi_\tau$. For instance, let $n = 9$, $k = 3$, $k^* = 2$ and let the binary sequences that correspond to $s^{*(n)}$ and $u^{*(n)}$ be

$$\begin{aligned}
x^{(n)} &= -1 \quad 1 \quad 1 \quad -1 \quad 1 \quad -1 \quad -1 \quad 1 \quad 1 \quad -1 \quad -1 \quad -1 \\
\tilde{x}^{(n)} &= -1 \quad 1 \quad 1 \quad 1 \quad -1 \quad -1 \quad -1 \quad -1 \quad 1 \quad -1 \quad -1 \quad -1,
\end{aligned}$$

then $r(k, k^*) = 3$ and θ, $\psi \in \mathbb{S}_2^3$. We have $\sum_{t=1}^{n} \mathbb{I}\{s_t^* \neq u_t^*\} = 5$. Looking at subsequences of length $k+1 = 4$ (because θ and ψ are $k+1$ bit long) and comparing them across $x^{(n)}$ and $\tilde{x}^{(n)}$ over $1 \leq t \leq n$ we obtain $\sum_{t=1}^{n} \mathbb{I}\{\theta_t \neq \psi_t\} = 8$ and indeed $8 \leq r(3, 2) \cdot 5 = 15$. From (A.61) it follows that $G_{j,\gamma}^{(\gamma)}$ is Lipschitz with constant $r(k, k^*)$. So $G_{j,\gamma}^{(\gamma)}/r$ is Lipschitz with constant 1 and, from Section A.1, we may use (A.15) for $S^{*(n)}$ and $G_{j,\gamma}^{(\gamma)}/r$. We have

$$\begin{aligned}
\mathbb{P}\left(M_{j,\gamma}\left(\omega, \omega, \eta\right)\right) &\leq \mathbb{P}\left(\mathbb{E}\left[G_{j,\gamma}^{(\gamma)}(S^{*(n)})\right] > G_{j,\gamma}^{(\gamma)}\left(S^{*(n)}\right) + n\frac{\omega\kappa\left(\gamma, \omega, \eta\right)}{2}\right) \\
&= \mathbb{P}\left(\frac{\mathbb{E}\left[G_{j,\gamma}^{(\gamma)}(S^{*(n)})\right]}{r} > \frac{G_{j,\gamma}^{(\gamma)}\left(S^{*(n)}\right)}{r} + n\frac{\omega\kappa(\gamma, \omega, \eta)}{2r}\right) \\
&\leq \exp\left\{-\frac{n}{2}\left(\frac{\omega\kappa}{2\rho r}\right)^2\right\}. \qquad \text{(A.62)}
\end{aligned}$$

Plugging the choice of κ (A.41) in the right side of (A.62) gives

$$\mathbb{P}\left(M_{j,\gamma}\left(\omega, \omega, \eta\right)\right) \leq \frac{\eta}{\mathcal{N}_\gamma}. \qquad \text{(A.63)}$$

From (A.53) and (A.63), it follows that

$$\mathbb{P}\left(J\left(\gamma, \gamma, \eta\right)\right) \leq \sum_{j=1}^{\mathcal{N}\gamma} \frac{\eta}{2\mathcal{N}_\gamma}(1 - \omega_0) \sum_{l=0}^{\infty} \left(\frac{1}{2}\right)^l = \eta(1 - \omega_0). \qquad \text{(A.64)}$$

Now, substitute for ϵ_1 in (A.30) the following,

$$\epsilon_1 = \kappa\left(\frac{\gamma}{2}, \frac{\omega}{2}, \frac{(\omega-\omega_0)\gamma\eta}{2}\right). \tag{A.65}$$

Define the set $\Gamma_l \subset (0, \text{diam}\,(\mathbb{S}_k)]$ as follows,

$$\Gamma_l = \left[\left(\frac{1}{2}\right)^{l+1} \text{diam}\,(\mathbb{S}_k), \left(\frac{1}{2}\right)^{l} \text{diam}\,(\mathbb{S}_k)\right] \tag{A.66}$$

and the set $\bigcup_{l=0}^{\infty}\Gamma_l$ contains the possible range $(0, \text{diam}\,(\mathbb{S}_k)]$ for γ.

From (A.30), (A.40), (A.42) and (A.65), the probability in (A.30) is bounded from above by

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{\gamma\in(0,\text{diam}(\mathbb{S}_k)]} J\left(\frac{\gamma}{2},\gamma,\gamma\eta\right)\right) = \mathbb{P}\left(\bigcup_{l=0}^{\infty}\bigcup_{\gamma\in\Gamma_l} J\left(\frac{\gamma}{2},\gamma,\gamma\eta\right)\right)\\
&\leq \sum_{l=0}^{\infty}\mathbb{P}\left(\bigcup_{\gamma\in\Gamma_l} J\left(\frac{\gamma}{2},\gamma,\gamma\eta\right)\right)\\
&\leq \sum_{l=0}^{\infty}\mathbb{P}\left(J\left(\left(\frac{1}{2}\right)^{l+1}\text{diam}\,(\mathbb{S}_k), \left(\frac{1}{2}\right)^{l+1}\text{diam}\,(\mathbb{S}_k), \eta\left(\frac{1}{2}\right)^{l}\text{diam}\,(\mathbb{S}_k)\right)\right)
\end{aligned} \tag{A.67}$$

where (A.67) follows from (A.44) and the fact that for all $\gamma \in \Gamma_l$ we have $\gamma/2 \leq (\frac{1}{2})^{l+1}\text{diam}\,(\mathbb{S}_k) \leq \gamma$ and $\gamma \leq (\frac{1}{2})^{l}\text{diam}\,(\mathbb{S}_k)$.

From (A.64), the l^{th} term in (A.67) is bounded by $\eta(1-\omega_0)(1/2)^l\text{diam}\,(\mathbb{S}_k)$ and so (A.30) is bounded from above by

$$(1-\omega_0)\eta\text{diam}\,(\mathbb{S}_k)\sum_{l=0}^{\infty}\left(\frac{1}{2}\right)^{l} \leq 2(1-\omega_0)\eta\text{diam}\,(\mathbb{S}_k). \tag{A.68}$$

A.2.3. *Bounding (A.31).* We obtain a bound from above on the following probability,

$$\mathbb{P}\left(\exists h, \exists\gamma, \exists\omega:\ \hat{P}_n\left(A_{h,\gamma/2}\right) - P\left(\tilde{A}_{h,\gamma}\right) > \epsilon_1\hat{P}_n(A_{h,\gamma/2})/2P\left(\tilde{B}_{h,\gamma}\,\middle|\,\tilde{A}_{h,\gamma}\right),\ \nu > \omega n\right).$$

From (A.27), this is bounded from above by

$$\mathbb{P}\left(\exists h, \exists\gamma, \exists\omega:\ \hat{P}_n\left(A_{h,\gamma/2}\right) - P\left(\tilde{A}_{h,\gamma}\right) > \epsilon_1\omega/2P\left(\tilde{B}_{h,\gamma}\,\middle|\,\tilde{A}_{h,\gamma}\right)\right)$$

which in turn is bounded from above by

$$\mathbb{P}\left(\exists\gamma, \exists h, \exists\omega : \hat{P}_n(A_{h,\gamma/2}) - P(\tilde{A}_{h,\gamma}) > \epsilon_1\omega/2\right). \tag{A.69}$$

Define the set,

$$M'_{h,\gamma}(\omega_1,\omega_2,\eta) := \left\{x^{(n)} : \hat{P}_n\left(A_{h,\gamma/2}\right) > P\left(\tilde{A}_{h,\gamma}\right) + \frac{\omega_2\kappa(\gamma/2,\omega_1,\eta)}{2}\right\} \tag{A.70}$$

and let

$$M'_{j,\gamma}(\omega_1,\omega_2,\eta) := \left\{x^{(n)} : \hat{P}_n\left(A^{(\gamma)}_{j,2\gamma}\right) > P\left(A^{(\gamma)}_{j,\gamma/2}\right) + \frac{\omega_2\kappa(\gamma/2,\omega_1,\eta)}{2}\right\}.$$

From (A.20), Claim 20 and Claim 21, it follows that if there exists a γ and h such that $X^{(n)} \in M'_{h,\gamma}(\omega_1,\omega_2,\eta)$ then there exists a j such that $X^{(n)} \in M'_{j,\gamma}(\omega_1,\omega_2,\eta)$.

With the choice of ϵ_1 as in (A.65), then (A.69) becomes

$$\begin{aligned}
&\mathbb{P}\left(\exists\gamma,\ \exists h,\ \exists\omega : \hat{P}_n(A_{h,\gamma/2}) > P(\tilde{A}_{h,\gamma}) + \epsilon_1\omega/2\right)\\
&= \mathbb{P}\left(\exists\gamma,\ \exists h,\ \exists\omega : X^{(n)} \in M'_{h,\gamma}(\omega/2,\omega,(\omega-\omega_0)\gamma\eta/2)\right)\\
&\leq \mathbb{P}\left(\exists\gamma,\ \exists 1\leq j\leq\mathcal{N}_\gamma,\ \exists\omega:\ X^{(n)} \in M'_{j,\gamma}(\omega/2,\omega,(\omega-\omega_0)\gamma\eta/2)\right).
\end{aligned} \tag{A.71}$$

The probability in (A.71) is expressed as,

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{\gamma\in(0,\mathrm{diam}(\mathbb{S}_k)]}\left\{x^{(n)}:\exists 1\le j\le\mathcal{N}_\gamma,\exists\omega,\,x^{(n)}\in M'_{j,\gamma}\left(\omega/2,\omega,(\omega-\omega_0)\gamma\eta/2\right)\right\}\right)\\
&=\mathbb{P}\left(\bigcup_{l=0}^{\infty}\bigcup_{\gamma\in\Gamma_l}\left\{x^{(n)}:\exists 1\le j\le\mathcal{N}_\gamma,\exists\omega,\,x^{(n)}\in M'_{j,\gamma}\left(\omega/2,\omega,(\omega-\omega_0)\gamma\eta/2\right)\right\}\right)\\
&\le\sum_{l=0}^{\infty}\mathbb{P}\left(\bigcup_{\gamma\in\Gamma_l}\left\{x^{(n)}:\exists 1\le j\le\mathcal{N}_\gamma,\exists\omega,\,x^{(n)}\in M'_{j,\gamma}\left(\omega/2,\omega,(\omega-\omega_0)\gamma\eta/2\right)\right\}\right). \qquad \text{(A.72)}
\end{aligned}$$

Denote by

$$\varrho_l:=(1/2^l)\mathrm{diam}(\mathbb{S}_k). \qquad \text{(A.73)}$$

For every $\gamma\in\Gamma_l$, we have $\gamma\le\varrho_l\le 2\gamma$ hence for any

$$\theta=[s,x]\in A^{(\gamma)}_{j,2\gamma} \qquad \text{(A.74)}$$

we have

$$\begin{aligned}
\left|f_j^{(\gamma)}(s)\right| &\ge \hat{a}(2\gamma)\\
&\ge \hat{a}\left(\varrho_l\right)
\end{aligned}$$

because $\hat{a}$ is non-decreasing.

Fix any h and let $f_i^{(\varrho_l)}$ be an element in the ϱ_l-cover C_{ϱ_l} of $\mathcal{F}$ that is closest to f_h in the l_∞-norm (3.54) and let $f_j^{(\gamma)}$ be the closest to f_h in the γ-cover C_γ. Denote by

$$\dot{A}^{(\gamma)}_{i,\gamma'}:=\left\{\theta=[s,x]:\left|f_i^{(\gamma)}(s)\right|\ge\dot{a}(\gamma')\right\}. \qquad \text{(A.75)}$$

Then, for any $\gamma\in\Gamma_l$ and any $\theta=[s,x]\in A^{(\gamma)}_{j,2\gamma}$, if $f_j^{(\gamma)}(s)\le-\hat{a}(2\gamma)$ then

$$\begin{aligned}
f_i^{(\varrho_l)}(s) &\le f_h(s)+\varrho_l\\
&\le f_j^{(\gamma)}(s)+\varrho_l+\gamma\\
&\overset{(i)}{\le} -\hat{a}(2\gamma)+\varrho_l+\gamma\\
&\overset{(ii)}{\le} -\hat{a}(\varrho_l)+\varrho_l+\gamma\\
&\overset{(iii)}{\le} -\hat{a}(\varrho_l)+2\varrho_l\\
&\overset{(iv)}{\le} -\dot{a}(\varrho_l) \qquad \text{(A.76)}
\end{aligned}$$

where (i) follows from the definition of the set $A^{(\gamma)}_{j,2\gamma}$, (ii) is from the fact that $\hat{a}$ is a non-decreasing function and $\varrho_l\le 2\gamma$ for all $\gamma\in\Gamma_l$, (iii) follows from $\gamma\le\varrho_l$ for $\gamma\in\Gamma_l$, and (iv) follows from (A.20) where $-\hat{a}(\gamma)+2\gamma\le-\dot{a}(\gamma)$ holds in particular for $\gamma=\varrho_l$. If $f_j^{(\gamma)}(s)\ge\hat{a}(2\gamma)$ then

$$\begin{aligned}
f_i^{(\varrho_l)}(s) &\ge f_h(s)-\varrho_l\\
&\ge f_j^{(\gamma)}(s)-\varrho_l-\gamma\\
&\ge \hat{a}(2\gamma)-\varrho_l-\gamma\\
&\ge \hat{a}(\varrho_l)-\varrho_l-\gamma\\
&\ge \hat{a}(\varrho_l)-2\varrho_l\\
&\ge \dot{a}(\varrho_l). \qquad \text{(A.77)}
\end{aligned}$$

