Appendices

Table of Contents

A Additional Background on HMMs

In this section, we define in detail the HMM settings we are interested in, including the conditions for Markov chains to converge to unique stationary distributions. Recall that a HMM is characterized by the Markov chain's *initial state distribution* and its *state transitions*, along with the *emission probabilities* of an observation given the hidden state. With finitely many states and observations, without loss of generality, states take values in $\mathcal{X} = \{1, 2, \ldots, M\}$ while observations take values in $\mathcal{O} = \{1, 2, \ldots, L\}$. The initial state distribution is denoted as $\boldsymbol{\pi} \in \mathbb{R}^M$ with π_j the probability of starting in state j, the state transitions are describe by the matrix $\mathbf{A} \in \mathbb{R}^{M \times M}$ with elements a_{ij} the probability of transitioning to state j from state i, and the emission matrix $\mathbf{B} \in \mathbb{R}^{M \times L}$ contains b_{jl} the probability of observing l when in hidden state j. The triple $\lambda = (\boldsymbol{\pi}, \mathbf{A}, \mathbf{B})$ completely parameterizes a finite-alphabet HMM.

Let $\{X_1, X_2, ...\}$ denote a discrete-time Markov chain taking values in $\mathcal{X}$ with transition matrix $\mathbf{A}$. Let $p_{ij}^{(n)} = \mathbb{P}(X_{t+n} = j | X_t = i)$ denote the n-step transition probability between states $i, j \in \mathcal{X}$. State j is said to be *accessible* from state i if there exists an integer $n \geq 1$ such that $p_{ij}^{(n)} > 0$. A subset $\mathcal{C} \subseteq \mathcal{X}$ is called *irreducible* if every pair of states $i, j \in \mathcal{C}$ is mutually accessible. The *period* of state i is defined as $c(i) = \gcd\{n \geq 1 : p_{ii}^{(n)} > 0\}$, the greatest common divisor of all possible return times. State i is *aperiodic* if $c(i) = 1$. A Markov chain is termed *geometrically ergodic* if it is irreducible and aperiodic, which guarantees convergence to a unique *stationary distribution* $\boldsymbol{\mu} \in \mathbb{R}^M$ satisfying $\boldsymbol{\mu} = \boldsymbol{\mu}\mathbf{A}$. The *mixing rate* $\rho \in [0, 1)$ is such that for all states $i, j \in \mathcal{X}$, there exists a constant $C \geq 0$ for which $|p_{ij}^{(n)} - \mu_j| \leq C\rho^n$ for all $n \geq 1$. For a finite-alphabet HMM, ρ equals λ_2, the second-largest eigenvalue of $\mathbf{A}$. We run experiments on a few non-ergodic cases, while the majority of HMMs are with ergodic state transitions to avoid dependence on the initial state.

The *entropy* $H(X)$ of a discrete random variable X is defined as $H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$. A fundamental property of entropy is that conditioning reduces uncertainty: for any two random variables X and Y, we have $H(X|Y) \leq H(X)$, with equality holding if and only if X and Y are statistically independent [1]. By applying the chain rule of entropy, the joint entropy of a stochastic process can be expressed as $H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1)$. For a Markov chain with stationary distribution $\boldsymbol{\mu}$, the *entropy rate* is defined as $H(\mathcal{X}) = \lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n) = -\sum_{i,j} \mu_i a_{ij} \log a_{ij}$, which depends solely on the transition matrix $\mathbf{A}$. We additionally define the entropy of the emission matrix $\mathbf{B}$ as $-\sum_{j,l} \mu_j b_{jl} \log b_{jl}$, which quantifies the average uncertainty in observations given the underlying states. Although the entropy rate of the observation process in a HMM has no known closed-form expression, it can be bounded as $H(O_n | O_{n-1}, X_{n-1}, \ldots, O_1, X_1) \leq H(\mathcal{O}) \leq H(O_n | O_{n-1}, \ldots, O_1)$. As $\mathbf{A}$ defines transitions from X_t to X_{t+1}, and $\mathbf{B}$ determines sampling O_t from X_t, the entropies of $\mathbf{A}$ and $\mathbf{B}$ combined help us to control the entropy lower bound of the sampled HMM sequence.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

B Additional Details of Experimental Setup

Construct A with specific mixing rate, entropy, and steady state distribution. For an ergodic Markov chain that converges to a unique stationary distribution, the stochastic matrix $\mathbf{A}$ can be decomposed into eigenvalues and eigenvectors with the ordering shown in Figure 1, where $\vec{1} \in \mathbb{R}^M$ is a vector of ones, λ_2 is the second-largest eigenvalue of $\mathbf{A}$, and $\boldsymbol{\mu}$ is the stationary distribution [2]. We leverage this decomposition to construct $\mathbf{A}$ with predefined λ_2 and $\boldsymbol{\mu}$. To determine the remaining eigenvalues and eigenvectors, we formulate an optimization problem based on the following requirements: (1) all entries of $\mathbf{A}$ are non-negative; (2) each row of $\mathbf{A}$ sums to 1; (3) $\mathbf{U}^{-1}\mathbf{U} = 1$; and (4) all remaining eigenvalues have magnitudes not exceeding λ_2. The optimization problem has the following form, where we translate the constraints above into penality terms.

$$\min_{\lambda_{3:M},\mathbf{V}_2} \sum_{i,j=1}^{M} \max\{-a_{ij}, 0\} + \sum_{j=1}^{M} \left(\left(\sum_{i=1}^{M} a_{ij} \right) - 1 \right)^2 + \sum_{i,j=1}^{M} (\mathbf{VU} - \mathbf{I})_{ij}^2 + \sum_{i=3}^{M} \max\{\lambda_i - \lambda_2, 0\}$$

$$\text{s.t.} \quad \mathbf{A} = \mathbf{V}\text{diag}(1, \lambda_2, \lambda_3, ..., \lambda_M)\mathbf{U}, \quad \mathbf{V} = [\mathbf{1} \quad \mathbf{V}_2], \quad \mathbf{U} = \begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{V}_2^\dagger \end{bmatrix}$$

This is a nonconvex problem, which we solve using first order methods with `pytorch`. We randomly initialize the free variables $\lambda_3, ..., \lambda_M$ and $\mathbf{V}_2$ and then run 5000 iterations of Adam with step size 0.01 and default values for other parameters. After the optimizer terminates, we reject instances which do not satisfy the constraints exactly. By initializing with multiple random seeds, we generate matrices spanning the desired entropy spectrum.

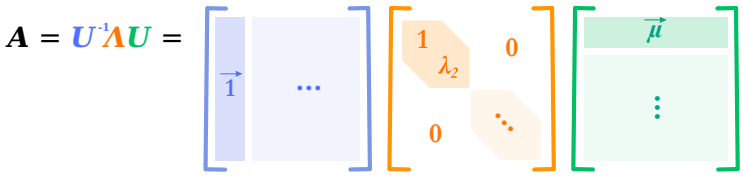

Figure 1: The singular value decomposition of ergodic unichain Markov matrix $\mathbf{A}$. The darker shaded region is pre-defined for our controlled experiments. The lighter shaded region is randomly initialized and calculated using a neural network.

Steady state distribution. We construct steady state distributions with varying skewness using the Beta distribution with $\alpha = 1$ and different values of β. When $\alpha = 1$ and $\beta = 1$, the resulting steady state distribution is uniform. As β increases, the distribution becomes increasingly skewed toward smaller state indices. Unless otherwise specified (Appendix D.3), we use a uniform steady state distribution as the default configuration.

Entropy for visualizations. The entropy definitions $H(\mathbf{A})$ and $H(\mathbf{B}, \boldsymbol{\mu})$ we introduced in Section 2.1 are used for constructing HMM parameters and sampling trajectories. For graphing main paper Figure 4 *(Left)*, we define normalized entropy considering both matrices:

$$\tilde{H}(\mathbf{A}, \mathbf{B}, \boldsymbol{\mu}) = \frac{H(\mathbf{A}) + H(\mathbf{B}, \boldsymbol{\mu})}{\log M + \log L}.$$

We define $\tilde{H}(\mathbf{A}) = H(\mathbf{A})/\log M$ and $\tilde{H}(\mathbf{B}, \boldsymbol{\mu}) = H(\mathbf{B}, \boldsymbol{\mu})/\log L$ for the main paper Figure 4 *(Middle)* and *(Right)*.

T is when LLM converges to Viterbi. The concept of "convergence", though intuitive to human eyes, requires a specific numerical definition for plots like the main paper Figure 4. We define convergence as the point where two conditions are simultaneously satisfied: (i) the accuracy difference between Viterbi and LLM is within 0.025, and (ii) LLM achieves at least 95% of Viterbi's accuracy. We use both constant and relative thresholds to ensure a strict convergence definition that accounts for different baseline performance levels across experimental conditions.

Hellinger Distance. For two discrete probability distributions $P, Q \in \mathbb{R}^L$, the Hellinger distance is defined as

$$D_{Hellinger}(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{L} (\sqrt{P_i} - \sqrt{Q_i})^2}.$$

C Details of Benchmark Models

In this section, we provide descriptions and pseudocode for the benchmark models we use (main paper Table 1). The executable code for all methods are included in supplemental materials.

Algorithm 1: Viterbi Algorithm

Input: States $\mathcal{X} = \{1, 2, \ldots, M\}$, initial distribution $\boldsymbol{\mu}$, transition matrix $\mathbf{A}$, emission matrix $\mathbf{B}$, and observation sequence $\{o_1, \ldots, o_T\}$.
Output: Most likely state sequence path $= \{x_1, \ldots, x_T\}$
Initialization: $\mathbf{P}[0][s] \leftarrow \boldsymbol{\mu}[s] \cdot \mathbf{B}[s][o_1]$ for all $s \in \mathcal{X}$;
Forward recursion: for $t = 1$ **to** $T - 1$ **do**
  for $s \in \mathcal{X}$ **do**
    $\mathbf{P}[t][s] \leftarrow \max_{r \in \mathcal{X}} \{\mathbf{P}[t-1][r] \cdot \mathbf{A}[r][s] \cdot \mathbf{B}[s][o_t]\}$;
    $\mathbf{Q}[t][s] \leftarrow \arg\max_{r \in \mathcal{X}} \{\mathbf{P}[t-1][r] \cdot \mathbf{A}[r][s] \cdot \mathbf{B}[s][o_t]\}$;
  end
end
Backtracking: path$[T-1] \leftarrow \arg\max_{s \in \mathcal{X}} \mathbf{P}[T-1][s]$;
path$[t] \leftarrow \mathbf{Q}[t+1]$[path$[t+1]$] for $t = T-2, \ldots, 0$;
return *path*

Viterbi algorithm. The Viterbi algorithm is a dynamic programming technique for efficiently finding the most likely sequence of hidden states in a Markov model, given a sequence of observations. It iteratively computes the highest probability path to each state at time t by considering all possible predecessor states at time $t-1$, their transition probabilities, and the emission probabilities of the current observation. Rather than exhaustively evaluating all M^T possible state sequences, Viterbi maintains only the M most promising paths at each time step, storing both their probabilities and the penultimate states that maximize these probabilities. After computing probabilities for all time steps, the algorithm traces backward from the most probable final state to reconstruct the optimal state sequence. We use the most probable final state and ground-truth $\mathbf{A}$ and $\mathbf{B}$ to calculate the prediction distribution of the next observation.

Algorithm 2: Compute $P(O_{t+1}|O_{t-k:t})$

Input: States $\mathcal{X} = \{1, 2, \ldots, M\}$, initial distribution $\boldsymbol{\mu}$, transition matrix $\mathbf{A}$, emission matrix $\mathbf{B}$, and observation sequence $\{o_{t-k}, \ldots, o_t\}$.
Output: Probability of next observation $P(o_{t+1}|o_{t-k:t})$
Forward pass over observation window: $\boldsymbol{\alpha}_{t-k}[s] \leftarrow \mathbf{B}[s][o_{t-k}] \cdot \boldsymbol{\mu}[s]$ for all $s \in \mathcal{X}$;
$\boldsymbol{\alpha}_i[s] \leftarrow \mathbf{B}[s][o_i] \cdot \sum_{r \in \mathcal{X}} \mathbf{A}[r][s] \cdot \boldsymbol{\alpha}_{i-1}[r]$ for $i = t-k+1, \ldots, t$, $s \in \mathcal{X}$;
Normalize to get posterior: $P(s|o_{t-k:t}) \leftarrow \frac{\boldsymbol{\alpha}_t[s]}{\sum_{s' \in \mathcal{X}} \boldsymbol{\alpha}_t[s']}$ for all $s \in \mathcal{X}$;
Prediction step: $P(s|o_{t-k:t}) \leftarrow \sum_{r \in \mathcal{X}} \mathbf{A}[r][s] \cdot P(r|o_{t-k:t})$ for all $s \in \mathcal{X}$;
Marginalize over states: $P(o_{t+1}|o_{t-k:t}) \leftarrow \sum_{s \in \mathcal{X}} \mathbf{B}[s][o_{t+1}] \cdot P(s|o_{t-k:t})$;
return $P(o_{t+1}|o_{t-k:t})$

Optimal inference with truncated memory $P(O_{t+1}|O_{t-k:t})$. The forward-based prediction algorithm computes the probability of the next observation in a hidden Markov model by using a three-step approach. First, it calculates the posterior distribution over current hidden states via the forward algorithm, recursively processing the observation window while accounting for transitions and emissions. Second, it projects this belief state forward by applying the transition matrix to compute the distribution over next possible states. Finally, it determines $P(o_{t+1}|o_{t-k:t})$ by marginalizing over all possible next states, weighting each by its emission probability.