Hence from (A.75), (A.76), and (A.77) it follows that $\theta\in\dot{A}^{(\varrho_l)}_{i,\varrho_l}$. Therefore for all $\gamma\in\Gamma_l$, we have

$$A^{(\gamma)}_{j,2\gamma}\subseteq\dot{A}^{(\varrho_l)}_{i,\varrho_l}. \qquad \text{(A.78)}$$

Now consider any $\theta \in A_{i,\varrho_l}^{(\varrho_l)}$. If $f_i^{(\varrho_l)}(s) \leq -\dot{a}(\varrho_l)$ then for every $\gamma \in \Gamma_l$ we have

$$\begin{aligned} f_j^{(\gamma)}(s) &\leq f_h(s) + \gamma \\ &\leq f_i^{(\varrho_l)}(s) + \gamma + \varrho_l \qquad (A.79) \\ &\overset{(i)}{\leq} -\dot{a}(\varrho_l) + \gamma + \varrho_l \\ &\overset{(ii)}{\leq} -\dot{a}(\gamma) + \gamma + \varrho_l \\ &\overset{(iii)}{\leq} -\dot{a}(\gamma) + 3\gamma \\ &\overset{(iv)}{\leq} -\hat{a}(\gamma/2) \qquad (A.80) \end{aligned}$$

where (i) follows from (A.75), (ii) and (iii) are from the fact that for all $\gamma \in \Gamma_l$, $\gamma \leq \varrho_l$, and $\varrho_l \leq 2\gamma$, respectively, and (iv) follows from (A.20). If $f_i^{(\varrho_l)}(s) \geq \dot{a}(\varrho_l)$ then for every $\gamma \in \Gamma_l$ we have

$$\begin{aligned} f_j^{(\gamma)}(s) &\geq f_h(s) - \gamma \\ &\geq f_i^{(\varrho_l)}(s) - \gamma - \varrho_l \qquad (A.81) \\ &\geq \dot{a}(\varrho_l) - \gamma - \varrho_l \\ &\geq \dot{a}(\gamma) - \gamma - \varrho_l \\ &\geq \dot{a}(\gamma) - 3\gamma \\ &\geq \hat{a}(\gamma/2). \qquad (A.82) \end{aligned}$$

From (A.80) and (A.82) it follows that

$$\dot{A}_{i,\varrho_l}^{(\varrho_l)} \subseteq A_{j,\gamma/2}^{(\gamma)}. \qquad (A.83)$$

Define by

$$\dot{M}_{i,l}(\omega_1, \omega_2, \eta) := \left\{ x^{(n)} : \hat{P}_n\left(\dot{A}_{i,\varrho_l}^{(\varrho_l)}\right) > P\left(\dot{A}_{i,\varrho_l}^{(\varrho_l)}\right) + \frac{\omega_2 \kappa(\varrho_l, \omega_1, \eta)}{2} \right\}. \qquad (A.84)$$

Then it follows from (A.78), (A.83) and from the fact that

$$\kappa(\gamma/2, \omega_1, \eta) \geq \kappa(\gamma, \omega_1, \eta) \geq \kappa(\varrho_l, \omega_1, \eta)$$

that for any $0 \leq l \leq \infty$ and any $\gamma \in \Gamma_l$, for all $h \in \mathcal{H}$ there exists $i \in \{1, \ldots, \mathcal{N}_{\varrho_l}\}$ and $j \in \{1, \ldots, \mathcal{N}_\gamma\}$ (both depending on h) such that the following holds,

$$M'_{j,\gamma}(\omega_1, \omega_2, \eta) \subseteq \dot{M}_{i,l}(\omega_1, \omega_2, \eta).$$

Therefore, the l^{th} term in the sum (A.72) is bounded from above as follows,

$$\begin{aligned} &\mathbb{P}\left(\bigcup_{\gamma \in \Gamma_l} \left\{ x^{(n)} : \exists 1 \leq j \leq \mathcal{N}_\gamma, \exists \omega, x^{(n)} \in M'_{j,\gamma}(\omega/2, \omega, (\omega - \omega_0)\gamma\eta/2) \right\} \right) \\ &\quad \leq \mathbb{P}\left(\left\{ x^{(n)} : \exists 1 \leq i \leq \mathcal{N}_{\varrho_l}, \exists \omega, x^{(n)} \in \dot{M}_{i,l}(\omega/2, \omega, (\omega - \omega_0)\varrho_l\eta/2) \right\} \right) \qquad (A.85) \end{aligned}$$

where we also used the fact that for $\gamma \in \Gamma_l$, $\gamma \leq \varrho_l$.

Now, for $\omega_1 \leq \omega \leq \omega_2$ we have

$$\dot{M}_{i,l}(\omega_1, \omega_2, \eta) \subseteq \dot{M}_{i,l}(\omega, \omega, \eta) \qquad (A.86)$$

which follows from the fact that $\kappa(\varrho_l, \omega_1, \eta) \geq \kappa(\varrho_l, \omega, \eta)$ and $\omega_2 \geq \omega$. Let us fix i. Using the sets (A.49) we have

$$\begin{aligned} &\mathbb{P}\left(\bigcup_{\omega \in [\omega_0, 1]} \dot{M}_{i,l}(\omega/2, \omega, (\omega - \omega_0)\varrho_l\eta/2) \right) \\ &= \mathbb{P}\left(\bigcup_{l'=0}^{\infty} \bigcup_{\omega \in \Delta_{l'}} \dot{M}_{i,l}(\omega/2, \omega, (\omega - \omega_0)\varrho_l\eta/2) \right) \\ &\leq \sum_{l'=0}^{\infty} \mathbb{P}\left(\dot{M}_{i,l}\left(\omega_0 + (1 - \omega_0)\left(\frac{1}{2}\right)^{l'+1}, \omega_0 + (1 - \omega_0)\left(\frac{1}{2}\right)^{l'+1}, \frac{\eta\varrho_l}{2}(1 - \omega_0)\left(\frac{1}{2}\right)^{l'} \right) \right) \qquad (A.87) \end{aligned}$$

where the inequality follows from (A.86) as for every $\omega \in \Delta_{l'}$ we have $\omega/2 \leq \omega_0 + (1-\omega_0)\left(\frac{1}{2}\right)^{l'+1} \leq \omega$. Define

$$q_{i,l}(\theta) := \mathbb{I}\left\{\left|f_i^{(\varrho_l)}(s)\right| \geq \dot{a}(\varrho_l)\right\} \tag{A.88}$$

and

$$Q_{i,l}(\Theta^{(n)}) := \sum_{t=1}^{n} q_{i,l}(\Theta_t).$$

Then,

$$\hat{P}_n\left(\dot{A}_{i,\varrho_l}^{(\varrho_l)}\right) = \frac{1}{n} Q_{i,l}\left(\Theta^{(n)}\right)$$

and

$$\begin{aligned}\frac{1}{n}\mathbb{E}\left[Q_{i,l}\left(\Theta^{(n)}\right)\right] &= \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\left[q_{i,l}\left(\Theta_t\right)\right] \\ &= \mathbb{P}\left(\dot{A}_{i,\varrho_l}^{(\varrho_l)}\right).\end{aligned}$$

As in (A.57), we obtain

$$\mathbb{P}\left(\dot{M}_{i,l}\left(\omega,\omega,\eta\right)\right) = \mathbb{P}\left(Q_{i,l}(S^{*(n)}) > \mathbb{E}\left[Q_{i,l}\left(S^{*(n)}\right)\right] + \frac{n\omega\kappa\left(\varrho_l,\omega,\eta\right)}{2}\right). \tag{A.89}$$

Replace $\hat{a}(\gamma/4)$, $f_j^{(\gamma)}$ in the sets (A.58) by $\dot{a}(\varrho_l)$, $f_i^{(\varrho_l)}$ then using the analysis (A.59)-(A.61), it follows that the functions $Q_{i,l}$ are Lipschitz with constant $r(k,k^*)$ (the bound there is actually looser in this case because the sum over the region I_4 vanishes). Then from the argument that leads to (A.63) it follows that (A.89) is bounded from above by $\eta/\mathcal{N}_{\varrho_l}$.

Therefore (A.87) is bounded from above by

$$\frac{\eta\varrho_l}{2\mathcal{N}_{\varrho_l}}(1-\omega_0)\sum_{l'=0}^{\infty}\left(\frac{1}{2}\right)^{l'} = \frac{\eta(1-\omega_0)\varrho_l}{\mathcal{N}_{\varrho_l}}.$$

Hence, (A.85) is bounded from above as follows,

$$\begin{aligned}&\sum_{i=1}^{\mathcal{N}_{\varrho_l}}\mathbb{P}\left(\exists\omega,\, X^{(n)} \in \dot{M}_{i,l}\left(\omega/2,\omega,(\omega-\omega_0)\varrho_l\eta/2\right)\right) \\ &\quad\leq \sum_{i=1}^{\mathcal{N}_{\varrho_l}}\sum_{l'=0}^{\infty}\mathbb{P}\left(\dot{M}_{i,l}\left(\omega_0+(1-\omega_0)\left(\frac{1}{2}\right)^{l'+1},\omega_0+(1-\omega_0)\left(\frac{1}{2}\right)^{l'+1},\frac{\eta\varrho_l}{2}(1-\omega_0)\left(\frac{1}{2}\right)^{l'}\right)\right) \\ &\quad\leq \sum_{i=1}^{\mathcal{N}_{\varrho_l}}\frac{\eta(1-\omega_0)\varrho_l}{\mathcal{N}_{\varrho_l}} \\ &\quad= \eta(1-\omega_0)\varrho_l.\end{aligned}$$

Substituting for ϱ_l according to (A.73) then it follows that (A.72) and, therefore, the left side of (A.71) and (A.31) are bounded from above by

$$\eta(1-\omega_0)\mathrm{diam}(\mathbb{S}_k)\sum_{l=0}^{\infty}\frac{1}{2^l} = 2\eta(1-\omega_0)\mathrm{diam}(\mathbb{S}_k).$$

Thus, with the conclusion of Section A.2.2, namely (A.68), we conclude that the sum of (A.30) and (A.31) is bounded from above by

$$4\eta(1-\omega_0)\mathrm{diam}\left(\mathbb{S}_k\right). \tag{A.90}$$

A.2.4. *Bounding (A.32).* We wish to bound from above (A.32), that is,

$$\mathbb{P}\left(\exists h, \exists \gamma : \tilde{L}^{(\gamma)}(h|\gamma) - L_m^{(b(2\gamma))}(h) > \epsilon_2\right). \tag{A.91}$$

This is done following the same steps taken in the previous section. The condition in (A.91) equals

$$\left(1 - \overline{\tilde{L}}^{(\gamma)}(h|\gamma)\right) - \left(1 - \overline{L}_m^{(b(2\gamma))}(h)\right) > \epsilon_2$$

which equals

$$\overline{L}_m^{(b(2\gamma))}(h) - \overline{\tilde{L}}^{(\gamma)}(h|\gamma) > \epsilon_2.$$

From (A.25), the right side is expressed as

$$\mathbb{P}\left(\overline{L}_m^{(b(2\gamma))}(h) - P\left(\tilde{B}_{h,\gamma}\middle|\tilde{A}_{h,\gamma}\right) > \epsilon_2\right). \tag{A.92}$$

We have,

$$\begin{aligned}\overline{L}_m^{(b(2\gamma))}(h) &= \frac{1}{m}\sum_{t=1}^{m}\mathbb{I}\{X_t f_h(S_{t-1}) \geq b(\gamma)\} \\ &= \frac{1}{\nu_0}\sum_{l:|f_h(S_{t_l-1})|\geq 0}\mathbb{I}\{X_t f_h(S_{t-1}) \geq b(\gamma)\}\end{aligned}$$

where ν_0 is defined as the number of times that the sequence $S^{(m)}$ enters a state s such that $|f_h(s)| \geq 0$, which obviously is always true, therefore

$$\nu_0 = m \tag{A.93}$$

(we choose this alternate representation in order to follow the same notation and steps as in the previous section for ease of understanding). Define

$$A_{h,0} := \{\theta : |f_h(S_{t_l-1})| \geq 0\}.$$

Denote by

$$\begin{aligned}\hat{P}_m(A_{h,0}) &= \frac{1}{m}\sum_{t=1}^{m}\mathbb{I}\{\Theta_t \in A_{h,0}\} \\ &= \frac{\nu_0}{m} = 1.\end{aligned} \tag{A.94}$$

Thus the expression inside the probability in (A.92) equals

$$\begin{aligned}\overline{L}_m^{(b(2\gamma))}(h) - P\left(\tilde{B}_{h,\gamma}\middle|\tilde{A}_{h,\gamma}\right) &= \frac{1}{\hat{P}_m(A_{h,0})}\left(\overline{L}_m^{(b(2\gamma))}(h)\hat{P}_m(A_{h,0}) - P\left(\tilde{B}_{h,\gamma},\tilde{A}_{h,\gamma}\right)\right) \\ &\quad + P\left(\tilde{B}_{h,\gamma},\tilde{A}_{h,\gamma}\right)\left(\frac{1}{\hat{P}_m(A_{h,0})} - \frac{1}{P\left(\tilde{A}_{h,\gamma}\right)}\right).\end{aligned} \tag{A.95}$$

For the second term of (A.95) we have,

$$\begin{aligned}&P\left(\tilde{B}_{h,\gamma},\tilde{A}_{h,\gamma}\right)\left(\frac{1}{\hat{P}_m(A_{h,0})} - \frac{1}{P\left(\tilde{A}_{h,\gamma}\right)}\right) \\ &= P\left(\tilde{B}_{h,\gamma}\middle|\tilde{A}_{h,\gamma}\right)P\left(\tilde{A}_{h,\gamma}\right)\left(\frac{P\left(\tilde{A}_{h,\gamma}\right) - \hat{P}_m(A_{h,0})}{\hat{P}_m(A_{h,0})P\left(\tilde{A}_{h,\gamma}\right)}\right) \\ &= \frac{P\left(\tilde{B}_{h,\gamma}\middle|\tilde{A}_{h,\gamma}\right)}{\hat{P}_m(A_{h,0})}\left(P\left(\tilde{A}_{h,\gamma}\right) - \hat{P}_m(A_{h,0})\right).\end{aligned}$$

Thus (A.92) is bounded from above by

$$\begin{aligned}&\mathbb{P}\left(\overline{L}_m^{(b(2\gamma))}(h)\hat{P}_m(A_{h,0}) - P\left(\tilde{B}_{h,\gamma},\tilde{A}_{h,\gamma}\right) > \epsilon_2\hat{P}_m(A_{h,0})/2\right) \\ &+ \mathbb{P}\left(P(\tilde{A}_{h,\gamma}) - \hat{P}_m(A_{h,0}) > \epsilon_2\hat{P}_m(A_{h,0})/2P\left(\tilde{B}_{h,\gamma}\middle|\tilde{A}_{h,\gamma}\right)\right).\end{aligned} \tag{A.96}$$

We have,

$$\overline{L}_m^{(b(2\gamma))}(h)\hat{P}_m\left(A_{h,0}\right) = \hat{P}_m\left(B_{h,2\gamma}, A_{h,0}\right)$$

thus (A.91) is bounded from above by

$$\mathbb{P}\left(\exists h, \exists\gamma : \hat{P}_m\left(B_{h,2\gamma}, A_{h,0}\right) - P\left(\tilde{B}_{h,\gamma}, \tilde{A}_{h,\gamma}\right) > \epsilon_2\hat{P}_m(A_{h,0})/2\right) \tag{A.97}$$

$$+ \quad \mathbb{P}\left(\exists h, \exists\gamma : P\left(\tilde{A}_{h,\gamma}\right) - \hat{P}_m\left(A_{h,0}\right) > \epsilon_2\hat{P}_m(A_{h,0})/2P\left(\tilde{B}_{h,\gamma}\left|\tilde{A}_{h,\gamma}\right.\right)\right). \tag{A.98}$$

Denote by

$$\dot{B}_{j,\gamma'}^{(\gamma)} := \left\{\theta : xf_j^{(\gamma)}(s) \geq \dot{b}(\gamma')\right\}.$$

We have the following.

Claim 22. For h and h_j such that f_h and f_j satisfy (3.56) with $\alpha = \gamma$, we have

$$\hat{P}_m\left(A_{h,0}, B_{h,2\gamma}\right) \leq \hat{P}_m\left(A_{j,\gamma}^{(\gamma)}, \dot{B}_{j,\gamma}^{(\gamma)}\right). \tag{A.99}$$

Proof. Consider $\theta = [s, x] \in A_{h,0} \bigcap B_{h,2\gamma}$. If $x = -1$, then $f_h(s) \leq -b(2\gamma)$ so

$$\begin{aligned} f_j^{(\gamma)}(s) &\leq f_h(s) + \gamma \\ &\leq -b(2\gamma) + \gamma \\ &\leq -\dot{b}(\gamma) \end{aligned}$$

where the last inequality follows from (A.20). If $x = 1$ then $f_h(s) \geq b(2\gamma)$ so

$$\begin{aligned} f_j^{(\gamma)}(s) &\geq f_h(s) - \gamma \\ &\geq b(2\gamma) - \gamma \\ &\geq \dot{b}(\gamma). \end{aligned}$$

Therefore $\theta \in \dot{B}_{h,\gamma}^{(\gamma)}$. And $\dot{b}(\gamma) \geq \hat{a}(\gamma)$ so $\theta \in A_{j,\gamma}^{(\gamma)}$ and therefore $\theta \in A_{j,\gamma}^{(\gamma)} \bigcap \dot{B}_{j,\gamma}^{(\gamma)}$. Hence $A_{h,0} \bigcap B_{h,2\gamma} \subseteq A_{j,\gamma}^{(\gamma)} \bigcap \dot{B}_{j,\gamma}^{(\gamma)}$. □

We have the following.

Claim 23. For h and h_j such that f_h and f_j satisfy (3.56) with $\alpha = \gamma$, we have

$$P\left(A_{j,\gamma}^{(\gamma)}, \dot{B}_{j,\gamma}^{(\gamma)}\right) \leq P\left(\tilde{A}_{h,\gamma}, \tilde{B}_{h,\gamma}\right). \tag{A.100}$$

Proof. Consider $\theta = [s, x] \in A_{j,\gamma}^{(\gamma)} \bigcap \dot{B}_{j,\gamma}^{(\gamma)}$. If $f_j^{(\gamma)}(s) \leq -\hat{a}(\gamma)$ then

$$\begin{aligned} f_h(s) &\leq f_j^{(\gamma)}(s) + \gamma \\ &\leq -\hat{a}(\gamma) + \gamma \\ &\leq -\tilde{a}(\gamma) \end{aligned}$$

where the last inequality follows from (A.20). If $f_j^{(\gamma)}(s) \geq \hat{a}(\gamma)$ then

$$\begin{aligned} f_h(s) &\geq f_j^{(\gamma)}(s) - \gamma \\ &\geq \hat{a}(\gamma) - \gamma \\ &\geq -\tilde{a}(\gamma) \end{aligned}$$

therefore $A_{j,\gamma}^{(\gamma)} \subseteq \tilde{A}_{h,\gamma}$. From (A.20), we have $\dot{b}(\gamma) \geq \tilde{b}(\gamma)$ therefore $\dot{B}_{j,\gamma}^{(\gamma)} \subseteq \tilde{B}_{h,\gamma}$. Hence, $A_{j,\gamma}^{(\gamma)} \bigcap \dot{B}_{j,\gamma}^{(\gamma)} \subseteq \tilde{A}_{h,\gamma} \bigcap \tilde{B}_{h,\gamma}$. □

A.2.5. *Bounding (A.97).* From (A.93) and (A.94) it follows that $\hat{P}_m(A_{h,0}) = 1$ therefore (A.97) is expressed as follows,

$$\mathbb{P}\left(\exists h, \exists \gamma : \hat{P}_m\left(B_{h,2\gamma}, A_{h,0}\right) - P\left(\tilde{B}_{h,\gamma}, \tilde{A}_{h,\gamma}\right) > \epsilon_2/2\right). \tag{A.101}$$

Let

$$\kappa\left(\gamma, \eta\right) := 2r(k, k^*)\rho(k^*, \mathfrak{l}_0)\sqrt{\frac{2}{m}\ln\left(\frac{\mathcal{N}_\gamma}{\eta}\right)} \tag{A.102}$$

where ρ and r are defined in (3.53) and (3.13). Define

$$J(\gamma_1, \gamma_2, \eta) := \left\{x^{(m)} : \exists h,\ \hat{P}_m\left(B_{h,\gamma_2}, A_{h,0}\right) > P\left(\tilde{B}_{h,\gamma_1}, \tilde{A}_{h,\gamma_1}\right) + \frac{\kappa\left(\gamma_1, \eta\right)}{2}\right\}. \tag{A.103}$$

For $\gamma_1 \leq \gamma \leq \gamma_2$, we claim $\hat{P}_m\left(B_{h,\gamma_2}, A_{h,0}\right) \leq \hat{P}_m\left(B_{h,\gamma}, A_{h,0}\right)$. We have $B_{h,\gamma_2} \subseteq B_{h,\gamma}$ as $b(\gamma_2) \geq b(\gamma)$ since b is non-decreasing with γ. Hence, $B_{h,\gamma_2} \bigcap A_{h,0} \subseteq B_{h,\gamma} \bigcap A_{h,0}$ from which the claim follows. Next we claim that $P\left(\tilde{B}_{h,\gamma_1}, \tilde{A}_{h,\gamma_1}\right) \geq P\left(\tilde{B}_{h,\gamma}, \tilde{A}_{h,\gamma}\right)$ as we next show: we have $\tilde{a}(\gamma) \geq \tilde{a}(\gamma_1)$ therefore $|f_h(s)| \geq \tilde{a}(\gamma)$ implies that $|f_h(s)| \geq \tilde{a}(\gamma_1)$ hence $\tilde{A}_{h,\gamma} \subseteq \tilde{A}_{h,\gamma_1}$. Also, $\tilde{b}(\gamma) \geq \tilde{b}(\gamma_1)$ so $|f_h(s)| \geq \tilde{b}(\gamma)$ implies that $|f_h(s)| \geq \tilde{b}(\gamma_1)$ hence $\tilde{B}_{h,\gamma} \subseteq \tilde{B}_{h,\gamma_1}$. So we have $\tilde{B}_{h,\gamma} \bigcap \tilde{A}_{h,\gamma} \subseteq \tilde{B}_{h,\gamma_1} \bigcap \tilde{A}_{h,\gamma_1}$ and the claim follows. Also,

$$\kappa\left(\gamma, \eta\right) \leq \kappa\left(\gamma_1, \eta\right) \tag{A.104}$$

because $\mathcal{N}_\alpha$ is non-decreasing with decreasing α. It follows that

$$J\left(\gamma_1, \gamma_2, \eta\right) \subseteq J(\gamma, \gamma, \eta). \tag{A.105}$$

Define

$$M_{h,\gamma}(\eta) := \left\{x^{(m)} : \hat{P}_m\left(B_{h,2\gamma}, A_{h,0}\right) > P\left(\tilde{B}_{h,\gamma}, \tilde{A}_{h,\gamma}\right) + \frac{\kappa\left(\gamma, \eta\right)}{2}\right\}$$

and

$$M'_{j,\gamma}(\eta) := \left\{x^{(m)} : \hat{P}_m\left(A^{(\gamma)}_{j,\gamma}, \dot{B}^{(\gamma)}_{j,\gamma}\right) > P\left(A^{(\gamma)}_{j,\gamma}, \dot{B}^{(\gamma)}_{j,\gamma}\right) + \frac{\kappa\left(\gamma, \eta\right)}{2}\right\}.$$

From Claim 22 and Claim 23 it follows that if there exists an h and γ such that $X^{(m)} \in M_{h,\gamma}$ then there exists a $j \in \{1, \ldots, \mathcal{N}_\gamma\}$ such that $X^{(m)} \in M'_{j,\gamma}$. Therefore we have,

$$\mathbb{P}\left(J\left(\gamma, \gamma, \eta\right)\right) \leq \mathbb{P}\left(\exists 1 \leq j \leq \mathcal{N}_\gamma,\ X^{(m)} \in M'_{j,\gamma}\left(\eta\right)\right) \leq \sum_{j=1}^{\mathcal{N}_\gamma} \mathbb{P}\left(M'_{j,\gamma}(\eta)\right). \tag{A.106}$$

Let

$$g^{(\gamma)}_{j,\gamma}(\theta) := \mathbb{I}\left\{\left|f^{(\gamma)}_j(s)\right| \geq \hat{a}(\gamma),\ x f^{(\gamma)}_j(s) \geq \dot{b}(\gamma)\right\} \tag{A.107}$$

and

$$G^{(\gamma)}_{j,\gamma}(\Theta^{(n)}) := \sum_{t=1}^{n} g^{(\gamma)}_{j,\gamma}(\Theta_t).$$

Then

$$\begin{aligned} G^{(\gamma)}_{j,\gamma}(\Theta^{(m)}) &= \sum_{t=1}^{m} \mathbb{I}\left\{\left|f^{(\gamma)}_j\left(\langle\Theta_t\rangle^k_1\right)\right| \geq \hat{a}(\gamma),\ \langle\Theta_t\rangle_0 f^{(\gamma)}_j\left(\langle\Theta_t\rangle^k_1\right) \geq \dot{b}(\gamma)\right\} \\ &= m\hat{P}_m\left(A^{(\gamma)}_{j,\gamma}, \dot{B}^{(\gamma)}_{j,\gamma}\right) \end{aligned} \tag{A.108}$$

and

$$\begin{aligned} \mathbb{E}\left[G^{(\gamma)}_{j,\gamma}\left(\Theta^{(m)}\right)\right] &= \mathbb{E}\left[m\hat{P}_m\left(A^{(\gamma)}_{j,\gamma}, \dot{B}^{(\gamma)}_{j,\gamma}\right)\right] \\ &= m\mathbb{P}\left(A^{(\gamma)}_{j,\gamma}, \dot{B}^{(\gamma)}_{j,\gamma}\right) \end{aligned} \tag{A.109}$$

and hence we express the probability of the event $M'_{j,\gamma}$ as follows,

$$\mathbb{P}\left(M'_{j,\gamma}\left(\eta\right)\right) = \mathbb{P}\left(G^{(\gamma)}_{j,\gamma}(S^{*(m)}) > \mathbb{E}\left[G^{(\gamma)}_{j,\gamma}\left(S^{*(m)}\right)\right] + \frac{m\kappa\left(\gamma, \eta\right)}{2}\right). \tag{A.110}$$