Baum-Welch algorithm. The Baum-Welch algorithm is an expectation-maximization method for estimating hidden Markov model parameters. It iteratively alternates between computing state posteriors $\gamma_t(s)$ and transition posteriors $\xi_t(s, r)$ via forward-backward recursion (E-step), and updating parameters to maximize likelihood (M-step): setting the initial distribution to γ_1, transition

Algorithm 3: Baum-Welch Algorithm

Input: States $\mathcal{X} = \{1, 2, \ldots, M\}$, observations $\mathcal{Y} = \{1, 2, \ldots, L\}$, observation sequence $\{o_1, \ldots, o_T\}$, initial parameters $\boldsymbol{\mu}^{(0)}$, $\mathbf{A}^{(0)}$, $\mathbf{B}^{(0)}$, and threshold ϵ.
Output: Refined parameters $\boldsymbol{\mu}$, $\mathbf{A}$, $\mathbf{B}$
Initialize: $\boldsymbol{\mu} \leftarrow \boldsymbol{\mu}^{(0)}$, $\mathbf{A} \leftarrow \mathbf{A}^{(0)}$, $\mathbf{B} \leftarrow \mathbf{B}^{(0)}$, $\mathcal{L}_{\text{prev}} \leftarrow -\infty$;
repeat
  $\mathcal{L}_{\text{prev}} \leftarrow \mathcal{L}$;
  E-Step:
  Forward pass: $\boldsymbol{\alpha}_1[s] \leftarrow \mathbf{B}[s][o_1] \cdot \boldsymbol{\mu}[s]$ for all $s \in \mathcal{X}$;
  $\boldsymbol{\alpha}_t[s] \leftarrow \mathbf{B}[s][o_t] \cdot \sum_r \mathbf{A}[r][s] \cdot \boldsymbol{\alpha}_{t-1}[r]$ for $t = 2, \ldots, T$, $s \in \mathcal{X}$;
  $\mathcal{L} \leftarrow \sum_s \boldsymbol{\alpha}_T[s]$;
  Backward pass: $\boldsymbol{\beta}_T[s] \leftarrow 1$ for all $s \in \mathcal{X}$;
  $\boldsymbol{\beta}_t[s] \leftarrow \sum_r \mathbf{A}[s][r] \cdot \mathbf{B}[r][o_{t+1}] \cdot \boldsymbol{\beta}_{t+1}[r]$ for $t = T-1, \ldots, 1$, $s \in \mathcal{X}$;
  Expected counts: $\gamma_t[s] \leftarrow \frac{\boldsymbol{\alpha}_t[s] \cdot \boldsymbol{\beta}_t[s]}{\mathcal{L}}$ for $t = 1, \ldots, T$, $s \in \mathcal{X}$;
  $\xi_t[s][r] \leftarrow \frac{\boldsymbol{\alpha}_t[s] \cdot \mathbf{A}[s][r] \cdot \mathbf{B}[r][o_{t+1}] \cdot \boldsymbol{\beta}_{t+1}[r]}{\mathcal{L}}$ for $t = 1, \ldots, T-1$, $s, r \in \mathcal{X}$;
  M-Step: $\boldsymbol{\mu}[s] \leftarrow \gamma_1[s]$ for all $s \in \mathcal{X}$;
  $\mathbf{A}[s][r] \leftarrow \frac{\sum_{t=1}^{T-1} \xi_t[s][r]}{\sum_{t=1}^{T-1} \gamma_t[s]}$ for all $s, r \in \mathcal{X}$;
  $\mathbf{B}[s][v] \leftarrow \frac{\sum_{t=1}^{T} \gamma_t[s] \cdot \mathbf{1}(o_t = v)}{\sum_{t=1}^{T} \gamma_t[s]}$ for all $s \in \mathcal{X}$, $v \in \mathcal{Y}$;
until $|\mathcal{L} - \mathcal{L}_{prev}| < \epsilon$;
return $\boldsymbol{\mu}$, $\mathbf{A}$, $\mathbf{B}$

probabilities to normalized expected transitions, and emission probabilities to normalized observation counts per state. This process continues until the log-likelihood converges, yielding locally optimal parameters that maximize the probability of generating the observed sequence. We use the learned parameters to predict next observation similar to the Viterbi algorithm.

Algorithm 4: n-gram Based Next-Observation Prediction

Input: Observation sequence $O = \{o_1, \ldots, o_T\}$, context length $n-1$, smoothing parameter δ
Output: n-gram model for predicting $P(o_t | o_{t-(n-1):t-1})$
Count extraction: $\text{counts}_n, \text{counts}_{n-1} \leftarrow$ empty associative arrays;
for $t = n-1$ **to** $T-1$ **do**
  context $\leftarrow [o_{t-(n-1)}, \ldots, o_{t-1}]$;
  Increment $\text{counts}_n[\text{context} \oplus o_t]$ and $\text{counts}_{n-1}[\text{context}]$;
end
Model construction with smoothing: $V \leftarrow$ number of unique symbols in O;
for *each observed context* c *in* counts_{n-1} **do**
  for *each unique observation* o *in* O **do**
    $\text{model}[c, o] \leftarrow \frac{\text{counts}_n[c \oplus o] + \delta}{V \cdot \delta + \text{counts}_{n-1}[c]}$;
  end
end
Back-off for unseen contexts: $P_{unif}(o) \leftarrow \frac{1}{V}$ for all o;

$$P(o_t | o_{t-(n-1):t-1}) \leftarrow \begin{cases} \text{model}[[o_{t-(n-1)}, \ldots, o_{t-1}], o_t], & \text{if context observed} \\ P_{unif}(o_t), & \text{otherwise} \end{cases};$$

return *model*

n-gram. n-gram models provide an elegant, computationally efficient framework for next-observation prediction in Markov chain processes by directly estimating conditional probabilities from observed sequences. These models embody the Markov assumption that $P(O_{t+1} | O_{1:t}) \approx P(O_{t+1} | O_{t-n+2:t})$, making them particularly effective for stochastic processes where future states depend only on a limited history of previous states. For first-order Markov chains, bigram models ($n = 2$) precisely capture the underlying transition dynamics, while higher-order dependencies can be modeled by increasing n.

Algorithm 5: LSTM for Single Sequence Prediction

Input: Observation sequence $O = \{o_1, \ldots, o_T\}$, vocabulary size V, embedding dimension d, hidden dimension h, layers L, learning rate α, epochs E
Output: Trained LSTM model for $P(o_{t+1}|o_{1:t})$
Architecture: Initialize embedding layer: Embedding : $\mathbb{Z} \to \mathbb{R}^d$;
Initialize LSTM layers: LSTM : $\mathbb{R}^d \to \mathbb{R}^h$ with L layers;
Initialize output projection: Linear : $\mathbb{R}^h \to \mathbb{R}^V$;
Initialize optimizer with learning rate α;
Training: for *epoch* $= 1$ **to** E **do**
  for $t = 1$ **to** $T - 1$ **do**
    $x \leftarrow O_{1:t}$; `// Use all previous observations as context`
    $y \leftarrow O_{t+1}$; `// Next observation as target`
    Update model to maximize $P(y|x)$ via gradient descent;
  end
end
Inference: For prefix $O_{1:t}$, compute $P(o_{t+1}|O_{1:t}) = \text{softmax}(\text{model}(O_{1:t}))$;
return *model*

RNN LSTM. LSTM networks are specialized recurrent neural architectures designed to model sequential data through memory cells regulated by input, forget, and output gates. These gates control information flow, allowing LSTMs to selectively retain relevant historical patterns while discarding irrelevant information. LSTMs excel at next-observation prediction tasks by capturing both short-term correlations and long-term dependencies in the observation history. The network processes a window of prior observations sequentially, updating its hidden state to encode temporal patterns, then projects this state through a softmax layer to generate a probability distribution over possible next observations—making LSTMs particularly effective for forecasting future values in time series where the prediction depends on complex patterns spanning multiple time scales.

D Additional Synthetic Experiment Results

This section presents additional results from synthetic experiments. All methods are evaluated using the average performance over 4,096 sequences, with the exception of LSTM, which is evaluated on 16 sequences due to its high computational cost. Consequently, the LSTM results exhibit higher variance. Nonetheless, in metrics such as Hellinger distance—which account for the full output distribution rather than relying solely on the argmax for accuracy—LSTM underperforms compared to the LLM most of the time.

D.1 Varying Entropy of A

In this section, we present detailed results on varying the entropy of A matrix over 4/8/16 states and emissions, reporting accuracies and Hellinger distances.

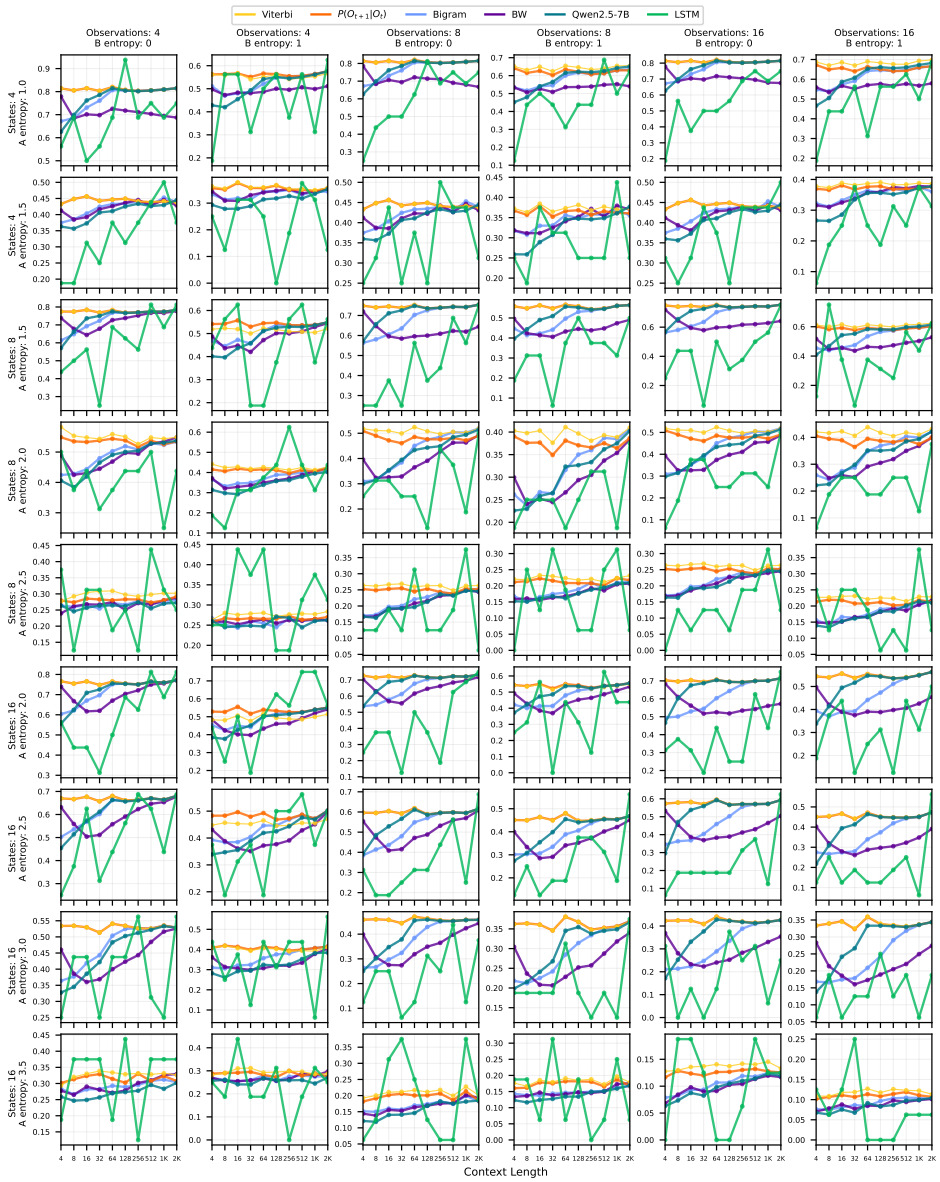

Figure 2: Accuracies of six methods across different **A** entropy, **B** entropy, number of states, and number of emissions with $\lambda_2 = 0.75$ and uniform steady state distribution.