We now show that the functions $G_{j,\gamma}^{(\gamma)}$ are Lipschitz. For two state sequences $s^{*(m)}$ and $u^{*(m)} \in \mathbb{S}_{k^*}^m$ we use the Hamming metric (A.2). Consider $s^{*(m)} = (s_1^*, \ldots, s_m^*)$ and $u^{*(m)} = (u_1^*, \ldots, u_m^*) \in \mathbb{S}_{k^*}^m$ such that

$$d_H(s^{*(m)}, u^{*(m)}) \quad \leq \quad \Delta$$

and define

$$\theta_t \quad = \quad (s^*_{t-r(k,k^*)+1}, \ldots, s^*_{t-1}, s^*_t)$$

and

$$\psi_t \quad = \quad (u^*_{t-r(k,k^*)+1}, \ldots, u^*_{t-1}, u^*_t).$$

Define the following subsets of $\{1, \ldots, m\}$,

$$\begin{aligned}
I_1 &:= \left\{l : \left|f_j^{(\gamma)}(< \theta_{t_l} >_1^k)\right| < \hat{a}(\gamma), \left|f_j^{(\gamma)}(< \psi_{t_l} >_1^k)\right| < \hat{a}(\gamma)\right\} \\
I_2 &:= \left\{l : \left|f_j^{(\gamma)}(< \theta_{t_l} >_1^k)\right| < \hat{a}(\gamma), \left|f_j^{(\gamma)}(< \psi_{t_l} >_1^k)\right| \geq \hat{a}(\gamma)\right\} \\
I_3 &:= \left\{l : \left|f_j^{(\gamma)}(< \theta_{t_l} >_1^k)\right| \geq \hat{a}(\gamma), \left|f_j^{(\gamma)}(< \psi_{t_l} >_1^k)\right| < \hat{a}(\gamma)\right\} \\
I_4 &:= \left\{l : \left|f_j^{(\gamma)}(< \theta_{t_l} >_1^k)\right| \geq \hat{a}(\gamma), \left|f_j^{(\gamma)}(< \psi_{t_l} >_1^k)\right| \geq \hat{a}(\gamma)\right\}.
\end{aligned} \tag{A.111}$$

Fix any $\gamma \in (0, \mathrm{diam}(\mathbb{S}_k)]$, $j \in \{1, \ldots, \mathcal{N}_\gamma\}$ we have

$$\begin{aligned}
&\left|G_{j,\gamma}^{(\gamma)}(s^{*(m)}) - G_{j,\gamma}^{(\gamma)}(u^{*(m)})\right| \leq \sum_{t=1}^m \left|g_{j,\gamma}^{(\gamma)}(\theta_t) - g_{j,\gamma}^{(\gamma)}(\psi_t)\right| \\
&= \sum_{l\in I_1} \left|g_{j,\gamma}^{(\gamma)}(\theta_{t_l}) - g_{j,\gamma}^{(\gamma)}(\psi_{t_l})\right| + \sum_{l\in I_2} \left|g_{j,\gamma}^{(\gamma)}(\theta_{t_l}) - g_{j,\gamma}^{(\gamma)}(\psi_{t_l})\right| + \sum_{l\in I_3} \left|g_{j,\gamma}^{(\gamma)}(\theta_{t_l}) - g_{j,\gamma}^{(\gamma)}(\psi_{t_l})\right| \\
&\quad + \sum_{l\in I_4} \left|g_{j,\gamma}^{(\gamma)}(\theta_{t_l}) - g_{j,\gamma}^{(\gamma)}(\psi_{t_l})\right| \\
&= \sum_{l\in I_2} \left|0 - \mathbb{I}\left\{< \psi_{t_l} >_0 f_j^{(\gamma)}(< \psi_{t_l} >_1^k) \geq \dot{b}(\gamma)\right\}\right| + \sum_{l\in I_3} \left|\mathbb{I}\left\{< \theta_{t_l} >_0 f_j^{(\gamma)}(< \theta_{t_l} >_1^k) \geq \dot{b}(\gamma)\right\} - 0\right| \\
&\quad + \sum_{l\in I_4} \left|\mathbb{I}\left\{< \theta_{t_l} >_0 f_j^{(\gamma)}(< \theta_{t_l} >_1^k) \geq \dot{b}(\gamma)\right\} - \mathbb{I}\left\{< \psi_{t_l} >_0 f_j^{(\gamma)}(< \psi_{t_l} >_1^k) \geq \dot{b}(\gamma)\right\}\right| \qquad \text{(A.112)} \\
&\leq \sum_{l\in I_2} \mathbb{I}\{\theta_{t_l} \neq \psi_{t_l}\} + \sum_{l\in I_3} \mathbb{I}\{\theta_{t_l} \neq \psi_{t_l}\} + \sum_{l\in I_4} \mathbb{I}\{\theta_{t_l} \neq \psi_{t_l}\} \\
&\leq \sum_{t=1}^m \mathbb{I}\{\theta_t \neq \psi_t\} \\
&\leq r(k,k^*) \sum_{t=1}^m \mathbb{I}\{s_t^* \neq q_t^*\} \qquad \text{(A.113)} \\
&= r(k,k^*) d_H(s^{*(m)}, q^{*(m)}) \\
&\leq r(k,k^*)\Delta \qquad \text{(A.114)}
\end{aligned}$$

From (A.114) it follows that $G_{j,\gamma}^{(\gamma)}$ is Lipschitz with constant $r(k,k^*)$. So $G_{j,\gamma}^{(\gamma)}/r$ is Lipschitz with constant 1 and, from Section A.1, we may use (A.14) for $S^{*(m)}$ and $G_{j,\gamma}^{(\gamma)}/r$. We have

$$\begin{aligned}
\mathbb{P}\left(M'_{j,\gamma}(\eta)\right) &\leq \mathbb{P}\left(G_{j,\gamma}^{(\gamma)}(S^{*(m)}) > \mathbb{E}\left[G_{j,\gamma}^{(\gamma)}\left(S^{*(m)}\right)\right] + \frac{m\kappa(\gamma,\eta)}{2}\right) \\
&= \mathbb{P}\left(\frac{G_{j,\gamma}^{(\gamma)}(S^{*(m)})}{r} > \frac{\mathbb{E}\left[G_{j,\gamma}^{(\gamma)}\left(S^{*(m)}\right)\right]}{r} + \frac{m\kappa(\gamma,\eta)}{2r}\right) \\
&\leq \exp\left\{-\frac{m}{2}\left(\frac{\kappa}{2\rho r}\right)^2\right\}. \qquad \text{(A.115)}
\end{aligned}$$

Plugging the choice of κ (A.102) in the right side of (A.115) gives

$$\mathbb{P}\left(M'_{j,\gamma}(\omega,\omega,\eta)\right) \leq \frac{\eta}{\mathcal{N}_\gamma}. \tag{A.116}$$

From (A.106) and (A.115), it follows that

$$\mathbb{P}\left(J(\gamma,\gamma,\eta)\right) \leq \sum_{j=1}^{\mathcal{N}_\gamma} \frac{\eta}{\mathcal{N}_\gamma} = \eta. \tag{A.117}$$

Now, substitute for ϵ_2 in (A.101) the following,

$$\epsilon_2 = \kappa(\gamma, \gamma\eta). \tag{A.118}$$

Then it follows that (A.101) equals

$$\mathbb{P}\left(\bigcup_{\gamma\in(0,\mathrm{diam}(\mathbb{S}_k)]} J(\gamma, 2\gamma, \gamma\eta)\right). \tag{A.119}$$

Define the set $\Gamma_l \subset (0, \mathrm{diam}(\mathbb{S}_k)]$ as follows,

$$\Gamma_l = \left[\left(\frac{1}{2}\right)^{l+1} \mathrm{diam}(\mathbb{S}_k), \left(\frac{1}{2}\right)^{l} \mathrm{diam}(\mathbb{S}_k)\right] \tag{A.120}$$

and the set $\bigcup_{l=0}^{\infty}\Gamma_l$ contains the possible range $(0, \mathrm{diam}(\mathbb{S}_k)]$ for γ. Therefore (A.119) is bounded from above as follows,

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{\gamma\in(0,\mathrm{diam}(\mathbb{S}_k)]} J(\gamma,2\gamma,\gamma\eta)\right) &= \mathbb{P}\left(\bigcup_{l=0}^{\infty}\bigcup_{\gamma\in\Gamma_l} J(\gamma,2\gamma,\gamma\eta)\right) \\
&\leq \sum_{l=0}^{\infty}\mathbb{P}\left(\bigcup_{\gamma\in\Gamma_l} J(\gamma,2\gamma,\gamma\eta)\right) \\
&\leq \sum_{l=0}^{\infty}\mathbb{P}\left(J\left(\left(\frac{1}{2}\right)^{l}\mathrm{diam}(\mathbb{S}_k), \left(\frac{1}{2}\right)^{l}\mathrm{diam}(\mathbb{S}_k), \eta\left(\frac{1}{2}\right)^{l}\mathrm{diam}(\mathbb{S}_k)\right)\right)
\end{aligned} \tag{A.121}$$

where (A.121) follows from (A.105) and the fact that for all $\gamma \in \Gamma_l$ we have $\gamma \leq (\frac{1}{2})^l \mathrm{diam}(\mathbb{S}_k) \leq 2\gamma$ and $\gamma \leq (\frac{1}{2})^l \mathrm{diam}(\mathbb{S}_k)$.

From (A.117), the l^{th} term in (A.121) is bounded by $\eta(1/2)^l \mathrm{diam}(\mathbb{S}_k)$ and so (A.101) is bounded from above by

$$\eta\mathrm{diam}(\mathbb{S}_k)\sum_{l=0}^{\infty}\left(\frac{1}{2}\right)^{l} \leq 2\eta\mathrm{diam}(\mathbb{S}_k). \tag{A.122}$$

A.2.6. *Bounding (A.98).* The following probability

$$\mathbb{P}\left(\exists h, \exists\gamma,\ P\left(\tilde{A}_{h,\gamma}\right) - \hat{P}_m(A_{h,0}) > \epsilon_2 \hat{P}_m(A_{h,0})/2P\left(\tilde{B}_{h,\gamma}\middle|\tilde{A}_{h,\gamma}\right)\right) \tag{A.123}$$

is bounded from above by

$$\mathbb{P}\left(\exists h, \exists\gamma,\ P\left(\tilde{A}_{h,\gamma}\right) - \hat{P}_m(A_{h,0}) > \epsilon_2 \hat{P}_m(A_{h,0})/2\right). \tag{A.124}$$

Recall from (A.94) that $\hat{P}_m(A_{h,0}) = 1$, hence (A.124) is bounded from above by

$$\mathbb{P}\left(\exists h, \exists\gamma,\ P\left(\tilde{A}_{h,\gamma}\right) > 1 + \epsilon_2/2\right).$$

This latter probability equals zero because the inequality $P\left(\tilde{A}_{h,\gamma}\right) > 1 + \epsilon_2/2$ does not hold for any h and γ. Therefore, (A.123) and hence (A.98) vanish.

A.2.7. *Finalizing.* The conclusion of Section A.2.6, together with the fact that (A.101) is bounded from above by (A.122), and with the choice for ϵ_2 in (A.118) and κ as defined in (A.102), it follows that (A.91) is bounded from above by $2\eta\text{diam}\left(\mathbb{S}_k\right)$. And together with the bound of (A.90) on the sum of (A.30) and (A.31), it follows that (A.17) is bounded from above by

$$4\eta(1-\omega_0)\text{diam}\left(\mathbb{S}_k\right)+2\eta\text{diam}\left(\mathbb{S}_k\right)=2\eta\left(2\left(1-\omega_0\right)+1\right)\text{diam}\left(\mathbb{S}_k\right). \tag{A.125}$$

From (3.58) it follows that

$$\mathcal{N}_{\gamma/2}\leq\left(2\left\lceil\frac{6\text{diam}\left(\mathbb{S}_k\right)}{\gamma}\right\rceil+1\right)^{N_{\gamma/6}}. \tag{A.126}$$

Let us substitute the right side of (A.126) for $\mathcal{N}_{\gamma/2}$ in the expression (A.41) for $\kappa\left(\gamma/2,\omega/2,(\omega-\omega_0)\gamma\eta/2\right)$ used in (A.65). We obtain

$$\begin{aligned}\epsilon_1&=\frac{4r(k,k^*)\rho(k^*,\mathfrak{l}_0)}{\omega}\sqrt{\frac{2}{n}\ln\left(\frac{2\mathcal{N}_{\gamma/2}}{(\omega-\omega_0)\gamma\eta}\right)}\\&\leq\frac{4r(k,k^*)\rho(k^*,\mathfrak{l}_0)}{\omega}\sqrt{\frac{2}{n}\left(N_{\gamma/6}\ln\left(2\left\lceil\frac{6\text{diam}\left(\mathbb{S}_k\right)}{\gamma}\right\rceil+1\right)+\ln\left(\frac{2}{(\omega-\omega_0)\gamma\eta}\right)\right)}.\end{aligned} \tag{A.127}$$

Substitute γ for α in (3.58) and place the right side in the expression (A.102) for $\kappa\left(\gamma,\gamma\eta\right)$ used in (A.118). We obtain

$$\begin{aligned}\epsilon_2&=2r(k,k^*)\rho(k^*,\mathfrak{l}_0)\sqrt{\frac{2}{m}\ln\left(\frac{\mathcal{N}_\gamma}{\gamma\eta}\right)}\\&\leq 2r(k,k^*)\rho(k^*,\mathfrak{l}_0)\sqrt{\frac{2}{m}\left(N_{\gamma/3}\ln\left(2\left\lceil\frac{3\text{diam}\left(\mathbb{S}_k\right)}{\gamma}\right\rceil+1\right)+\ln\left(\frac{1}{\gamma\eta}\right)\right)}.\end{aligned} \tag{A.128}$$

Substitute the sum (A.24) for ϵ in (A.17). With (A.127), (A.128), it follows that if we let

$$\begin{aligned}\epsilon=&\frac{4r(k,k^*)\rho(k^*,\mathfrak{l}_0)}{\omega}\sqrt{\frac{2}{n}\left(N_{\gamma/6}\ln\left(2\left\lceil\frac{6\text{diam}\left(\mathbb{S}_k\right)}{\gamma}\right\rceil+1\right)+\ln\left(\frac{2}{(\omega-\omega_0)\gamma\eta}\right)\right)}\\&+2r(k,k^*)\rho(k^*,\mathfrak{l}_0)\sqrt{\frac{2}{m}\left(N_{\gamma/3}\ln\left(2\left\lceil\frac{3\text{diam}\left(\mathbb{S}_k\right)}{\gamma}\right\rceil+1\right)+\ln\left(\frac{1}{\gamma\eta}\right)\right)}\end{aligned} \tag{A.129}$$

then (A.17) is bounded from above by (A.125) which is expressed as $2\eta\left(3-2\omega_0\right)\text{diam}\left(\mathbb{S}_k\right)$. Setting this bound to δ, solving for η, substituting for η in (A.129) and using (3.16) yields the statement of the theorem.

A.3. **Proof of Theorem 16.** We need to show that

$$\mathbb{P}\left(\exists h\in\mathcal{H},\,\exists 0<\gamma\leq\text{diam}(\mathbb{S}_k),\exists\omega\in(\ell/m,1],\,L(h|\gamma)>\mathcal{L}_{\nu^{(a)}}^{(m,\gamma/6)}(h)+\xi(m,\gamma,\omega,\delta),\nu^{(a)}\geq\omega m\right) \tag{A.130}$$

is no larger than δ where ξ is defined in (6.3).

Define

$$\dot{a}(\gamma):=8\gamma,\,\hat{a}(\gamma):=10\gamma,$$

and

$$\hat{b}(\gamma):=43\gamma,$$

all of which are positive for $\gamma>0$ and satisfy

$$\begin{aligned}
-b(\gamma) &< -b(\gamma)+\gamma \\
&\leq -\hat{b}(\gamma) < -\hat{b}(\gamma)+\gamma \\
&\leq -a(2\gamma) \\
&< -a(\gamma) < -a(\gamma)+\gamma \\
&\leq -\hat{a}(2\gamma) \\
&\leq -\hat{a}(\gamma) < -\hat{a}(\gamma)+2\gamma \\
&\leq -\dot{a}(\gamma) < -\dot{a}(\gamma)+3\gamma \\
&< -\hat{a}(\gamma/2) < -\hat{a}(\gamma/2)+\gamma \\
&\leq -a(\gamma/6) \\
&< 0. \qquad (A.131)
\end{aligned}$$

Let us fix γ, h and ω in the expression whose probability is (A.130). We consider the event $L(h|\gamma) - \mathcal{L}_{\nu}^{(m,\gamma/6)}(h) > \epsilon$, $\nu > \omega m$, and bound its probability

$$\mathbb{P}\left(L(h|\gamma) - \mathcal{L}_{\nu^{(a)}}^{(m,\gamma/6)}(h) > \epsilon,\, \nu \geq \omega m\right). \qquad (A.132)$$

For $\theta \in \mathbb{S}_{k+1}$ we write $\theta = [s,x]$ and define the sets $A_{h,\gamma}$, $B_{h,\gamma} \subseteq \mathbb{S}_{k+1}$ by

$$A_{h,\gamma} := \{\theta :\, |f_h(s)| \geq a(\gamma)\},\; B_{h,\gamma} := \{\theta :\, x f_h(s) \geq b(\gamma)\}. \qquad (A.133)$$

Define the counterpart of (6.1) as follows,

$$\overline{\mathcal{L}}_{\nu}^{(m,\gamma/6)}(h) := \frac{1}{\nu} \sum_{l:\left|f_h(S_{t_l-1})\right| \geq a(\gamma/6)} \mathbb{I}\{X_{t_l} f_h(S_{t_l-1}) \geq b(\gamma)\}. \qquad (A.134)$$

From (3.31), (3.32) we have,

$$L(h|\gamma) := \mathbb{P}\left(X_t f_h(S_{t-1}) < 0 \middle| |f_h(S_{t-1})| \geq a(\gamma)\right).$$

Denote by

$$\overline{L}(h|\gamma) := \mathbb{P}\left(X_t f_h(S_{t-1}) \geq 0 \middle| |f_h(S_{t-1})| \geq a(\gamma)\right).$$

Therefore, since $b(0) = 0$ then with (A.133) we have

$$\overline{L}(h|\gamma) = P(B_{h,0}|A_{h,\gamma}).$$

We have,

$$\begin{aligned}
L(h|\gamma) - \mathcal{L}_{\nu}^{(m,\gamma/6)}(h) &= 1 - \overline{L}(h|\gamma) - 1 + \overline{\mathcal{L}}_{\nu}^{(m,\gamma/6)}(h) \\
&= \overline{\mathcal{L}}_{\nu}^{(m,\gamma/6)}(h) - \overline{L}(h|\gamma) \\
&= \overline{\mathcal{L}}_{\nu}^{(m,\gamma/6)}(h) - P(B_{h,0}|A_{h,\gamma}).
\end{aligned}$$

Denote by

$$\begin{aligned}
\hat{P}_m(A_{h,\gamma/6}) &:= \frac{1}{m}\sum_{t=1}^{m} \mathbb{I}\{\Theta_t \in A_{h,\gamma/6}\} \\
&= \frac{\nu}{m}. \qquad (A.135)
\end{aligned}$$

We have,

$$\begin{aligned}
\overline{\mathcal{L}}_{\nu}^{(m,\gamma/6)}(h) - P(B_{h,0}|A_{h,\gamma}) &= \frac{1}{\hat{P}_m(A_{h,\gamma/6})}\left(\overline{\mathcal{L}}_{\nu}^{(m,\gamma/6)}(h)\hat{P}_m(A_{h,\gamma/6}) - P(B_{h,0},A_{h,\gamma})\right) \\
&+ P(B_{h,0},A_{h,\gamma})\left(\frac{1}{\hat{P}_m(A_{h,\gamma/6})} - \frac{1}{P(A_{h,\gamma})}\right). \qquad (A.136)
\end{aligned}$$

We have,

$$\begin{aligned}
&P\left(B_{h,0}, A_{h,\gamma}\right)\left(\frac{1}{\hat{P}_m\left(A_{h,\gamma/6}\right)}-\frac{1}{P\left(A_{h,\gamma}\right)}\right)\\
&=P\left(B_{h,0}\,|A_{h,\gamma}\right)P\left(A_{h,\gamma}\right)\left(\frac{P\left(A_{h,\gamma}\right)-\hat{P}_m\left(A_{h,\gamma/6}\right)}{\hat{P}_m\left(A_{h,\gamma/6}\right)P\left(A_{h,\gamma}\right)}\right)\\
&=\frac{P\left(B_{h,0}\,|A_{h,\gamma}\right)}{\hat{P}_m\left(A_{h,\gamma/6}\right)}\left(P\left(A_{h,\gamma}\right)-\hat{P}_m\left(A_{h,\gamma/6}\right)\right).
\end{aligned}$$

Thus (A.132) is bounded from above by

$$\begin{aligned}
&\mathbb{P}\left(\overline{\mathcal{L}}_{\nu}^{(m,\gamma/6)}(h)\hat{P}_m\left(A_{h,\gamma/6}\right)-P\left(B_{h,0},A_{h,\gamma}\right)>\epsilon\hat{P}_m(A_{h,\gamma/6})/2,\,\nu>\omega m\right)\\
&\quad+\quad\mathbb{P}\left(P(A_{h,\gamma})-\hat{P}_m(A_{h,\gamma/6})>\epsilon\hat{P}_m(A_{h,\gamma/6})/2P\left(B_{h,0}\,|A_{h,\gamma}\right),\,\nu>\omega m\right). \qquad \text{(A.137)}
\end{aligned}$$

For convenience, we denote by

$$\hat{P}_m\left(B_{h,\gamma},A_{h,\gamma/6}\right):=\overline{\mathcal{L}}_{\nu}^{(m,\gamma/6)}(h)\hat{P}_m\left(A_{h,\gamma/6}\right)$$

since $\overline{\mathcal{L}}_{\nu}^{(m,\gamma/6)}(h)$ is the empirical probability of the set $B_{h,\gamma}$ conditioned on the state being in $A_{h,\gamma/6}$.

Therefore (A.130) is bounded from above by

$$\begin{aligned}
&\mathbb{P}\left(\exists h,\exists\gamma,\exists\omega:\hat{P}_m\left(B_{h,\gamma},A_{h,\gamma/6}\right)-\,P\left(B_{h,0},A_{h,\gamma}\right)>\epsilon\hat{P}_m(A_{h,\gamma/6})/2,\,\nu>\omega m\right) \qquad \text{(A.138)}\\
&\quad+\quad\mathbb{P}\left(\exists h,\exists\gamma,\exists\omega:P\left(A_{h,\gamma}\right)-\hat{P}_m\left(A_{h,\gamma/6}\right)>\epsilon\hat{P}_m(A_{h,\gamma/6})/2P\left(B_{h,0}\,|A_{h,\gamma}\right),\,\nu>\omega m\right). \qquad \text{(A.139)}
\end{aligned}$$

Denote by

$$A_{j,\gamma'}^{(\gamma)}:=\left\{\theta=[s,x]:\left|f_j^{(\gamma)}(s)\right|\geq\hat{a}(\gamma')\right\}, \qquad \text{(A.140)}$$

and

$$B_{j,\gamma'}^{(\gamma)}:=\left\{\theta:\,xf_j^{(\gamma)}(s)\geq\hat{b}(\gamma')\right\}.$$

We have the following:

Claim 24. For h and h_j such that f_h and f_j satisfy (3.56) with $\alpha=\gamma$, we have

$$\hat{P}_m\left(B_{h,\gamma},A_{h,\gamma/6}\right)\leq\hat{P}_m\left(B_{j,\gamma}^{(\gamma)},A_{j,\gamma}^{(\gamma)}\right). \qquad \text{(A.141)}$$

Proof. Consider $\theta=[s,x]\in B_{h,\gamma}\bigcap A_{h,\gamma/6}$, then $xf_h(s)\geq b(\gamma)$ and $|f_h(s)|\geq a(\gamma/6)$. If $x=-1$ then

$$f_h(s)\leq-b(\gamma)$$

and we have,

$$\begin{aligned}
f_j^{(\gamma)}(s)&\leq f_h(s)+\gamma\\
&\leq-b(\gamma)+\gamma\\
&\leq-\hat{b}(\gamma)\\
&\leq-\hat{a}(\gamma)
\end{aligned}$$

which follow from (A.131). If $x=1$ then $f_h(s)\geq b(\gamma)$ and we have,

$$\begin{aligned}
f_j^{(\gamma)}(s)&\geq f_h(s)-\gamma\\
&\geq b(\gamma)-\gamma\\
&\geq\hat{b}(\gamma)\\
&\geq\hat{a}(\gamma)
\end{aligned}$$

therefore $xf_j^{(\gamma)}(s)\geq\hat{b}(\gamma)$ and $\left|f_j^{(\gamma)}\right|\geq\hat{a}(\gamma)$. Therefore $xf_j^{(\gamma)}(s)\geq\hat{b}(\gamma)$ and $\left|f_j^{(\gamma)}\right|\geq\hat{a}(\gamma)$. Hence $\theta\in B_{j,\gamma}^{(\gamma)}\cap A_{j,\gamma}^{(\gamma)}$. Therefore, $B_{h,\gamma}\bigcap A_{h,\gamma/6}\subseteq B_{j,\gamma}^{(\gamma)}\bigcap A_{j,\gamma}^{(\gamma)}$ from which (A.141) follows. □

Next, we have the following:

Claim 25. For h and h_j such that f_h and $f_j^{(\gamma)}$ satisfy (3.56) with $\alpha = \gamma$, we have

$$P\left(B_{j,\gamma}^{(\gamma)}, A_{j,\gamma}^{(\gamma)}\right) \leq P\left(B_{h,0}, A_{h,2\gamma}\right). \tag{A.142}$$

Proof. Consider any $\theta = [s, x] \in A_{j,\gamma}^{(\gamma)} \bigcap B_{j,\gamma}^{(\gamma)}$. If $x = -1$ then $f_j^{(\gamma)}(s) \leq -\hat{b}(\gamma)$. From (3.56) we have

$$\begin{aligned} f_h(s) &\leq f_j^{(\gamma)}(s) + \gamma \\ &\leq -\hat{b}(\gamma) + \gamma \\ &\overset{(i)}{<} -a(2\gamma) \\ &\overset{(ii)}{<} 0 \end{aligned}$$

where (i) and (ii) follow from (A.131). Hence for this θ, $f_h(s) < -a(2\gamma)$ and $x f_h(s) > 0$ (since $x = -1$) and therefore $\theta \in A_{h,2\gamma} \bigcap B_{h,0}$. If $x = 1$ then $f_j^{(\gamma)}(s) \geq \hat{b}(\gamma)$ and we have

$$\begin{aligned} f_h(s) &\geq f_j^{(\gamma)}(s) - \gamma \\ &\geq \hat{b}(\gamma) - \gamma \\ &> a(2\gamma) \\ &> 0 \end{aligned}$$

therefore $\theta \in A_{h,2\gamma} \cap B_{h,0}$. Hence

$$A_{j,\gamma}^{(\gamma)} \bigcap B_{j,\gamma}^{(\gamma)} \subseteq A_{h,2\gamma} \bigcap B_{h,0}. \tag{A.143}$$

Hence (A.142) follows. □

We have the following claim.