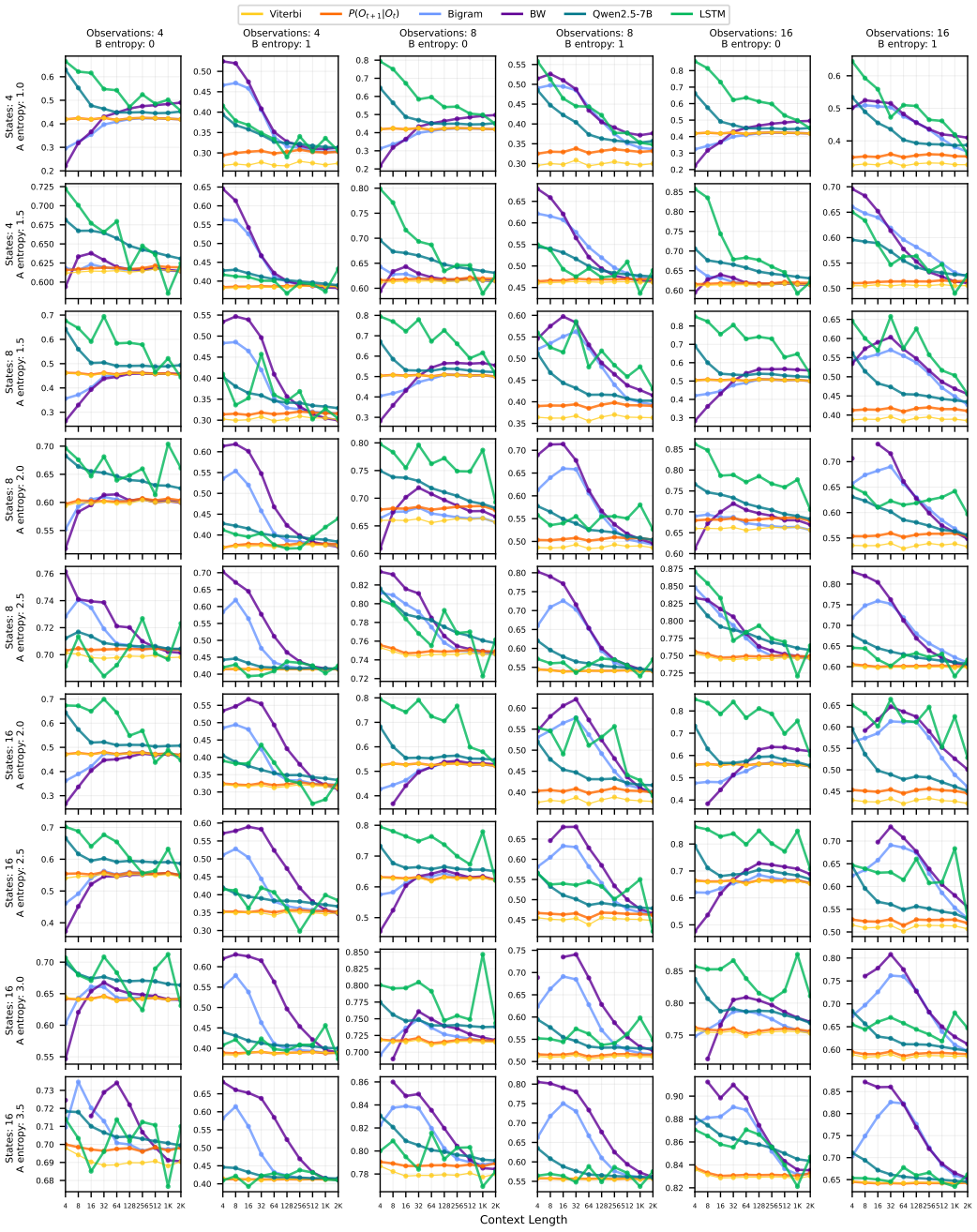

Figure 3: Hellinger distances of six methods across different $\mathbf{A}$ entropy, $\mathbf{B}$ entropy, number of states, and number of emissions with $\lambda_2 = 0.75$ and uniform steady state distribution.

D.2 Varying Mixing Rate of A

In this section, we present detailed results on varying the mixing rate (λ_2) of A matrix over 4/8/16 states and emissions, reporting accuracies and Hellinger distances.

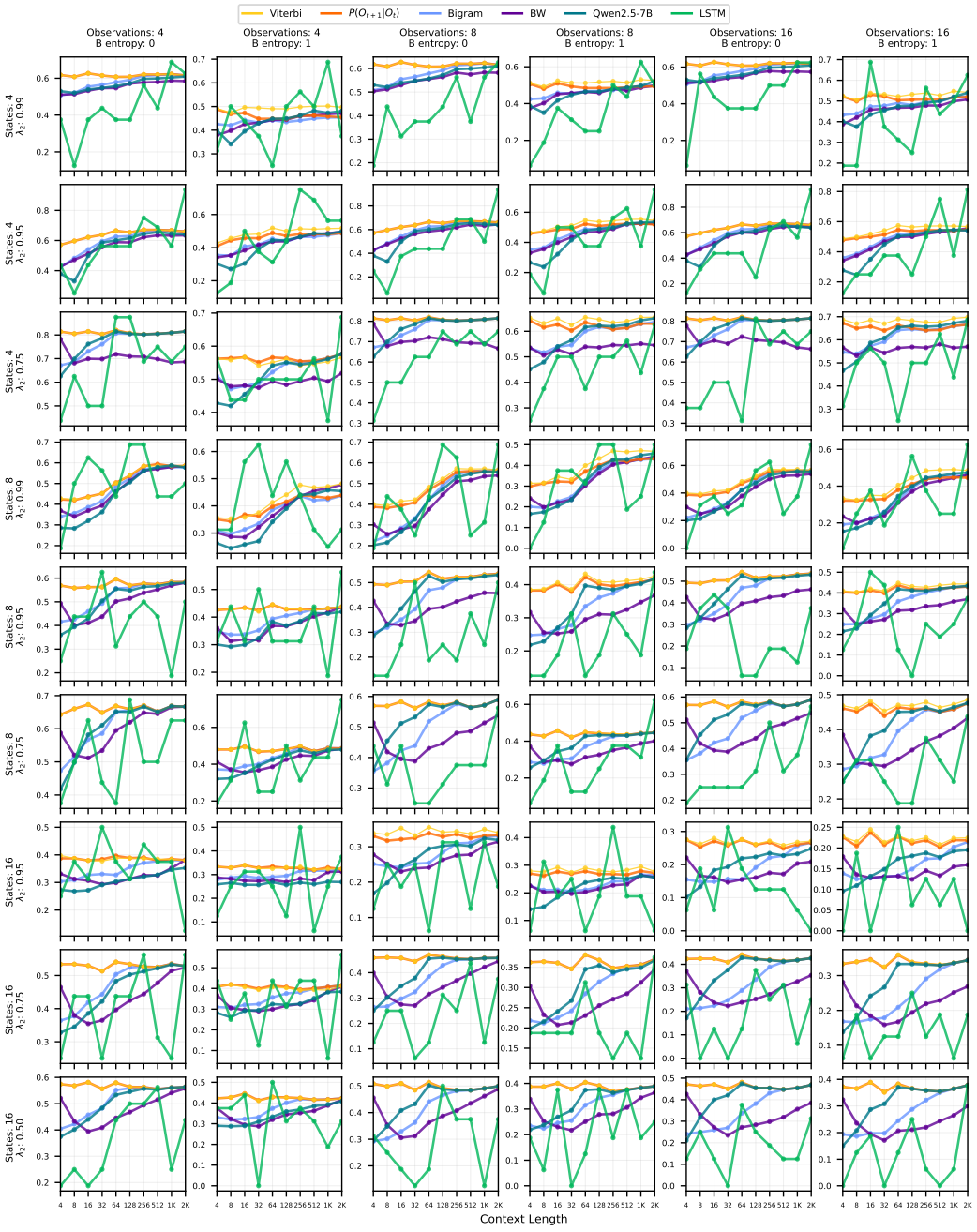

Figure 4: Accuracies of six methods across different mixing rates (λ_2), $\mathbf{B}$ entropy, number of states, and number of emissions with uniform steady state distribution and (1, 2, 3) $\mathbf{A}$ entropy for (4, 8, 16) states respectively.

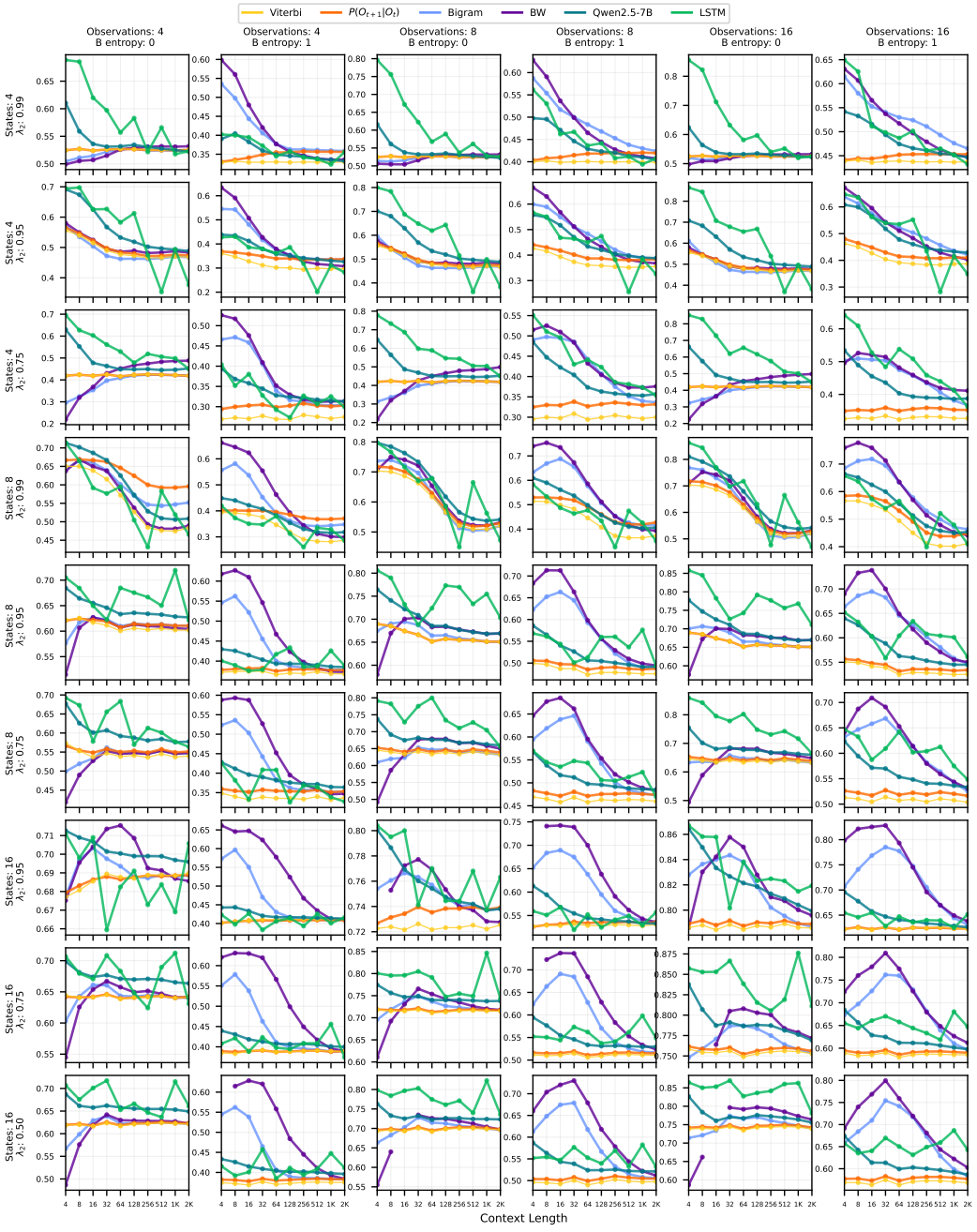

Figure 5: Hellinger distances of six methods across different mixing rates (λ_2), $\mathbf{B}$ entropy, number of states, and number of emissions with uniform steady state distribution and (1, 2, 3) $\mathbf{A}$ entropy for (4, 8, 16) states respectively.

D.3 Varying Steady State Distribution of A

In this section, we present detailed results on varying the steady state distributions of A matrix over 4/8/16 states and emissions, reporting accuracies and Hellinger distances. We construct steady states with different skewness using Beta distribution with $\alpha = 1$. Notably, with $\alpha = 1$ and $\beta = 1$, the steady state distribution is uniform. As we increase β, the distribution becomes more skewed. We test with $\beta = 1, 2, 3$, representing uniform, skewed, and very skewed respectively.