Claim 26. For h and h_j such that f_h and $f_j^{(\gamma)}$ satisfy (3.56) with $\alpha = \gamma$,

$$P(A_{j,2\gamma}^{(\gamma)}) \geq P(A_{h,\gamma}). \tag{A.144}$$

Proof. Consider $\theta \in A_{h,\gamma}$. If $f_h(s) \leq -a(\gamma)$ then we have

$$\begin{aligned} f_j^{(\gamma)}(s) &\leq f_h(s) + \gamma \\ &\leq -a(\gamma) + \gamma \\ &\leq -\hat{a}(2\gamma) \end{aligned}$$

where the last inequality follows from (A.131). It follows that $\theta \in A_{j,2\gamma}^{(\gamma)}$. If $f_h(s) \geq a(\gamma)$ then we have

$$\begin{aligned} f_j^{(\gamma)}(s) &\geq f_h(s) - \gamma \\ &\geq a(\gamma) - \gamma \\ &\geq \hat{a}(2\gamma) \end{aligned}$$

so $\theta \in A_{j,2\gamma}^{(\gamma)}$. This proves (A.144). □

Claim 27. For h and h_j such that f_h and $f_j^{(\gamma)}$ satisfy (3.56) with $\alpha = \gamma$,

$$\hat{P}_m\left(A_{h,\gamma/6}\right) \geq \hat{P}_m\left(A_{j,\gamma/2}^{(\gamma)}\right). \tag{A.145}$$

Proof. Consider any $\theta = [s, x] \in A_{j,\gamma/2}^{(\gamma)}$. If $f_j^{(\gamma)}(s) \leq -\hat{a}(\gamma/2)$ then

$$\begin{aligned} f_h(s) &\leq f_j^{(\gamma)}(s) + \gamma \\ &< -\hat{a}(\gamma/2) + \gamma \\ &\leq -a(\gamma/6) \end{aligned}$$

which follows from (A.131) and hence $\theta \in A_{h,\gamma/6}$. If $f_j^{(\gamma)}(s) \geq \hat{a}(\gamma/2)$ then

$$\begin{aligned} f_h(s) &\geq f_j^{(\gamma)}(s) - \gamma \\ &\geq \hat{a}(\gamma/2) - \gamma \\ &\geq a(\gamma/6) \end{aligned}$$

so $\theta \in A_{h,\gamma/6}$. Therefore $A^{(\gamma)}_{j,\gamma/2} \subseteq A_{h,\gamma/6}$ and (A.145) is proved. □

A.3.1. *Bounding (A.138).* Denote by

$$\lambda := \frac{\nu}{m} = \hat{P}_m(A_{h,\gamma/6}).$$

The probability in (A.138) is expressed as follows,

$$\mathbb{P}\left(\exists h, \exists \gamma, \exists \omega : \hat{P}_m\left(B_{h,\gamma}, A_{h,\gamma/6}\right) - P\left(B_{h,0}, A_{h,\gamma}\right) > \epsilon\lambda/2,\, \lambda > \omega\right)$$

which is bounded from above by

$$\mathbb{P}\left(\exists h, \exists \gamma, \exists \omega : \hat{P}_m\left(B_{h,\gamma}, A_{h,\gamma/6}\right) - P\left(B_{h,0}, A_{h,\gamma}\right) > \epsilon\omega/2\right). \tag{A.146}$$

Let

$$\kappa\left(\gamma, \omega, \eta\right) := \frac{2r(k,k^*)\rho(k^*, \mathfrak{l}_0)}{\omega}\sqrt{\frac{2}{m}\ln\left(\frac{\mathcal{N}_\gamma}{\eta}\right)} \tag{A.147}$$

where ρ and r are defined in (3.53) and (3.13), and denote by

$$\omega_0 := \frac{\ell}{m}.$$

Henceforth, for conciseness, we write $\exists\omega$ for $\exists\omega \in (\omega_0, 1]$. Define

$$J(\gamma_1, \gamma_2, \eta) := \left\{x^{(m)} : \exists h\, \exists\omega, \hat{P}_m\left(B_{h,\gamma_2}, A_{h,\gamma_2/6}\right) > P\left(B_{h,0}, A_{h,2\gamma_1}\right) + \frac{\omega\kappa\left(\gamma_1, \frac{\omega}{2}, \frac{(\omega-\omega_0)\eta}{2}\right)}{2}\right\}. \tag{A.148}$$

For $\gamma_1 \le \gamma \le \gamma_2$, we claim $\hat{P}_m\left(B_{h,\gamma_2}, A_{h,\gamma_2/6}\right) \le \hat{P}_m\left(B_{h,\gamma}, A_{h,\gamma/6}\right)$: to see this, note that $A_{h,\gamma_2/6} \subseteq A_{h,\gamma/6}$ because $a(\gamma_2/6) \ge a(\gamma/6)$ as a is non-decreasing with γ. And, $B_{h,\gamma_2} \subseteq B_{h,\gamma}$ as $b(\gamma_2) \ge b(\gamma)$ since b is non-decreasing with γ. Hence, $B_{h,\gamma_2} \bigcap A_{h,\gamma_2/6} \subseteq B_{h,\gamma} \bigcap A_{h,\gamma/6}$ from which the claim follows. Next we claim $P\left(B_{h,0}, A_{h,2\gamma_1}\right) \ge P\left(B_{h,0}, A_{h,2\gamma}\right)$ as we next show: we have $a(2\gamma) \ge a(2\gamma_1)$ therefore $|f_h(s)| \ge a(2\gamma)$ implies that $|f_h(s)| \ge a(2\gamma_1)$ hence $A_{h,2\gamma} \subseteq A_{h,2\gamma_1}$. So we have $B_{h,0} \bigcap A_{h,2\gamma} \subseteq B_{h,0} \bigcap A_{h,2\gamma_1}$ and the claim follows. Also,

$$\kappa\left(\gamma, \frac{\omega}{2}, \frac{(\omega-\omega_0)\eta}{2}\right) \le \kappa\left(\gamma_1, \frac{\omega}{2}, \frac{(\omega-\omega_0)\eta}{2}\right) \tag{A.149}$$

because $\mathcal{N}_\alpha$ is non-decreasing with decreasing α. It follows that

$$J\left(\gamma_1, \gamma_2, \eta\right) \subseteq J(\gamma, \gamma, \eta). \tag{A.150}$$

Let

$$M_{h,\gamma}\left(\omega_1, \omega_2, \eta\right) := \left\{x^{(m)} : \hat{P}_m\left(B_{h,\gamma}, A_{h,\gamma/6}\right) > P\left(B_{h,0}, A_{h,2\gamma}\right) + \frac{\omega_2\kappa\left(\gamma, \omega_1, \eta\right)}{2}\right\} \tag{A.151}$$

then for $\omega_1 \le \omega \le \omega_2$ we have

$$M_{h,\gamma}\left(\omega_1, \omega_2, \eta\right) \subseteq M_{h,\gamma}\left(\omega, \omega, \eta\right) \tag{A.152}$$

which follows from the fact that $\kappa\left(\gamma, \omega_1, \eta\right) \ge \kappa\left(\gamma, \omega, \eta\right)$ and $\omega_2 \ge \omega$, and for $\eta_a \le \eta_b$ we have

$$M_{h,\gamma}(\omega, \omega, \eta_a) \subseteq M_{h,\gamma}(\omega, \omega, \eta_b). \tag{A.153}$$

We have,

$$\mathbb{P}\left(J\left(\gamma, \gamma, \eta\right)\right) = \mathbb{P}\left(\left\{x^{(m)} : \exists h,\, x^{(m)} \in \bigcup_{\omega_0 \le \omega \le 1} M_{h,\gamma}\left(\frac{\omega}{2}, \omega, \frac{(\omega-\omega_0)\eta}{2}\right)\right\}\right). \tag{A.154}$$

Define the set $\Delta_l \subset [0, 1]$ as follows,

$$\Delta_l = \left[\omega_0 + (1-\omega_0)\left(\frac{1}{2}\right)^{l+1}, \omega_0 + (1-\omega_0)\left(\frac{1}{2}\right)^{l}\right]. \tag{A.155}$$

We have $[\omega_0, 1] \subseteq \bigcup_{l=0}^{\infty} \Delta_l$. The right side of (A.154) is no larger than

$$\mathbb{P}\left(\exists h,\, X^{(m)} \in \bigcup_{l=0}^{\infty} \bigcup_{\omega \in \Delta_l} M_{h,\gamma}\left(\frac{\omega}{2}, \omega, \frac{(\omega-\omega_0)\eta}{2}\right)\right). \tag{A.156}$$

Now for any $h \in \mathcal{H}$ there exists a $j \in \{1, \ldots, \mathcal{N}_\gamma\}$ such that (3.56) is satisfied with $\alpha = \gamma$. Define

$$M_{j,\gamma}(\omega_1,\omega_2,\eta) := \left\{ x^{(m)} : \hat{P}_m\left(B_{j,\gamma}^{(\gamma)}A_{j,\gamma}^{(\gamma)}\right) > P\left(B_{j,\gamma}^{(\gamma)}A_{j,\gamma}^{(\gamma)}\right) + \frac{\omega_2\kappa(\gamma,\omega_1,\eta)}{2} \right\}. \tag{A.157}$$

For the same reason that (A.152) and (A.153) hold, we have for $\omega_1 \le \omega \le \omega_2$, $\eta_a \le \eta_b$,

$$M_{j,\gamma}(\omega_1,\omega_2,\eta) \subseteq M_{j,\gamma}(\omega,\omega,\eta) \tag{A.158}$$

and

$$M_{j,\gamma}(\omega,\omega,\eta_a) \subseteq M_{j,\gamma}(\omega,\omega,\eta_b).$$

Then from Claim 24 and Claim 25 it follows that if there exists an h such that $X^{(m)} \in M_{h,\gamma}(\omega_1,\omega_2,\eta)$ then there exists a $j \in \{1,\ldots,\mathcal{N}_\gamma\}$ such that $X^{(m)} \in M_{j,\gamma}(\omega_1,\omega_2,\eta)$. Hence if there exists $\omega \in (\omega_0,1]$ such t-hat there is an h with $X^{(m)} \in M_{h,\gamma}\left(\frac{\omega}{2},\omega,\eta\right)$ then there exists a $j \in \{1,\ldots,\mathcal{N}_\gamma\}$ such that $X^{(m)} \in M_{j,\gamma}\left(\frac{\omega}{2},\omega,\eta\right)$.

Therefore from (A.156) we have,

$$\begin{aligned}\mathbb{P}(J(\gamma,\gamma,\eta)) &\le \mathbb{P}\left(\exists 1 \le j \le \mathcal{N}_\gamma,\ X^{(m)} \in \bigcup_{l=0}^{\infty}\bigcup_{\omega\in\Delta_l} M_{j,\gamma}\left(\frac{\omega}{2},\omega,\frac{(\omega-\omega_0)\eta}{2}\right)\right)\\ &\le \sum_{j=1}^{\mathcal{N}_\gamma}\sum_{l=0}^{\infty}\mathbb{P}\left(M_{j,\gamma}\left(\omega_0+(1-\omega_0)\left(\frac{1}{2}\right)^{l+1},\omega_0+(1-\omega_0)\left(\frac{1}{2}\right)^{l+1},\frac{\eta}{2}(1-\omega_0)\left(\frac{1}{2}\right)^{l}\right)\right)\end{aligned} \tag{A.159}$$

where the inequality follows from (A.158) as for every $\omega \in \Delta_l$ we have $\omega/2 \le \omega_0 + (1-\omega_0)\left(\frac{1}{2}\right)^{l+1} \le \omega$.

Let

$$g_{j,\gamma}^{(\gamma)}(\theta) := \mathbb{I}\left\{\left|f_j^{(\gamma)}(s)\right| \ge \hat{a}(\gamma),\ x f_j^{(\gamma)}(s) \ge \hat{b}(\gamma)\right\} \tag{A.160}$$

and

$$G_{j,\gamma}^{(\gamma)}(\Theta^{(m)}) := \sum_{t=1}^{m} g_{j,\gamma}^{(\gamma)}(\Theta_t).$$

Then

$$\begin{aligned}G_{j,\gamma}^{(\gamma)}(\Theta^{(m)}) &= \sum_{t=1}^{m}\mathbb{I}\left\{\left|f_j^{(\gamma)}\left(\langle\Theta_t\rangle_1^k\right)\right| \ge \hat{a}(\gamma),\ \langle\Theta_t\rangle_0 f_j^{(\gamma)}\left(\langle\Theta_t\rangle_1^k\right) \ge \hat{b}(\gamma)\right\}\\ &= m\hat{P}_m\left(A_{j,\gamma}^{(\gamma)},B_{j,\gamma}^{(\gamma)}\right)\end{aligned} \tag{A.161}$$

and

$$\begin{aligned}\mathbb{E}\left[G_{j,\gamma}^{(\gamma)}\left(\Theta^{(m)}\right)\right] &= \mathbb{E}\left[m\hat{P}_m\left(A_{j,\gamma}^{(\gamma)},B_{j,\gamma}^{(\gamma)}\right)\right]\\ &= \mathbb{E}\left[\sum_{l=1}^{m}\mathbb{I}\left\{\Theta_{t_l} \in A_{j,\gamma}^{(\gamma)}\bigcap B_{j,\gamma}^{(\gamma)}\right\}\right]\\ &= mP\left(A_{j,\gamma}^{(\gamma)}B_{j,\gamma}^{(\gamma)}\right)\end{aligned} \tag{A.162}$$

and therefore,

$$\mathbb{P}(M_{j,\gamma}(\omega,\omega,\eta)) = \mathbb{P}\left(G_{j,\gamma}^{(\gamma)}(S^{*(m)}) > \mathbb{E}\left[G_{j,\gamma}^{(\gamma)}\left(S^{*(m)}\right)\right] + \frac{m\omega\kappa(\gamma,\omega,\eta)}{2}\right). \tag{A.163}$$

As is shown in Section A.2.1, here also, the functions $G_{j,\gamma}^{(\gamma)}$ are Lipschitz with constant $r(k,k^*)$. So $G_{j,\gamma}^{(\gamma)}/r$ is Lipschitz with constant 1 and, from Section A.1, we may use (A.14) for $S^{*(m)}$ and $G_{j,\gamma}^{(\gamma)}/r$. We have

$$\begin{aligned}\mathbb{P}(M_{j,\gamma}(\omega,\omega,\eta)) &\le \mathbb{P}\left(G_{j,\gamma}^{(\gamma)}\left(S^{*(m)}\right) > \mathbb{E}\left[G_{j,\gamma}^{(\gamma)}(S^{*(m)})\right] + m\frac{\omega\kappa(\gamma,\omega,\eta)}{2}\right)\\ &= \mathbb{P}\left(\frac{G_{j,\gamma}^{(\gamma)}(S^{*(m)})}{r} > \frac{\mathbb{E}\left[G_{j,\gamma}^{(\gamma)}(S^{*(m)})\right]}{r} + m\frac{\omega\kappa(\gamma,\omega,\eta)}{2r}\right)\\ &\le \exp\left\{-\frac{m}{2}\left(\frac{\omega\kappa}{2\rho r}\right)^2\right\}.\end{aligned} \tag{A.164}$$

Plugging the choice of κ (A.147) in the right side of (A.164) gives

$$\mathbb{P}\left(M_{j,\gamma}\left(\omega,\omega,\eta\right)\right) \leq \frac{\eta}{\mathcal{N}_\gamma}. \tag{A.165}$$

From (A.159) and (A.165), it follows that

$$\mathbb{P}\left(J\left(\gamma,\gamma,\eta\right)\right) \leq \sum_{j=1}^{\mathcal{N}_\gamma} \frac{\eta}{2\mathcal{N}_\gamma}(1-\omega_0)\sum_{l=0}^{\infty}\left(\frac{1}{2}\right)^l = \eta(1-\omega_0). \tag{A.166}$$

Now, substitute for ϵ in (A.138) the following,

$$\epsilon = \kappa\left(\frac{\gamma}{2}, \frac{\omega}{2}, \frac{(\omega-\omega_0)\gamma\eta}{2}\right). \tag{A.167}$$

Define the set $\Gamma_l \subset (0, \mathrm{diam}\left(\mathbb{S}_k\right)]$ as follows,

$$\Gamma_l = \left[\left(\frac{1}{2}\right)^{l+1}\mathrm{diam}\left(\mathbb{S}_k\right), \left(\frac{1}{2}\right)^{l}\mathrm{diam}\left(\mathbb{S}_k\right)\right] \tag{A.168}$$

and the set $\bigcup_{l=0}^{\infty}\Gamma_l$ contains the possible range $(0, \mathrm{diam}\left(\mathbb{S}_k\right)]$ for γ.

We have

$$J\left(\frac{\gamma}{2},\gamma,\eta\right) := \left\{x^{(m)} : \exists h\, \exists\omega, \hat{P}_m\left(B_{h,\gamma}, A_{h,\gamma/6}\right) > P\left(B_{h,0}, A_{h,\gamma}\right) + \frac{\omega\kappa\left(\gamma/2, \frac{\omega}{2}, \frac{(\omega-\omega_0)\gamma\eta}{2}\right)}{2}\right\}.$$

From (A.138), (A.146), (A.148) and (A.167), the probability in (A.138) is bounded from above by

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{\gamma\in(0,\mathrm{diam}(\mathbb{S}_k)]} J\left(\frac{\gamma}{2},\gamma,\gamma\eta\right)\right) &= \mathbb{P}\left(\bigcup_{l=0}^{\infty}\bigcup_{\gamma\in\Gamma_l} J\left(\frac{\gamma}{2},\gamma,\gamma\eta\right)\right) \\
&\leq \sum_{l=0}^{\infty}\mathbb{P}\left(\bigcup_{\gamma\in\Gamma_l} J\left(\frac{\gamma}{2},\gamma,\gamma\eta\right)\right) \\
&\leq \sum_{l=0}^{\infty}\mathbb{P}\left(J\left(\left(\frac{1}{2}\right)^{l+1}\mathrm{diam}\left(\mathbb{S}_k\right), \left(\frac{1}{2}\right)^{l+1}\mathrm{diam}\left(\mathbb{S}_k\right), \eta\left(\frac{1}{2}\right)^{l}\mathrm{diam}\left(\mathbb{S}_k\right)\right)\right)
\end{aligned} \tag{A.169}$$

where (A.169) follows from (A.150) and the fact that for all $\gamma \in \Gamma_l$ we have $\gamma/2 \leq (\frac{1}{2})^{l+1}\mathrm{diam}\left(\mathbb{S}_k\right) \leq \gamma$ and $\gamma \leq (\frac{1}{2})^{l}\mathrm{diam}\left(\mathbb{S}_k\right)$.

From (A.166), the l^{th} term in (A.169) is bounded by $\eta(1-\omega_0)(1/2)^l\mathrm{diam}\left(\mathbb{S}_k\right)$ and so (A.138) is bounded from above by

$$(1-\omega_0)\eta\mathrm{diam}\left(\mathbb{S}_k\right)\sum_{l=0}^{\infty}\left(\frac{1}{2}\right)^l \leq 2(1-\omega_0)\eta\mathrm{diam}\left(\mathbb{S}_k\right). \tag{A.170}$$

A.3.2. *Bounding (A.139).* We obtain a bound from above on the following probability,

$$\mathbb{P}\left(\exists h, \exists\gamma, \exists\omega : P\left(A_{h,\gamma}\right) - \hat{P}_m\left(A_{h,\gamma/6}\right) > \epsilon\hat{P}_m(A_{h,\gamma/6})/2P\left(B_{h,0}\,|A_{h,\gamma}\right), \nu > \omega m\right)$$

From (A.135), this is bounded from above by

$$\mathbb{P}\left(\exists h, \exists\gamma, \exists\omega : P\left(A_{h,\gamma}\right) - \hat{P}_m\left(A_{h,\gamma/6}\right) > \epsilon\omega/2P\left(B_{h,0}\,|A_{h,\gamma}\right)\right)$$

which in turn is bounded from above by

$$\mathbb{P}\left(\exists\gamma, \exists h, \exists\omega : P(A_{h,\gamma}) - \hat{P}_m(A_{h,\gamma/6}) > \epsilon\omega/2\right). \tag{A.171}$$

Define the set,

$$M'_{h,\gamma}\left(\omega_1,\omega_2,\eta\right) := \left\{x^{(m)} : P\left(A_{h,\gamma}\right) > \hat{P}_m\left(A_{h,\gamma/6}\right) + \frac{\omega_2\kappa\left(\gamma/2,\omega_1,\eta\right)}{2}\right\} \tag{A.172}$$

and let

$$M'_{j,\gamma}(\omega_1,\omega_2,\eta) := \left\{ x^{(m)} : P\left(A^{(\gamma)}_{j,2\gamma}\right) > \hat{P}_m\left(A^{(\gamma)}_{j,\gamma/2}\right) + \frac{\omega_2 \kappa(\gamma/2,\omega_1,\eta)}{2} \right\}.$$

From (A.131), Claim 26 and Claim 27, it follows that if there exists a γ and h such that $X^{(m)} \in M'_{h,\gamma}(\omega_1,\omega_2,\eta)$ then there exists a j such that $X^{(m)} \in M'_{j,\gamma}(\omega_1,\omega_2,\eta)$.

With the choice of ϵ as in (A.167), then (A.171) becomes

$$\begin{aligned}
&\mathbb{P}\left(\exists\gamma,\, \exists h,\, \exists\omega : P(A_{h,\gamma}) > \hat{P}_m(A_{h,\gamma/6}) + \epsilon\omega/2\right) \\
&= \mathbb{P}\left(\exists\gamma,\, \exists h,\, \exists\omega : X^{(m)} \in M'_{h,\gamma}(\omega/2,\omega,(\omega-\omega_0)\gamma\eta/2)\right) \\
&\leq \mathbb{P}\left(\exists\gamma,\, \exists 1 \leq j \leq \mathcal{N}_\gamma,\, \exists\omega :\, X^{(m)} \in M'_{j,\gamma}(\omega/2,\omega,(\omega-\omega_0)\gamma\eta/2)\right). \qquad \text{(A.173)}
\end{aligned}$$

The probability in (A.173) is expressed as,

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{\gamma\in(0,\mathrm{diam}(\mathbb{S}_k)]} \left\{ x^{(m)} :\, \exists 1 \leq j \leq \mathcal{N}_\gamma,\, \exists\omega,\, x^{(m)} \in M'_{j,\gamma}(\omega/2,\omega,(\omega-\omega_0)\gamma\eta/2) \right\}\right) \\
&= \mathbb{P}\left(\bigcup_{l=0}^{\infty}\bigcup_{\gamma\in\Gamma_l} \left\{ x^{(m)} :\, \exists 1 \leq j \leq \mathcal{N}_\gamma,\, \exists\omega,\, x^{(m)} \in M'_{j,\gamma}(\omega/2,\omega,(\omega-\omega_0)\gamma\eta/2) \right\}\right) \\
&\leq \sum_{l=0}^{\infty} \mathbb{P}\left(\bigcup_{\gamma\in\Gamma_l} \left\{ x^{(m)} :\, \exists 1 \leq j \leq \mathcal{N}_\gamma,\, \exists\omega,\, x^{(m)} \in M'_{j,\gamma}(\omega/2,\omega,(\omega-\omega_0)\gamma\eta/2) \right\}\right). \qquad \text{(A.174)}
\end{aligned}$$

Denote by

$$\varrho_l := (1/2^l)\mathrm{diam}(\mathbb{S}_k). \qquad \text{(A.175)}$$

For every $\gamma \in \Gamma_l$, we have $\gamma \leq \varrho_l \leq 2\gamma$ hence for any

$$\theta = [s,x] \in A^{(\gamma)}_{j,2\gamma} \qquad \text{(A.176)}$$

we have

$$\begin{aligned}
\left|f^{(\gamma)}_j(s)\right| &\geq \hat{a}(2\gamma) \\
&\geq \hat{a}(\varrho_l)
\end{aligned}$$

because $\hat{a}$ is non-decreasing.

Fix any h and let $f^{(\varrho_l)}_i$ be an element in the ϱ_l-cover C_{ϱ_l} of $\mathcal{F}$ that is closest to f_h in the l_∞-norm (3.54) and let $f^{(\gamma)}_j$ be the closest to f_h in the γ-cover C_γ. Denote by

$$\dot{A}^{(\gamma)}_{i,\gamma'} := \left\{ \theta = [s,x] : \left|f^{(\gamma)}_i(s)\right| \geq \dot{a}(\gamma') \right\}. \qquad \text{(A.177)}$$

Then, for any $\gamma \in \Gamma_l$ and any $\theta = [s,x] \in A^{(\gamma)}_{j,2\gamma}$, if $f^{(\gamma)}_j(s) \leq -\hat{a}(2\gamma)$ then

$$\begin{aligned}
f^{(\varrho_l)}_i(s) &\leq f_h(s) + \varrho_l \\
&\leq f^{(\gamma)}_j(s) + \varrho_l + \gamma \\
&\overset{(i)}{\leq} -\hat{a}(2\gamma) + \varrho_l + \gamma \\
&\overset{(ii)}{\leq} -\hat{a}(\varrho_l) + \varrho_l + \gamma \\
&\overset{(iii)}{\leq} -\hat{a}(\varrho_l) + 2\varrho_l \\
&\overset{(iv)}{\leq} -\dot{a}(\varrho_l) \qquad \text{(A.178)}
\end{aligned}$$

where (i) follows from the definition of the set $A^{(\gamma)}_{j,2\gamma}$, (ii) is from the fact that $\hat{a}$ is a non-decreasing function and $\varrho_l \leq 2\gamma$ for all $\gamma \in \Gamma_l$, (iii) follows from $\gamma \leq \varrho_l$ for $\gamma \in \Gamma_l$, and (iv) follows from (A.131) where

$-\hat{a}(\gamma)+2\gamma \leq -\dot{a}(\gamma)$ holds in particular for $\gamma=\varrho_l$. If $f_j^{(\gamma)}(s)\geq \hat{a}(2\gamma)$ then

$$\begin{aligned} f_i^{(\varrho_l)}(s) &\geq f_h(s)-\varrho_l \\ &\geq f_j^{(\gamma)}(s)-\varrho_l-\gamma \\ &\geq \hat{a}(2\gamma)-\varrho_l-\gamma \\ &\geq \hat{a}(\varrho_l)-\varrho_l-\gamma \\ &\geq \hat{a}(\varrho_l)-2\varrho_l \\ &\geq \dot{a}(\varrho_l). \end{aligned} \tag{A.179}$$

Hence from (A.177), (A.178), and (A.179) it follows that $\theta\in \dot{A}_{i,\varrho_l}^{(\varrho_l)}$. Therefore for all $\gamma\in\Gamma_l$, we have

$$A_{j,2\gamma}^{(\gamma)}\subseteq \dot{A}_{i,\varrho_l}^{(\varrho_l)}. \tag{A.180}$$