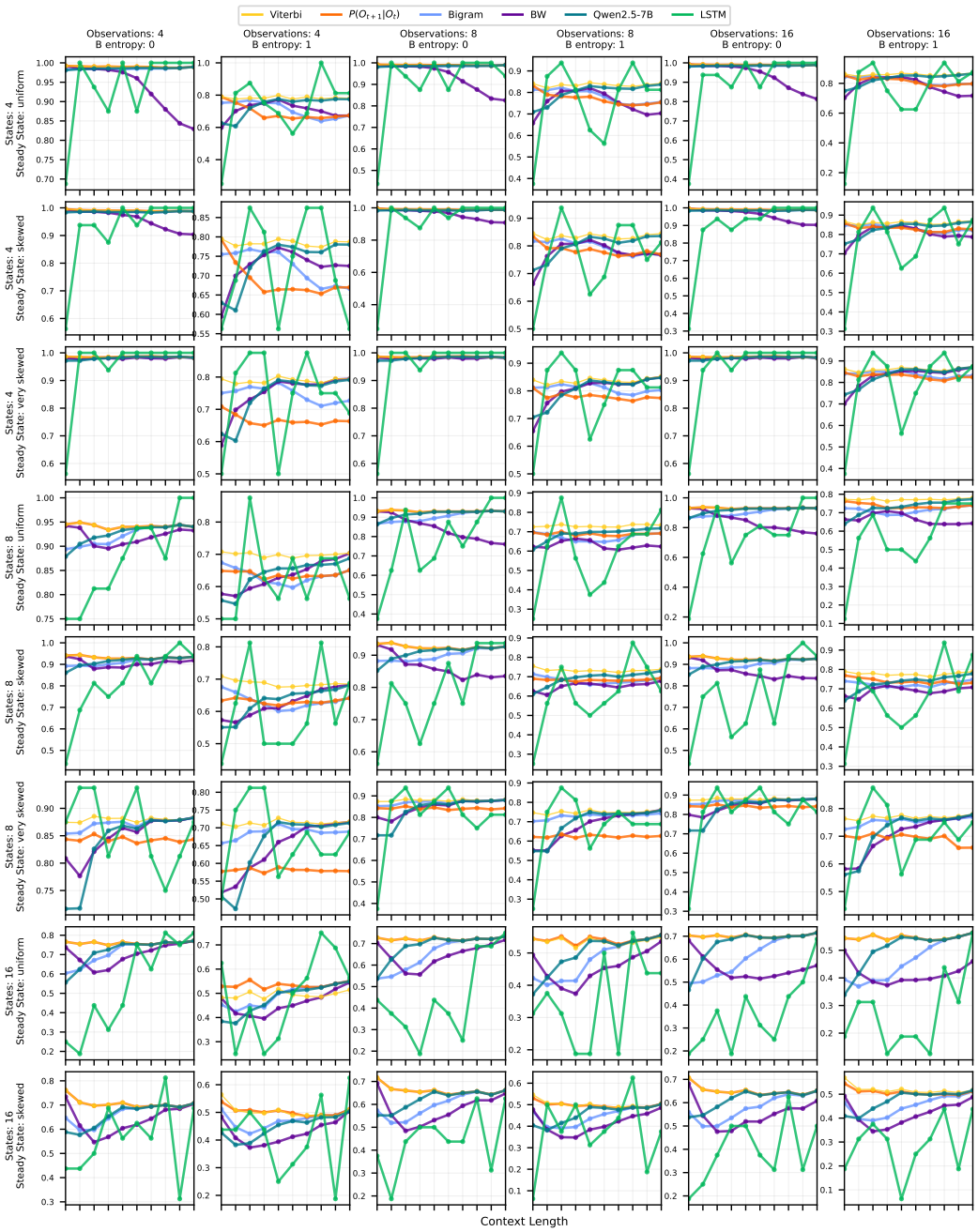

Figure 6: Accuracies of six methods across different steady state distributions, **B** entropy, number of states, and number of emissions with (0, 0.5, 2) **A** entropy for (4, 8, 16) states respectively and (0.99, 0.95, 0.75) λ_2 for (4, 8, 16) states respectively.

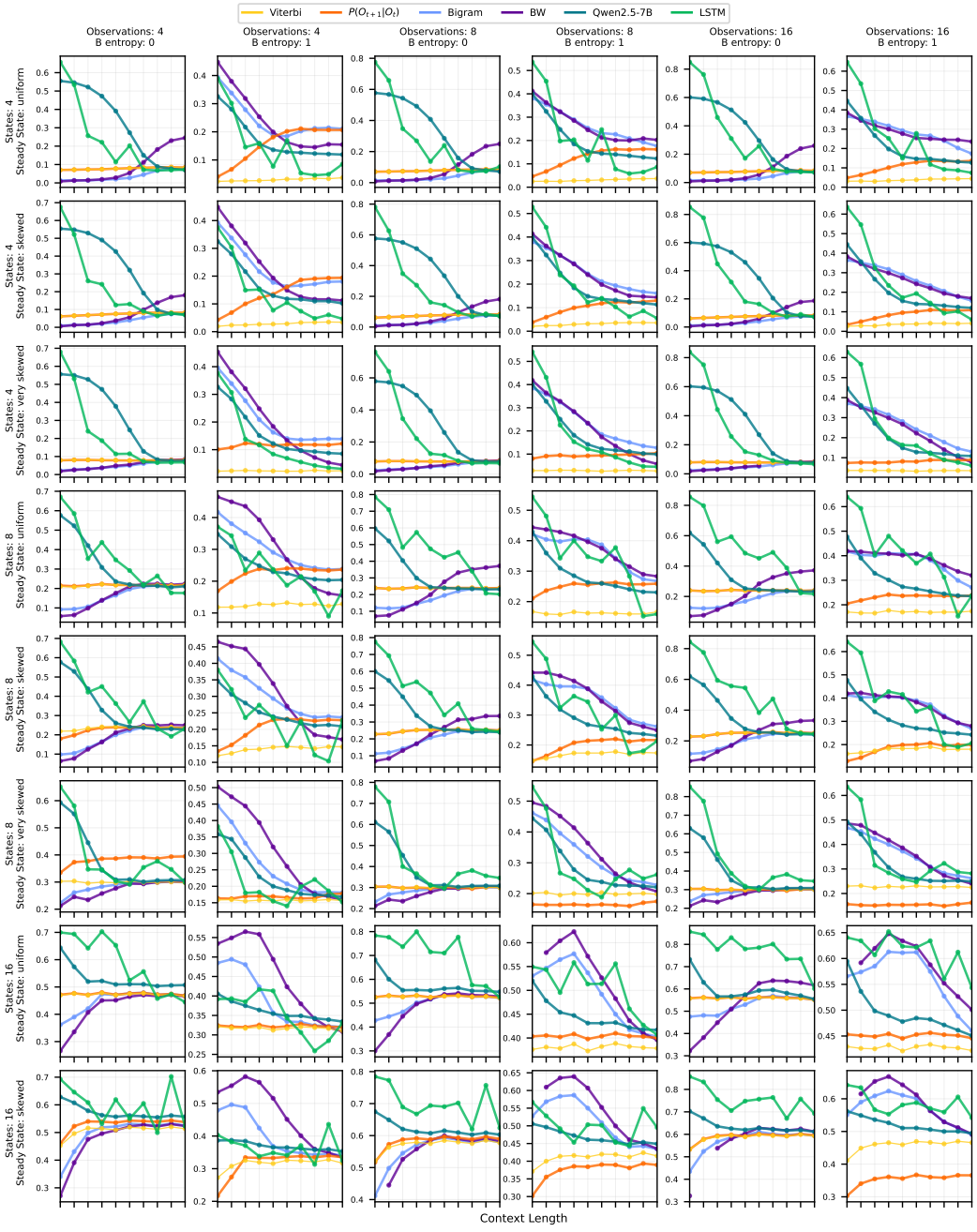

Figure 7: Hellinger distances of six methods across different steady state distributions, **B** entropy, number of states, and number of emissions with (0, 0.5, 2) **A** entropy for (4, 8, 16) states respectively and (0.99, 0.95, 0.75) λ_2 for (4, 8, 16) states respectively.

D.4 Discussions

When LLMs fail to converge. While LLMs converge to Viterbi performance efficiently under most HMM parameter settings (scaling trends summarized in Section 3.1), we identify two conditions where convergence fails or proceeds exceptionally slowly. First, when entropy of $\mathbf{A}$ or $\mathbf{B}$ is approaches its maximum ($\log M$ or $\log L$ respectively), the prediction accuracy gap ε at context length 2048 remains substantial. For instance, in the last row of Figure 2 with $M = 16$ and the entropy of $\mathbf{A}$ is 3.5 (near the maximum of $\log 16 = 4$), the LLM (Qwen2.5-7B) exhibits gradual convergence with a persistent gap. Second, when mixing is slow (λ_2 approaches 1), such as in the third-to-last row of Figure 4 with $M = 16$ and $\lambda_2 = 0.95$, a performance gap persists even at maximum context length.

Importantly, the Viterbi algorithm also struggles under these challenging conditions. Under high entropy, as shown in the last row of Figure 2, Viterbi accuracy barely exceeds random prediction (0.25/0.125/0.0625 for $L = 4/8/16$). Under slow mixing, such as in the fourth row of Figure 4 with $M = 8$ and $\lambda_2 = 0.99$, Viterbi algorithm requires context length 512 to achieve peak performance. These results demonstrate that LLM performance degradation under high entropy and slow mixing conditions reflects fundamental limits of stochastic system learnability—arising from random dynamics and long-range dependencies—that affect even optimal inference methods.

Monotonicity of LLM performance with respect to context length. We observe that LLM performance almost always improves monotonically with longer context length—a property notably absent in other learning baselines. Even excluding LSTM from this comparison (due to high variance from averaging over fewer sequences, as discussed in the first paragraph in Appendix D), both Baum-Welch and bigram models lack monotonic convergence behavior. For Baum-Welch, the accuracy graphs (Figures 2, 4, and 6) reveal multiple cases where performance "dips" and recovers, or deteriorates as context length increases. The Hellinger distance graphs (Figures 3, 5, and 7) provide clearer evidence that both BW and bigram exhibit non-monotonic learning patterns. In most cases, LLM Hellinger distance decreases monotonically, while BW and bigram display erratic behavior: sometimes experiencing early-context "bumps", other times starting very close to the ground truth emission distribution (occasionally even closer than the oracle Viterbi by empirical chance) before gradually converging to statistically sound distributions. Importantly, when BW or bigram achieve lower Hellinger distances, this does not necessarily indicate better performance—the corresponding prediction accuracy graphs often show poor results, highlighting the distinction between distributional similarity and predictive capability.

When (normalized) entropy $\tilde{H}$ is held constant, varying the number of states does not affect the LLM convergence rate. We provide concrete evidence for this claim in Figure 2, where rows 1, 3, and 6 all have the same normalized entropy $\tilde{H}(\mathbf{A}) = 0.5$. Across each column, the Qwen2.5-7B convergence curves for these three rows exhibit nearly identical shapes, demonstrating that the convergence rate depends primarily on normalized entropy rather than absolute state space size.

We emphasize that convergence rate differs from convergence target—the Viterbi performance. While the rate of improvement remains consistent across different state space sizes (when normalized entropy is fixed), larger state spaces result in lower achievable prediction accuracy due to increased task difficulty.

E Ablations on LLMs

In this section, we provide the results on the families, sizes, and tokenization of the LLMs.

E.1 LLM Size

We compare Qwen and Llama model families with seven different models. We found that their performances are similar, with slight degradation when the model size is small.

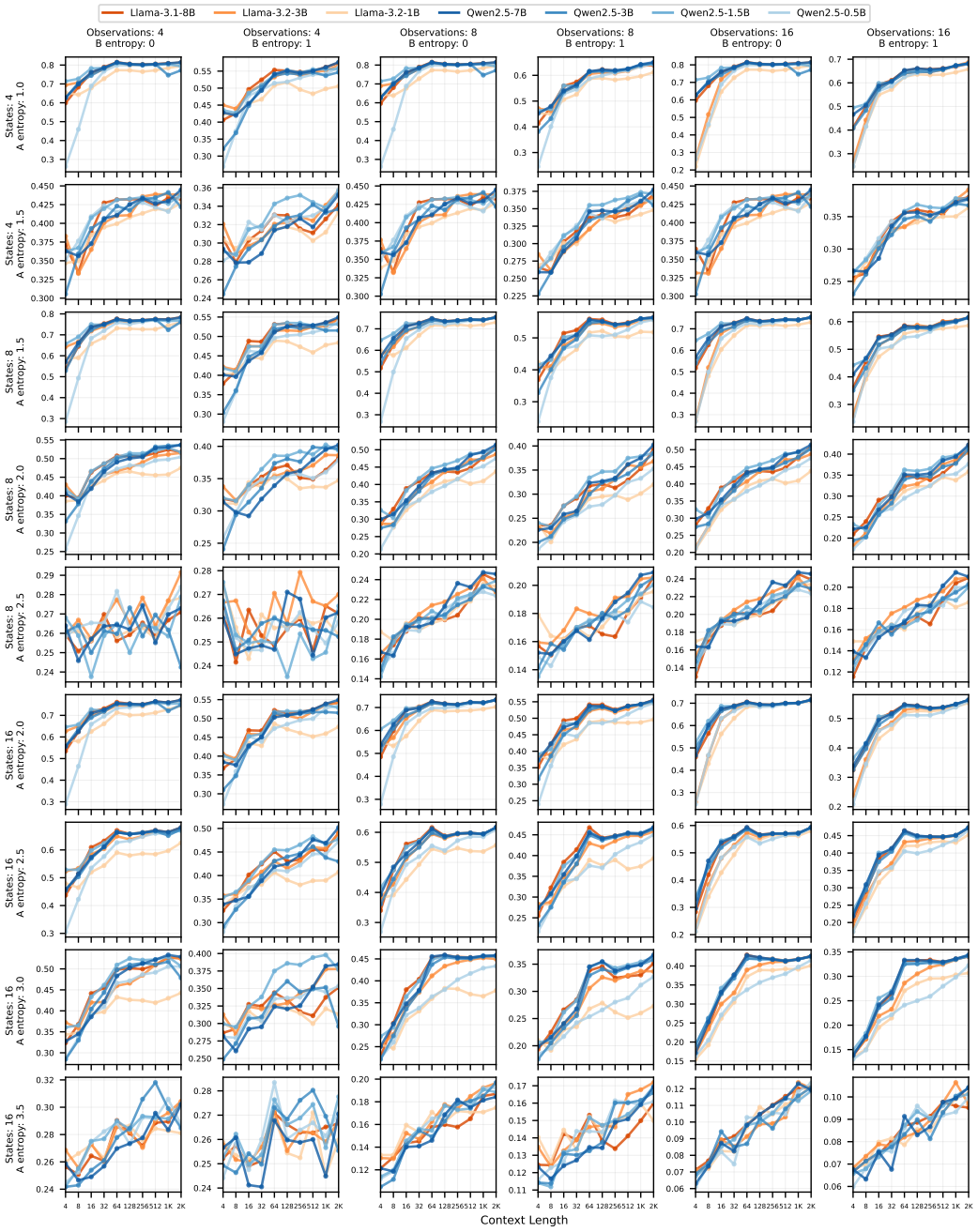

Figure 8: Accuracies of seven models across different **A** entropy, **B** entropy, number of states, and number of emissions with $\lambda_2 = 0.75$ and uniform steady state distribution. Lighter color represents smaller models. The two smallest models from each family have suboptimal performance.