Now consider any $\theta\in A_{i,\varrho_l}^{(\varrho_l)}$. If $f_i^{(\varrho_l)}(s)\leq -\dot{a}(\varrho_l)$ then for every $\gamma\in\Gamma_l$ we have

$$\begin{aligned} f_j^{(\gamma)}(s) &\leq f_h(s)+\gamma \\ &\leq f_i^{(\varrho_l)}(s)+\gamma+\varrho_l \qquad (A.181) \\ &\overset{(i)}{\leq} -\dot{a}(\varrho_l)+\gamma+\varrho_l \\ &\overset{(ii)}{\leq} -\dot{a}(\gamma)+\gamma+\varrho_l \\ &\overset{(iii)}{\leq} -\dot{a}(\gamma)+3\gamma \\ &\overset{(iv)}{\leq} -\hat{a}(\gamma/2) \qquad (A.182) \end{aligned}$$

where (i) follows from (A.177), (ii) and (iii) are from the fact that for all $\gamma\in\Gamma_l$, $\gamma\leq\varrho_l$, and $\varrho_l\leq 2\gamma$, respectively, and (iv) follows from (A.131). If $f_i^{(\varrho_l)}(s)\geq \dot{a}(\varrho_l)$ then for every $\gamma\in\Gamma_l$ we have

$$\begin{aligned} f_j^{(\gamma)}(s) &\geq f_h(s)-\gamma \\ &\geq f_i^{(\varrho_l)}(s)-\gamma-\varrho_l \qquad (A.183) \\ &\geq \dot{a}(\varrho_l)-\gamma-\varrho_l \\ &\geq \dot{a}(\gamma)-\gamma-\varrho_l \\ &\geq \dot{a}(\gamma)-3\gamma \\ &\geq \hat{a}(\gamma/2). \qquad (A.184) \end{aligned}$$

From (A.182) and (A.184) it follows that

$$\dot{A}_{i,\varrho_l}^{(\varrho_l)}\subseteq A_{j,\gamma/2}^{(\gamma)}. \tag{A.185}$$

Define by

$$\dot{M}_{i,l}\left(\omega_1,\omega_2,\eta\right):=\left\{x^{(m)}:P\left(\dot{A}_{i,\varrho_l}^{(\varrho_l)}\right)>\hat{P}_m\left(\dot{A}_{i,\varrho_l}^{(\varrho_l)}\right)+\frac{\omega_2\kappa\left(\varrho_l,\omega_1,\eta\right)}{2}\right\}. \tag{A.186}$$

Then it follows from (A.180), (A.185) and from the fact that

$$\kappa\left(\gamma/2,\omega_1,\eta\right)\geq\kappa\left(\gamma,\omega_1,\eta\right)\geq\kappa\left(\varrho_l,\omega_1,\eta\right)$$

that for any $0\leq l\leq\infty$ and any $\gamma\in\Gamma_l$, for all $h\in\mathcal{H}$ there exists $i\in\{1,\ldots,\mathcal{N}_{\varrho_l}\}$ and $j\in\{1,\ldots,\mathcal{N}_\gamma\}$ (both depending on h) such that the following holds,

$$M'_{j,\gamma}(\omega_1,\omega_2,\eta)\subseteq\dot{M}_{i,l}(\omega_1,\omega_2,\eta).$$

Therefore, the l^{th} term in the sum (A.174) is bounded from above as follows,

$$\begin{aligned} &\mathbb{P}\left(\bigcup_{\gamma\in\Gamma_l}\left\{x^{(m)}:\exists 1\leq j\leq\mathcal{N}_\gamma,\exists\omega:\,x^{(m)}\in M'_{j,\gamma}\left(\omega/2,\omega,(\omega-\omega_0)\gamma\eta/2\right)\right\}\right) \\ &\leq \mathbb{P}\left(\left\{x^{(m)}:\exists 1\leq i\leq\mathcal{N}_{\varrho_l},\exists\omega,\,x^{(m)}\in\dot{M}_{i,l}\left(\omega/2,\omega,(\omega-\omega_0)\varrho_l\eta/2\right)\right\}\right) \end{aligned} \tag{A.187}$$

where we also used the fact that for $\gamma\in\Gamma_l$, $\gamma\leq\varrho_l$.

Now, for $\omega_1 \leq \omega \leq \omega_2$ we have

$$\dot{M}_{i,l}(\omega_1, \omega_2, \eta) \subseteq \dot{M}_{i,l}(\omega, \omega, \eta) \tag{A.188}$$

which follows from the fact that $\kappa(\gamma, \omega_1, \eta) \geq \kappa(\gamma, \omega, \eta)$ and $\omega_2 \geq \omega$. Let us fix i. Using the sets (A.155) we have

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{\omega\in[\omega_0,1]} \dot{M}_{i,l}\left(\omega/2, \omega, (\omega-\omega_0)\varrho_l\eta/2\right)\right) \\
&= \mathbb{P}\left(\bigcup_{l'=0}^{\infty}\bigcup_{\omega\in\Delta_{l'}} \dot{M}_{i,l}\left(\omega/2, \omega, (\omega-\omega_0)\varrho_l\eta/2\right)\right) \\
&\leq \sum_{l'=0}^{\infty}\mathbb{P}\left(\dot{M}_{i,l}\left(\omega_0+(1-\omega_0)\left(\frac{1}{2}\right)^{l'+1}, \omega_0+(1-\omega_0)\left(\frac{1}{2}\right)^{l'+1}, \frac{\eta\varrho_l}{2}(1-\omega_0)\left(\frac{1}{2}\right)^{l'}\right)\right)
\end{aligned} \tag{A.189}$$

where the inequality follows from (A.188) as for every $\omega \in \Delta_{l'}$ we have $\omega/2 \leq \omega_0 + (1-\omega_0)\left(\frac{1}{2}\right)^{l'+1} \leq \omega$. Define

$$q_{i,l}(\theta) := \mathbb{I}\left\{\left|f_i^{(\varrho_l)}(s)\right| \geq \dot{a}(\varrho_l)\right\} \tag{A.190}$$

and

$$Q_{i,l}(\Theta^{(m)}) := \sum_{t=1}^{m} q_{i,l}(\Theta_t). \tag{A.191}$$

Then,

$$\hat{P}_m\left(\dot{A}_{i,\varrho_l}^{(\varrho_l)}\right) = \frac{1}{m}Q_{i,l}\left(\Theta^{(m)}\right)$$

and

$$\begin{aligned}
\frac{1}{m}\mathbb{E}\left[Q_{i,l}\left(\Theta^{(m)}\right)\right] &= \frac{1}{m}\sum_{t=1}^{m}\mathbb{E}\left[q_{i,l}\left(\Theta_t\right)\right] \\
&= \mathbb{P}\left(\dot{A}_{i,\varrho_l}^{(\varrho_l)}\right).
\end{aligned}$$

We obtain

$$\mathbb{P}\left(\dot{M}_{i,l}\left(\omega,\omega,\eta\right)\right) = \mathbb{P}\left(\mathbb{E}\left[Q_{i,l}\left(S^{*(m)}\right)\right] > Q_{i,l}(S^{*(m)}) + \frac{m\omega\kappa\left(\varrho_l,\omega,\eta\right)}{2}\right). \tag{A.192}$$

From Section A.2.3, the functions $Q_{i,l}$ defined in (A.191) are Lipschitz with constant $r(k,k^*)$ hence from the argument that leads to (A.165) it follows that (A.192) is bounded from above by $\eta/\mathcal{N}_{\varrho_l}$.

Therefore (A.189) is bounded from above by

$$\frac{\eta\varrho_l}{2\mathcal{N}_{\varrho_l}}(1-\omega_0)\sum_{l'=0}^{\infty}\left(\frac{1}{2}\right)^{l'} = \frac{\eta(1-\omega_0)\varrho_l}{\mathcal{N}_{\varrho_l}}.$$

Hence, (A.187) is bounded from above as follows,

$$\begin{aligned}
&\sum_{i=1}^{\mathcal{N}_{\varrho_l}}\mathbb{P}\left(\exists\omega,\, X^{(m)} \in \dot{M}_{i,l}\left(\omega/2,\omega,(\omega-\omega_0)\varrho_l\eta/2\right)\right) \\
&\leq \sum_{i=1}^{\mathcal{N}_{\varrho_l}}\sum_{l'=0}^{\infty}\mathbb{P}\left(\dot{M}_{i,l}\left(\omega_0+(1-\omega_0)\left(\frac{1}{2}\right)^{l'+1}, \omega_0+(1-\omega_0)\left(\frac{1}{2}\right)^{l'+1}, \frac{\eta\varrho_l}{2}(1-\omega_0)\left(\frac{1}{2}\right)^{l'}\right)\right) \\
&\leq \sum_{i=1}^{\mathcal{N}_{\varrho_l}}\frac{\eta(1-\omega_0)\varrho_l}{\mathcal{N}_{\varrho_l}} \\
&= \eta(1-\omega_0)\varrho_l.
\end{aligned}$$

Substituting for ϱ_l according to (A.175) then it follows that (A.174) and, therefore, the left side of (A.173) and (A.139) are bounded from above by

$$\eta(1-\omega_0)\mathrm{diam}(\mathbb{S}_k)\sum_{l=0}^{\infty}\frac{1}{2^l} = 2\eta(1-\omega_0)\mathrm{diam}(\mathbb{S}_k).$$

Thus, with the conclusion of Section A.3.1, namely (A.170), we conclude that the sum of (A.138) and (A.139) is bounded from above by

$$4\eta(1-\omega_0)\text{diam}\left(\mathbb{S}_k\right). \tag{A.193}$$

A.3.3. *Finalizing.* From (3.58) it follows that

$$\mathcal{N}_{\gamma/2} \leq \left(2\left\lceil\frac{6\text{diam}\left(\mathbb{S}_k\right)}{\gamma}\right\rceil+1\right)^{N_{\gamma/6}}. \tag{A.194}$$

Let us substitute the right side of (A.194) for $\mathcal{N}_{\gamma/2}$ in the expression (A.147) for $\kappa\left(\gamma/2,\omega/2,(\omega-\omega_0)\gamma\eta/2\right)$ used in (A.167). We obtain

$$\begin{aligned}&\frac{4r(k,k^*)\rho(k^*,\mathfrak{l}_0)}{\omega}\sqrt{\frac{2}{m}\ln\left(\frac{2\mathcal{N}_{\gamma/2}}{(\omega-\omega_0)\gamma\eta}\right)}\\&\leq\frac{4r(k,k^*)\rho(k^*,\mathfrak{l}_0)}{\omega}\sqrt{\frac{2}{m}\left(N_{\gamma/6}\ln\left(2\left\lceil\frac{6\text{diam}\left(\mathbb{S}_k\right)}{\gamma}\right\rceil+1\right)+\ln\left(\frac{2}{(\omega-\omega_0)\gamma\eta}\right)\right)}.\end{aligned} \tag{A.195}$$

Therefore if in (A.132) we let ϵ be the right side of (A.195) then (A.132) is bounded from above by (A.193). Setting this bound to δ, solving for η, substituting for η in (A.195), using (3.16) and replacing ϵ in (A.132) by $\xi(m,\gamma,\omega,\delta)$ yields the statement of the theorem.

□

References

[1] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[2] M. Anthony and J. Ratsaby. Learning bounds via sample width for classifiers on finite metric spaces. *Theoretical Computer Science*, 529:2–10, 2014.

[3] E. Behrends. *Introduction to Markov Chains: With Special Emphasis on Rapid Mixing*. Advanced Lectures in Mathematics. Springer Fachmedien Wiesbaden, 2000.

[4] Z. Blázsik and Z. Kása. Dominating sets in de Bruijn graphs. *PU.M.A., Pure Math. Appl.*, 13(1-2):79–85, 2003.

[5] J. Chaskalovic and J. Ratsaby. Interaction of a self vibrating beam with chaotic external forces. *Comptes Rendus Mecanique*, 338(1):33–39, 2010.

[6] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley-Interscience, New York, NY, USA, 2006.

[7] N. G. de Bruijn. A combinatorial problem. *Koninklijke Nederlandsche Akademie Van Wetenschappen*, 49(6):758–764, June 1946.

[8] A. Kontorovich and R. Weiss. Uniform Chernoff and Dvoretzky-Kiefer-Wolfowitz-type inequalities for Markov chains and related processes. *Journal of Applied Probability*, 51(4):1100–1113, 2014.

[9] R. M. May. *Stability and complexity in model ecosystems*. Princeton University Press, 1974.

[10] C. D. Meyer, editor. *Matrix Analysis and Applied Linear Algebra*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2000.

[11] P. Nicolas. *Understanding Markov Chains, Examples and Applications*. Springer, 2013.

[12] J. Ratsaby. On deterministic finite state machines in random environments. *Probability in the Engineering and Informational Sciences*, pages 1–36.

[13] J. Ratsaby. An algorithmic complexity interpretation of Lin's third law of information theory. *Entropy*, 10(1):6–14, 2008.

[14] J. Ratsaby. An empirical study of the complexity and randomness of prediction error sequences. *Communications in Nonlinear Science and Numerical Simulation*, 16:2832–2844, 2011.

[15] J. Ratsaby. On the descriptional complexity of systems and their output response. *Mathematics in Engineering, Science and Aerospace*, 2(3):587–298, 2011.

[16] J. Ratsaby and I. Chaskalovic. On the algorithmic complexity of static structures. *Journal of Systems Science and Complexity*, 23(6):1037–1053, 2010.

[17] N. P. Suh. Complexity in engineering. *CIRP Annals - Manufacturing Technology*, 54(2):46 – 63, 2005.

Ariel University, ISRAEL