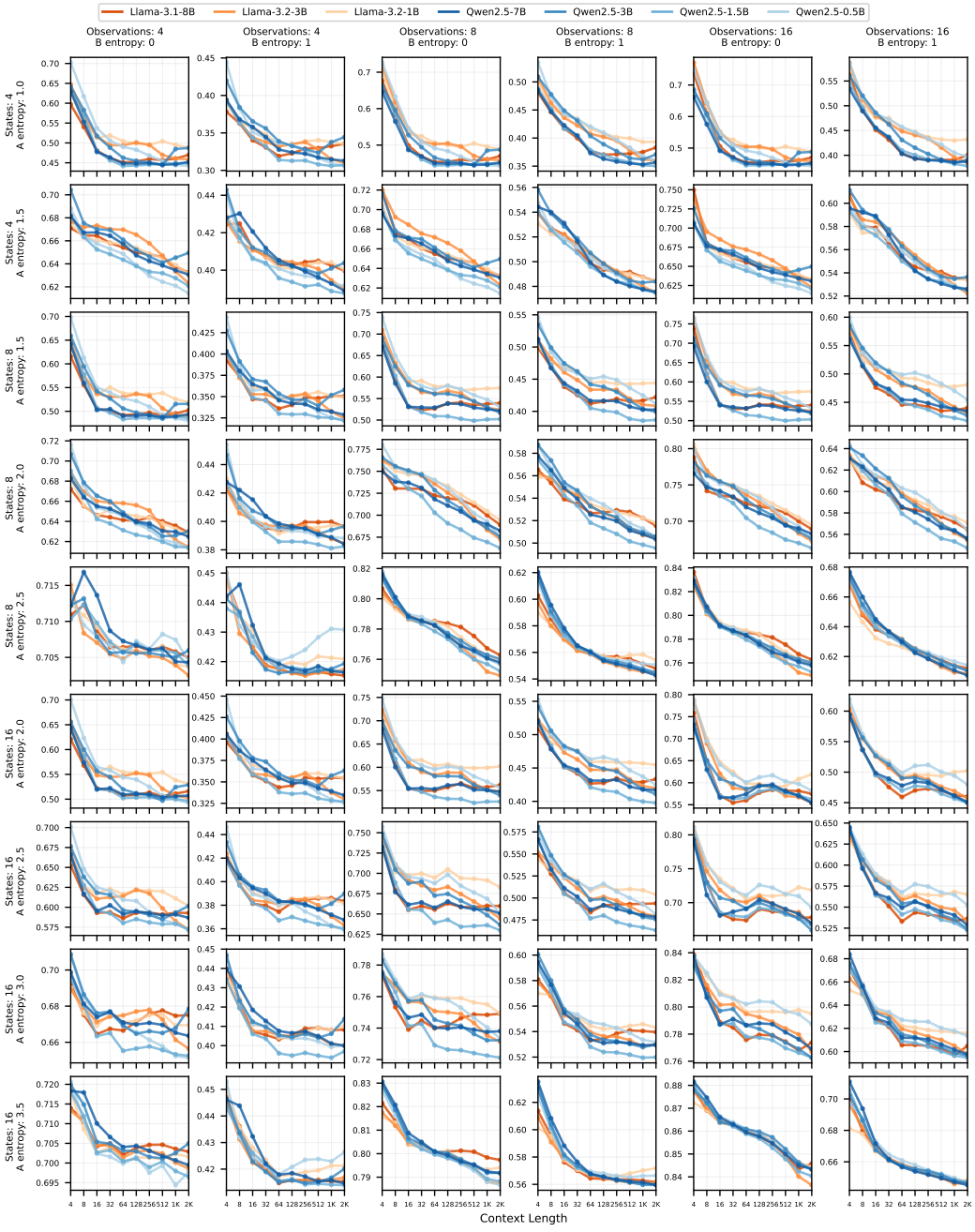

Figure 9: Hellinger distances of seven models across different **A** entropy, **B** entropy, number of states, and number of emissions with $\lambda_2 = 0.75$ and uniform steady state distribution. The models converge similarly, especially when entropy is high.

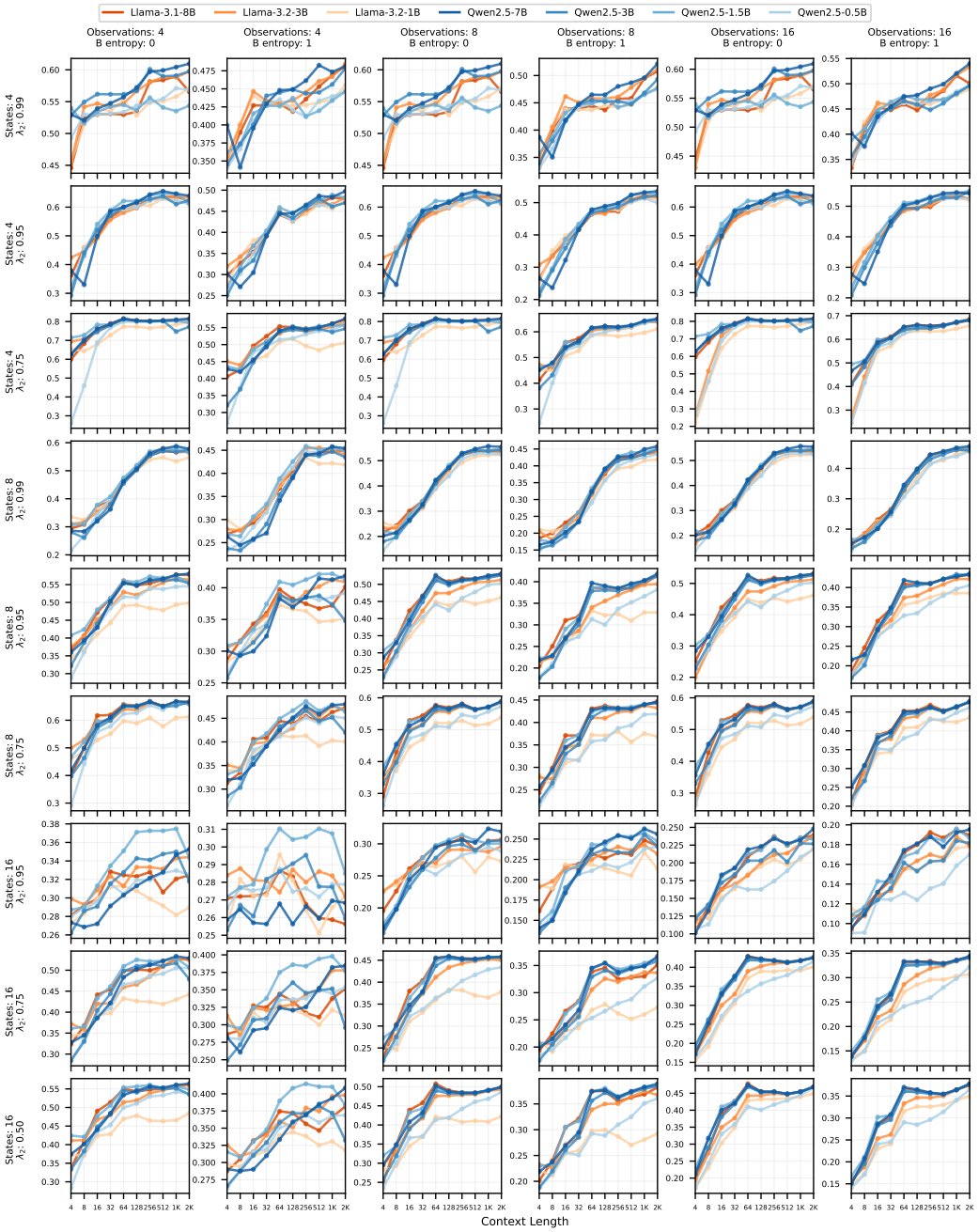

Figure 10: Accuracies of seven models across different mixing rates (λ_2), $\mathbf{B}$ entropy, number of states, and number of emissions with uniform steady state distribution and (1, 2, 3) $\mathbf{A}$ entropy for (4, 8, 16) states respectively. The two smallest models from each family have suboptimal performance, especially when mixing is fast.

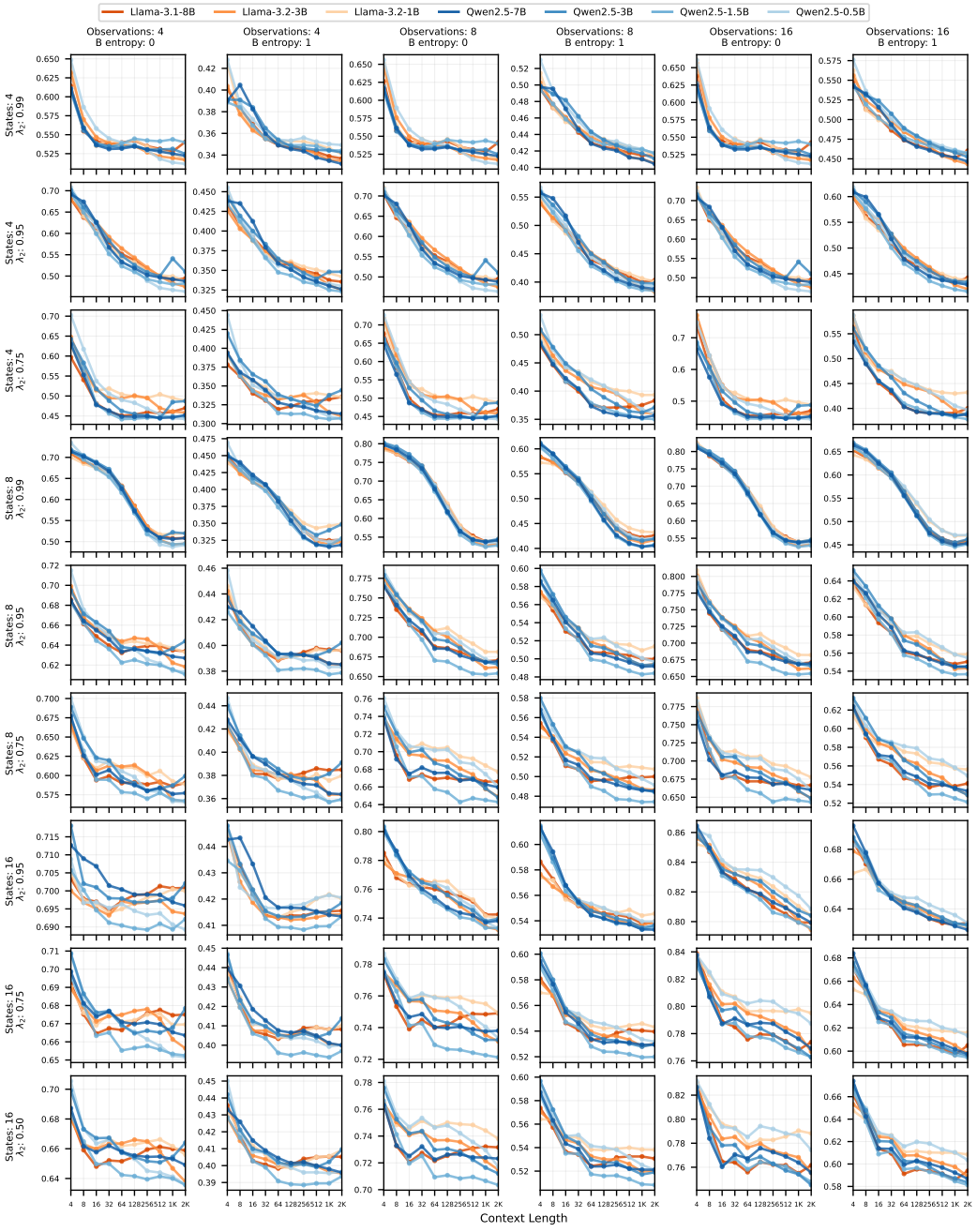

Figure 11: Hellinger distances of seven models across different mixing rates (λ_2), **B** entropy, number of states, and number of emissions with uniform steady state distribution and (1, 2, 3) **A** entropy for (4, 8, 16) states respectively. The models converge similarly, especially when mixing is slow.

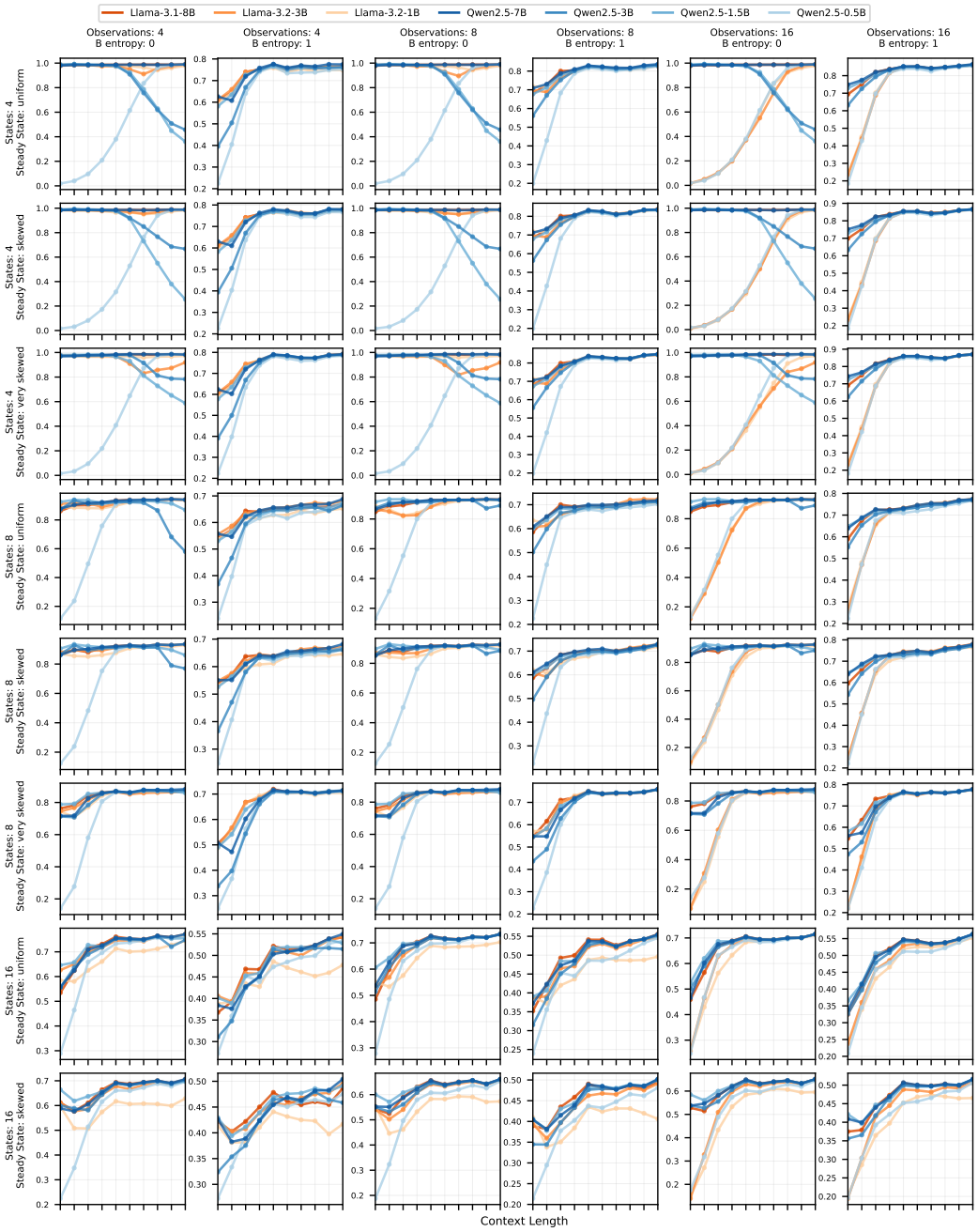

Figure 12: Accuracies of seven models across different steady state distributions, $\mathbf{B}$ entropy, number of states, and number of emissions with (0, 0.5, 2) $\mathbf{A}$ entropy for (4, 8, 16) states respectively and (0.99, 0.95, 0.75) λ_2 for (4, 8, 16) states respectively. The poor performance observed in smaller models at short context length under low $\mathbf{A}\&\mathbf{B}$ entropy settings may be attributed to the filtering of repeated n-grams during pretraining, as discussed in Appendix E.2.

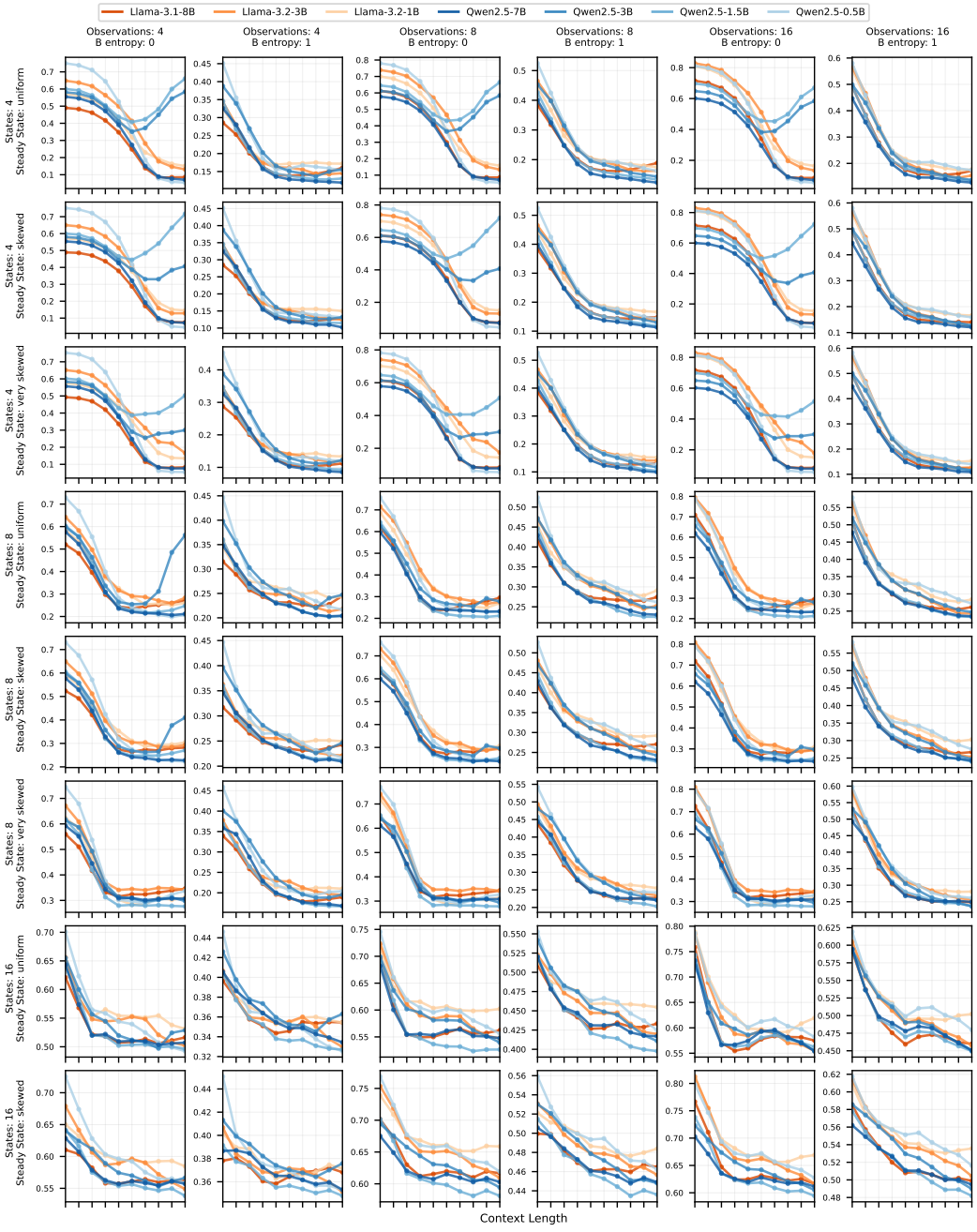

Figure 13: Hellinger distances of seven models across different steady state distributions, $\mathbf{B}$ entropy, number of states, and number of emissions with (0, 0.5, 2) $\mathbf{A}$ entropy for (4, 8, 16) states respectively and (0.99, 0.95, 0.75) λ_2 for (4, 8, 16) states respectively.

E.2 Tokenization

In this section, we evaluate three tokenization strategies: **ABC**, which encodes emissions as single letters; **123**, which encodes them as single digits; and **random**, which maps emissions to random tokens from the LLM's tokenizer. For the **random** strategy, we specifically map emissions to special tokens (!@#$). All experiments are conducted using the Qwen2.5-1.5B model, and the results are presented below.

We observe that all tokenization methods converge to similar performance levels in terms of accuracy, with **ABC** converging slightly faster when the entropy of A is large. This suggests that the choice of tokenization has limited impact on final performance. In our experiments, we adopt the **ABC** tokenization for maximum performance on the LLM. However, when the entropy of matrix A is low, **ABC** tokenization exhibits significantly lower initial accuracy and a higher Hellinger distance with short context length. We hypothesize that this is due to the increased likelihood of repetitive state sequences early in the sequence—for example, 'AAAAA...'. During pretraining, such repeated n-gram patterns are often filtered out, as they could cause loss spikes [7]. As a result, the model may have limited exposure to these patterns, leading to poor initial performance on such inputs.

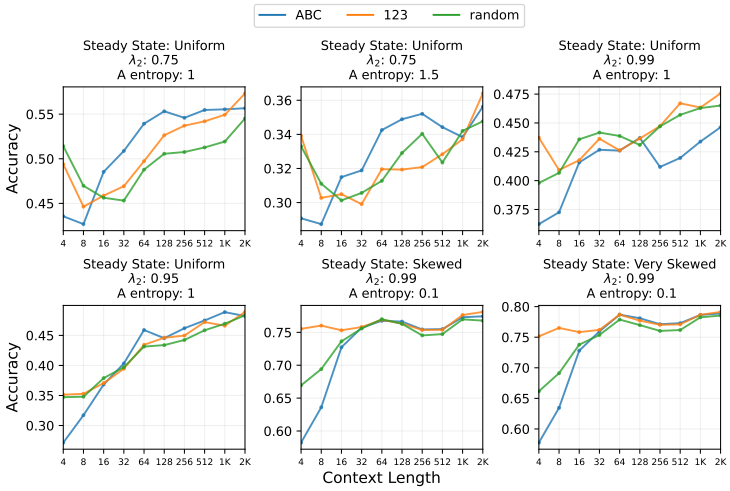

Figure 14: Accuracy of three tokenization methods across different mixing rates (λ_2), $\mathbf{A}$ entropy, and steady states with 4 states, 4 emissions, and 1 for $\mathbf{B}$ entropy.

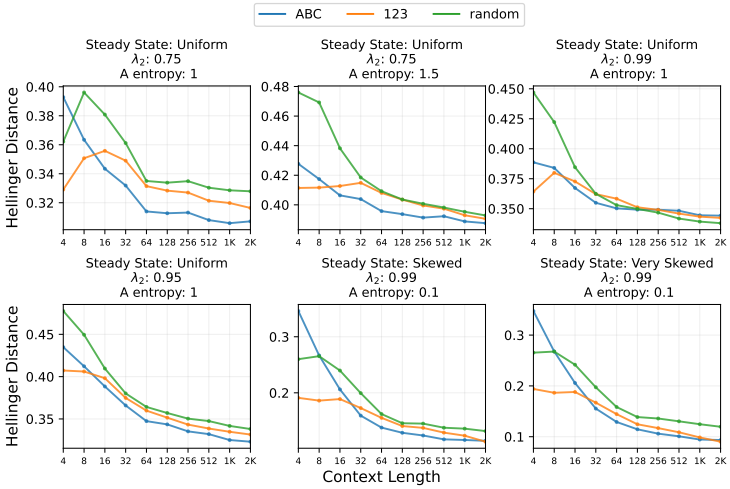

Figure 15: Hellinger distance of three tokenization methods across different mixing rates (λ_2), $\mathbf{A}$ entropy, and steady states with 4 states, 4 emissions, and 1 for $\mathbf{B}$ entropy.

F Spectral Learning HMMs for Prediction Task

Notations: We use $[\mathbf{X}]_{i,j}$ to denote the element of matrix $\mathbf{X}$ at its i-th row and j-th column. The indicator function $\mathbf{1}_{\{x=i\}}$ is 1 only when $x = i$ and is 0 otherwise. We use $\mathbf{1}_M$ to denote a vector of all 1's with dimension M. We use the notation $[L] = \{1, 2, \ldots, L\}$. $\|\cdot\|$ denotes the Frobenius norm for matrices, and depending on the context it denotes ℓ_1 or ℓ_2 norm for vectors.

Algorithm 6: Spectral Learning-Based Prediction

Input: Number of hidden states M, number of observations L, sequence $\{o_1, \ldots, o_N\}$
Output: Conditional probability distribution $\hat{P}(O_{N+1}|O_{1:N} = o_{1:N})$
Estimate empirical probabilities: for *all combinations* $i, j, n \in [L]$ **do**
  $[\hat{\mathbf{P}}_1]_i \leftarrow \frac{1}{N}\sum_{k=1}^{N} \mathbf{1}_{\{o_k=i\}}$;
  $[\hat{\mathbf{P}}_2]_{i,j} \leftarrow \frac{1}{N}\sum_{k=1}^{N} \mathbf{1}_{\{o_k=i,o_{k-1}=j\}}$;
  $[\hat{\mathbf{P}}_{3,n}]_{i,j} \leftarrow \frac{1}{N}\sum_{k=1}^{N} \mathbf{1}_{\{o_k=i,o_{k-1}=n,o_{k-2}=j\}}$;
end
Compute SVD for dimensionality reduction:
$\hat{\mathbf{U}} \leftarrow$ left singular vectors of $\hat{\mathbf{P}}_2$ corresponding to M largest singular values;
Estimate spectral parameters: $\hat{\mathbf{b}}_1 \leftarrow \hat{\mathbf{U}}^\top \hat{\mathbf{P}}_1$;
$\hat{\mathbf{b}}_\infty \leftarrow (\hat{\mathbf{P}}_2^\top \hat{\mathbf{U}})^\dagger \hat{\mathbf{P}}_1$;
for *each observation* $o \in [L]$ **do**
  $\hat{\mathbf{C}}_o \leftarrow \hat{\mathbf{U}}^\top \hat{\mathbf{P}}_{3,o} (\hat{\mathbf{U}}^\top \hat{\mathbf{P}}_2)^\dagger$;
end
Hidden state belief update: $\hat{\mathbf{b}}_1 \leftarrow$ initial belief;
for $\tau = 1$ **to** N **do**
  $\hat{\mathbf{b}}_{\tau+1} \leftarrow \frac{\hat{\mathbf{C}}_{o_\tau} \hat{\mathbf{b}}_\tau}{\hat{\mathbf{b}}_\infty^\top \hat{\mathbf{C}}_{o_\tau} \hat{\mathbf{b}}_\tau}$;
end
Conditional probability prediction: for *each possible next observation* $o_{N+1} \in [L]$ **do**
  $\hat{P}(O_{N+1} = o_{N+1}|O_{1:N} = o_{1:N}) \leftarrow \frac{\hat{\mathbf{b}}_\infty^\top \hat{\mathbf{C}}_{o_{N+1}} \hat{\mathbf{b}}_{N+1}}{\sum_{k=1}^{L} \hat{\mathbf{b}}_\infty^\top \hat{\mathbf{C}}_k \hat{\mathbf{b}}_{N+1}}$;
end
return $\hat{P}(O_{N+1}|O_{1:N} = o_{1:N})$

F.1 Preliminaries

For a Markov chain with transition matrix $\mathbf{A}$, we let $\boldsymbol{\pi} \in \mathbb{R}_+^M$ denote the initial state distribution. We assume that $\boldsymbol{\pi}$ is also the stationary distribution of the Markov chain. This can be achieved by taking samples after a burn-in time which is proportional to $\frac{1}{1-\lambda_2(\mathbf{A})}$. Note that $\boldsymbol{\pi}_t = (\mathbf{A}^t)^\top \boldsymbol{\pi}$ is essentially a convex combination of rows of matrix $\mathbf{A}^t$, then by triangle inequality, we have $\|\boldsymbol{\pi}_t - \boldsymbol{\pi}_\infty\|_1 \leq \max_{i\in[M]} \|([\mathbf{A}^t]_{i,:})^\top - \boldsymbol{\pi}_\infty\|_1$. Thus, for an ergodic Markov matrix $\mathbf{A}$, we define the following to quantify the convergence of $\|\boldsymbol{\pi}_t - \boldsymbol{\pi}_\infty\|_1$. For an ergodic Markov matrix $\mathbf{A} \in \mathbb{R}_+^{M\times M}$, let $\tau_{\mathrm{MC}} > 1$ and $\rho_{\mathrm{MC}} \in (\lambda_2(\mathbf{A}), 1)$ be two constants [5, Theorem 4.9] such that

$$\max_{i\in[M]} \|([\mathbf{A}^t]_{i,:})^\top - \boldsymbol{\pi}_\infty\|_1 \leq \tau_{\mathrm{MC}} \rho_{\mathrm{MC}}^t. \tag{1}$$

Furthermore, we define the mixing time of $\mathbf{A}$ as

$$t_{\mathrm{MC}}(\epsilon) := \min\left\{ t \in \mathbb{N} : \max_{i\in[M]} \frac{1}{2} \|([\mathbf{A}^t]_{i,:})^\top - \boldsymbol{\pi}_\infty\|_1 \leq \epsilon \right\}. \tag{2}$$

Note that $\tau(\mathbf{M})$ and τ_{MC} have similar roles except $\tau(\mathbf{M})$ is usually used to study state matrices while τ_{MC} is for Markov matrices. For a square $\mathbf{M}$, we have $\|\mathbf{M}^k\| \leq \tau(\mathbf{M})\rho(\mathbf{M})^k$, and for a Markov matrix, we have $\|\mathbf{A}^t - \mathbf{1}_M \boldsymbol{\pi}_\infty^\top\| \leq \tau_{\mathrm{MC}} \rho_{\mathrm{MC}}^t$.

F.2 Sample Complexity Analysis

In this section, we analyze the sample complexity of spectral learning algorithm (Alg 6) when the observation sequence is coming from a single trajectory. Our proof builds on [3] by modifying their analysis in Appendix A to incorporate single trajectory learning. We only present the Sample complexity analysis here and refer the reader to [3] for the remaining proofs.

F.3 Proof of Theorem 1

Fix $2 < T < N$, and recall from [3] that $[\mathbf{P}_1]_i = \mathbb{E}[\mathbf{1}_{\{o_T=i\}}]$, $[\mathbf{P}_2]_{i,j} = \mathbb{E}[\mathbf{1}_{\{o_T=i,o_{T-1}=j\}}]$, $[\mathbf{P}_{3,k}]_{i,j} = \mathbb{E}[\mathbf{1}_{\{o_T=i,o_{T-1}=k,o_{T-2}=j\}}]$, for all $k \in [L]$, when the initial distribution $\boldsymbol{\pi}$ is the stationary distribution of the Markov chain. In the following, we will present three different estimators for each of these quantities and analyze their convergence.

• **Estimation of $\mathbf{P}_1$:** Let $\bar{N} := \lfloor \frac{N}{T} \rfloor$, and without loss of generality, suppose $\frac{N}{T}$ is an integer. Suppose $\{o_T^{(1)}, \dots, o_T^{(\bar{N})}\}$ be the i.i.d. samples obtained from $\bar{N}$ independent trajectories of the HMM. We define the following three estimators of $\mathbf{P}_1$,

$$[\hat{\mathbf{P}}_1]_i = \frac{\sum_{k=1}^{N} \mathbf{1}_{\{o_k=i\}}}{N} \qquad [\hat{\mathbf{P}}_1^{(\ell)}]_i = \frac{\sum_{k=1}^{\bar{N}} \mathbf{1}_{\{o_{kT-\ell}=i\}}}{\bar{N}} \qquad [\hat{\mathbf{P}}_1^{(\perp)}]_i = \frac{\sum_{k=1}^{\bar{N}} \mathbf{1}_{\{o_T^{(k)}=i\}}}{\bar{N}}, \tag{3}$$

for all $\ell = 0, \dots, T-1$. By triangle inequality, we have

$$\|\hat{\mathbf{P}}_1 - \mathbf{P}_1\| \le \|\hat{\mathbf{P}}_1 - \hat{\mathbf{P}}_1^{(\perp)}\| + \|\hat{\mathbf{P}}_1^{(\perp)} - \mathbf{P}_1\|. \tag{4}$$

[3] showed that, with probability at least $1-\delta$, we have, $\|\hat{\mathbf{P}}_1^{(\perp)} - \mathbf{P}_1\| \lesssim \sqrt{\frac{\log(1/\delta)}{N}} + \sqrt{\frac{1}{N}}$. In the following, we will upper bound the term $\|\hat{\mathbf{P}}_1 - \hat{\mathbf{P}}_1^{(\perp)}\|$ by considering entry-wise concentration of each ℓ-th subtrajectory as follows: We have

$$[\hat{\mathbf{P}}_1^{(\ell)}]_i - [\hat{\mathbf{P}}_1^{(\perp)}]_i = \frac{\sum_{k=1}^{\bar{N}} \left(\mathbf{1}_{\{o_{kT-\ell}=i\}} - \mathbf{1}_{\{o_T^{(k)}=i\}} \right)}{\bar{N}}. \tag{5}$$

First, we observe that $\mathbb{E}\left[\mathbf{1}_{\{o_{kT-\ell}=i\}} - \mathbf{1}_{\{o_T^{(k)}=i\}}\right] 0$. Moreover, $|\mathbf{1}_{\{o_{kT-\ell}=i\}} - \mathbf{1}_{\{o_T^{(k)}=i\}}| \le 1$, almost surely. However, the summation in (5) has weakly dependent terms. Therefore, we use the Bernstein type inequality for a class of weakly dependent and bounded random variables proposed in [6]. Before that, we need to upper bound the variance of the summation in (5). Observing that $\mathbb{E}\left[[\hat{\mathbf{P}}_1^{(\ell)}]_i - [\hat{\mathbf{P}}_1^{(\perp)}]_i\right] = 0$, we have,

$$\begin{aligned}
\mathbf{Var}\left([\hat{\mathbf{P}}_1^{(\ell)}]_i - [\hat{\mathbf{P}}_1^{(\perp)}]_i\right) &:= \mathbb{E}\left[\left([\hat{\mathbf{P}}_1^{(\ell)}]_i - [\hat{\mathbf{P}}_1^{(\perp)}]_i\right)^2\right], \\
&= \mathbb{E}\left[\left([\hat{\mathbf{P}}_1^{(\ell)}]_i\right)^2\right] + \mathbb{E}\left[\left([\hat{\mathbf{P}}_1^{(\perp)}]_i\right)^2\right] - 2\,\mathbb{E}\left[[\hat{\mathbf{P}}_1^{(\ell)}]_i [\hat{\mathbf{P}}_1^{(\perp)}]_i\right].
\end{aligned} \tag{6}$$

In the following, we will upper bound each term in (6) separately. We begin with,

$$\begin{aligned}
\mathbb{E}\left[\left([\hat{\mathbf{P}}_1^{(\ell)}]_i\right)^2\right] &= \frac{1}{\bar{N}^2}\,\mathbb{E}\left[\sum_{k=1}^{\bar{N}}\sum_{k'=1}^{\bar{N}} \mathbf{1}_{\{o_{kT-\ell}=i\}} \mathbf{1}_{\{o_{k'T-\ell}=i\}}\right], \\
&= \frac{1}{\bar{N}^2}\sum_{k=1}^{\bar{N}}\sum_{k'=1}^{\bar{N}}\mathbb{E}\left[\mathbf{1}_{\{o_{kT-\ell}=i, o_{k'T-\ell}=i\}}\right], \\
&= \frac{1}{\bar{N}^2}\sum_{k=1}^{\bar{N}}\sum_{k'=1}^{\bar{N}}\mathbb{P}\left(O_{kT-\ell}=i, O_{k'T-\ell}=i\right), \\
&= \frac{[\mathbf{B}^\top \boldsymbol{\pi}]_i}{\bar{N}} + \frac{1}{\bar{N}^2}\sum_{k=1}^{\bar{N}}\sum_{\substack{k'=1 \\ k'\neq k}}^{\bar{N}}\left[\mathbf{B}^\top \mathbf{diag}(\boldsymbol{\pi}) \mathbf{A}^{|k-k'|T} \mathbf{B}\right]_{i,i}.
\end{aligned} \tag{7}$$

Next, we have,

$$\mathbb{E}\left[\left([\hat{\mathbf{P}}_1^{(\perp)}]_i\right)^2\right] = \frac{1}{\bar{N}^2}\mathbb{E}\left[\sum_{k=1}^{\bar{N}}\sum_{k'=1}^{\bar{N}}\mathbf{1}_{\{o_T^{(k)}=i\}}\mathbf{1}_{\{o_T^{(k')}=i\}}\right],$$
$$= \frac{1}{\bar{N}^2}\sum_{k=1}^{\bar{N}}\sum_{k'=1}^{\bar{N}}\mathbb{E}\left[\mathbf{1}_{\{o_T^{(k)}=i,o_T^{(k')}=i\}}\right],$$
$$= \frac{1}{\bar{N}^2}\sum_{k=1}^{\bar{N}}\sum_{k'=1}^{\bar{N}}\mathbb{P}\left(O_T^{(k)}=i, O_T^{(k')}=i\right) = \frac{[\mathbf{B}^\top\boldsymbol{\pi}]_i}{\bar{N}} + (\bar{N}-1)\frac{[\mathbf{B}^\top\boldsymbol{\pi}]_i^2}{\bar{N}}. \quad (8)$$

Lastly, we have

$$\mathbb{E}\left[\left([\hat{\mathbf{P}}_1^{(\ell)}]_i\right)\left([\hat{\mathbf{P}}_1^{(\perp)}]_i\right)\right] = \frac{1}{\bar{N}^2}\mathbb{E}\left[\sum_{k=1}^{\bar{N}}\sum_{k'=1}^{\bar{N}}\mathbf{1}_{\{o_{kT-\ell}=i\}}\mathbf{1}_{\{o_T^{(k')}=i\}}\right],$$
$$= \frac{1}{\bar{N}^2}\sum_{k=1}^{\bar{N}}\sum_{k'=1}^{\bar{N}}\mathbb{E}\left[\mathbf{1}_{\{o_{kT-\ell}=i,o_T^{(k')}=i\}}\right],$$
$$= \frac{1}{\bar{N}^2}\sum_{k=1}^{\bar{N}}\sum_{k'=1}^{\bar{N}}\mathbb{P}\left(O_{kT-\ell}=i, O_T^{(k')}=i\right) = [\mathbf{B}^\top\boldsymbol{\pi}]_i^2. \quad (9)$$

Combining (7), (8), and (9) into (6), we get

$$\mathbf{Var}\left([\hat{\mathbf{P}}_1^{(\ell)}]_i - [\hat{\mathbf{P}}_1^{(\perp)}]_i\right) = \frac{2[\mathbf{B}^\top\boldsymbol{\pi}]_i}{\bar{N}} + \frac{1}{\bar{N}^2}\sum_{k=1}^{\bar{N}}\sum_{\substack{k'=1\\k'\neq k}}^{\bar{N}}\left[\mathbf{B}^\top\mathbf{diag}(\boldsymbol{\pi})\mathbf{A}^{|k-k'|T}\mathbf{B}\right]_{i,i}$$
$$- (\bar{N}+1)\frac{[\mathbf{B}^\top\boldsymbol{\pi}]_i^2}{\bar{N}},$$
$$= \frac{2([\mathbf{B}^\top\boldsymbol{\pi}]_i - [\mathbf{B}^\top\boldsymbol{\pi}]_i^2)}{\bar{N}} + \frac{1}{\bar{N}^2}\sum_{k=1}^{\bar{N}}\sum_{\substack{k'=1\\k'\neq k}}^{\bar{N}}\left[\mathbf{B}^\top\mathbf{diag}(\boldsymbol{\pi})\mathbf{A}^{|k-k'|T}\mathbf{B}\right]_{i,i}$$
$$- (\bar{N}-1)\frac{[\mathbf{B}^\top\boldsymbol{\pi}]_i^2}{\bar{N}},$$
$$= \frac{2(\mathbf{b}_i^\top\boldsymbol{\pi} - (\mathbf{b}_i^\top\boldsymbol{\pi})^2)}{\bar{N}}$$
$$+ \frac{1}{\bar{N}^2}\sum_{k=1}^{\bar{N}}\sum_{\substack{k'=1\\k'\neq k}}^{\bar{N}}\mathbf{b}_i^\top\mathbf{diag}(\boldsymbol{\pi})\left(\mathbf{A}^{|k-k'|T} - \mathbf{1}_M\boldsymbol{\pi}^\top\right)\mathbf{b}_i,$$
$$\lesssim \frac{\mathbf{b}_i^\top\boldsymbol{\pi} - (\mathbf{b}_i^\top\boldsymbol{\pi})^2}{\bar{N}} + \frac{\|\mathbf{b}_i\|^2\tau_{\mathrm{MC}}\rho_{\mathrm{MC}}^T}{\bar{N}(1-\rho_{\mathrm{MC}}^T)} \lesssim \frac{\mathbf{b}_i^\top\boldsymbol{\pi} - (\mathbf{b}_i^\top\boldsymbol{\pi})^2}{\bar{N}}, \quad (10)$$

where $\mathbf{b}_i$ denotes the i-th column of $\mathbf{B}$ and we get the last inequality by choosing,

$$T \gtrsim \log\left(\frac{\|\mathbf{b}_i\|^2\tau_{\mathrm{MC}}}{(\mathbf{b}_i^\top\boldsymbol{\pi} - (\mathbf{b}_i^\top\boldsymbol{\pi})^2)(1-\rho_{\mathrm{MC}}^T)}\right)/(1-\rho). \quad (11)$$

Hence, using the Bernstein type inequality for weakly dependent and bounded random variables (Theorem 1 in [6]), together with (10) (11), and the observations we made right after (5), with probability at least $1-\delta$, we have

$$|[\hat{\mathbf{P}}_1^{(\ell)}]_i - [\hat{\mathbf{P}}_1^{(\perp)}]_i| \lesssim \sqrt{\frac{(\mathbf{b}_i^\top\boldsymbol{\pi} - (\mathbf{b}_i^\top\boldsymbol{\pi})^2)}{\bar{N}}\log\left(\frac{1}{\delta}\right)}. \quad (12)$$

Union bounding over all $i \in [L]$, and $\ell \in \{0, 1, \ldots, T-1\}$, with probability at least $1-\delta$, we have

$$\|[\hat{\mathbf{P}}_1^{(\ell)} - \hat{\mathbf{P}}_1^{(\perp)}\| \lesssim \sqrt{\frac{\mathbf{1}_L^\top \mathbf{B}^\top \boldsymbol{\pi} - \|\mathbf{B}^\top \boldsymbol{\pi}\|^2}{\bar{N}} \log\left(\frac{LT}{\delta}\right)}, \tag{13}$$

given $T \gtrsim \max_{i\in[L]} \left\{\log\left(\frac{\|\mathbf{b}_i\|^2 \tau_{\mathrm{MC}}}{(\mathbf{b}_i^\top \boldsymbol{\pi} - (\mathbf{b}_i^\top \boldsymbol{\pi})^2)(1-\rho_{\mathrm{MC}}^T)}\right)\right\} / (1-\rho)$. This further implies that, with probability at least $1-\delta$, the same upper bound also holds for $\|[\hat{\mathbf{P}}_1 - \hat{\mathbf{P}}_1^{(\perp)}\|$. Combining this with (4) and [3], with probability at least $1-\delta$, we have

$$\|\hat{\mathbf{P}}_1 - \mathbf{P}_1\| \lesssim \sqrt{\frac{\log(1/\delta)}{\bar{N}}} + \sqrt{\frac{1}{\bar{N}}} + \sqrt{\frac{\mathbf{1}_L^\top \mathbf{B}^\top \boldsymbol{\pi} - \|\mathbf{B}^\top \boldsymbol{\pi}\|^2}{\bar{N}} \log\left(\frac{LT}{\delta}\right)}. \tag{14}$$

• **Estimation of $\mathbf{P}_2$:** Here, we follow a similar line of reasoning as above. We begin with defining the three estimators of $\mathbf{P}_2$ as follows,

$$[\hat{\mathbf{P}}_2]_{i,j} = \frac{\sum_{k=1}^{N} \mathbf{1}_{\{o_k = i, o_{k-1} = j\}}}{N} \qquad [\hat{\mathbf{P}}_2^{(\ell)}]_{i,j} = \frac{\sum_{k=1}^{\bar{N}} \mathbf{1}_{\{o_{kT-\ell} = i, o_{kT-\ell-1} = j\}}}{\bar{N}},$$
$$[\hat{\mathbf{P}}_2^{(\perp)}]_{i,j} = \frac{\sum_{k=1}^{\bar{N}} \mathbf{1}_{\{o_T^{(k)} = i, o_{T-1}^{(k)} = j\}}}{\bar{N}} \tag{15}$$

Similar to $\mathbf{P}_1$, we consider the entry-wise concentration of each ℓ-th subtrajectory as follows,

$$[\hat{\mathbf{P}}_2^{(\ell)}]_{i,j} - [\hat{\mathbf{P}}_2^{(\perp)}]_{i,j} = \frac{\sum_{k=1}^{\bar{N}} \left(\mathbf{1}_{\{o_{kT-\ell} = i, o_{kT-\ell-1} = j\}} - \mathbf{1}_{\{o_T^{(k)} = i, o_{T-1}^{(k)} = j\}}\right)}{\bar{N}}. \tag{16}$$

Observing that $\mathbb{E}\left[[\hat{\mathbf{P}}_2^{(\ell)}]_{i,j} - [\hat{\mathbf{P}}_2^{(\perp)}]_{i,j}\right] = 0$, we have,

$$\begin{aligned}
\mathbf{Var}\left([\hat{\mathbf{P}}_2^{(\ell)}]_{i,j} - [\hat{\mathbf{P}}_2^{(\perp)}]_{i,j}\right) &= \mathbb{E}\left[\left([\hat{\mathbf{P}}_2^{(\ell)}]_{i,j} - [\hat{\mathbf{P}}_2^{(\perp)}]_{i,j}\right)^2\right], \\
&= \mathbb{E}\left[\left([\hat{\mathbf{P}}_2^{(\ell)}]_{i,j}\right)^2\right] + \mathbb{E}\left[\left([\hat{\mathbf{P}}_2^{(\perp)}]_{i,j}\right)^2\right] - 2\,\mathbb{E}\left[[\hat{\mathbf{P}}_2^{(\ell)}]_{i,j}[\hat{\mathbf{P}}_2^{(\perp)}]_{i,j}\right].
\end{aligned} \tag{17}$$

In the following, we will upper bound each term in (17) separately. We begin with,

$$\begin{aligned}
\mathbb{E}\left[\left([\hat{\mathbf{P}}_2^{(\ell)}]_{i,j}\right)^2\right] &= \frac{1}{\bar{N}^2}\,\mathbb{E}\left[\sum_{k=1}^{\bar{N}}\sum_{k'=1}^{\bar{N}} \mathbf{1}_{\{o_{kT-\ell}=i, o_{kT-\ell-1}=j\}} \mathbf{1}_{\{o_{k'T-\ell}=i, o_{k'T-\ell-1}=j\}}\right], \\
&= \frac{1}{\bar{N}^2}\sum_{k=1}^{\bar{N}}\sum_{k'=1}^{\bar{N}} \mathbb{E}\left[\mathbf{1}_{\{o_{kT-\ell}=i, o_{kT-\ell-1}=j, o_{k'T-\ell}=i, o_{k'T-\ell-1}=j\}}\right], \\
&= \frac{1}{\bar{N}^2}\sum_{k=1}^{\bar{N}}\sum_{k'=1}^{\bar{N}} \mathbb{P}\left(O_{kT-\ell}=i, O_{kT-\ell-1}=j, O_{k'T-\ell}=i, O_{k'T-\ell-1}=j\right), \\
&= \frac{\boldsymbol{\pi}^\top \mathbf{D}_{j,i} \mathbf{1}_M}{\bar{N}} + \frac{1}{\bar{N}^2}\sum_{k=1}^{\bar{N}}\sum_{\substack{k'=1 \\ k'\neq k}}^{\bar{N}} \boldsymbol{\pi}^\top \mathbf{D}_{j,i} \mathbf{A}^{|k-k'|T-1} \mathbf{D}_{j,i} \mathbf{1}_M,
\end{aligned} \tag{18}$$

where, given the i-th column $\mathbf{b}_i$, and the j-th column $\mathbf{b}_j$ of $\mathbf{B}$, we define

$$\mathbf{D}_{j,i} := \mathbf{diag}\left(\mathbf{b}_j\right) \mathbf{A}\, \mathbf{diag}\left(\mathbf{b}_i\right). \tag{19}$$

Next, we have

$$\begin{aligned}
\mathbb{E}\left[\left([\hat{\mathbf{P}}_2^{(\perp)}]_{i,j}\right)^2\right] &= \frac{1}{\bar{N}^2}\,\mathbb{E}\left[\sum_{k=1}^{\bar{N}}\sum_{k'=1}^{\bar{N}} \mathbf{1}_{\{o_T^{(k)}=i, o_{T-1}^{(k)}=j\}}\mathbf{1}_{\{o_T^{(k')}=i, o_{T-1}^{(k')}=j\}}\right], \\
&= \frac{1}{\bar{N}^2}\sum_{k=1}^{\bar{N}}\sum_{k'=1}^{\bar{N}}\mathbb{E}\left[\mathbf{1}_{\{o_T^{(k)}=i, o_{T-1}^{(k)}=j, o_T^{(k')}=i, o_{T-1}^{(k')}=j\}}\right], \\
&= \frac{1}{\bar{N}^2}\sum_{k=1}^{\bar{N}}\sum_{k'=1}^{\bar{N}}\mathbb{P}\left(O_T^{(k)}=i, O_{T-1}^{(k)}=j, O_T^{(k')}=i, O_{T-1}^{(k')}=j\right), \\
&= \frac{\boldsymbol{\pi}^\top\mathbf{D}_{j,i}\mathbf{1}_M}{\bar{N}} + (\bar{N}-1)\frac{\left(\boldsymbol{\pi}^\top\mathbf{D}_{j,i}\mathbf{1}_M\right)^2}{\bar{N}}. \qquad (20)
\end{aligned}$$

Lastly, we have

$$\begin{aligned}
\mathbb{E}\left[\left([\hat{\mathbf{P}}_2^{(\ell)}]_{i,j}\right)\left([\hat{\mathbf{P}}_2^{(\perp)}]_{i,j}\right)\right] &= \frac{1}{\bar{N}^2}\,\mathbb{E}\left[\sum_{k=1}^{\bar{N}}\sum_{k'=1}^{\bar{N}} \mathbf{1}_{\{o_{kT-\ell}=i, o_{kT-\ell-1}=j\}}\mathbf{1}_{\{o_T^{(k')}=i, o_{T-1}^{(k')}=j\}}\right], \\
&= \frac{1}{\bar{N}^2}\sum_{k=1}^{\bar{N}}\sum_{k'=1}^{\bar{N}}\mathbb{E}\left[\mathbf{1}_{\{o_{kT-\ell}=i, o_{kT-\ell-1}=j, o_T^{(k')}=i, o_{T-1}^{(k')}=j\}}\right], \\
&= \frac{1}{\bar{N}^2}\sum_{k=1}^{\bar{N}}\sum_{k'=1}^{\bar{N}}\mathbb{P}\left(O_{kT-\ell}=i, O_{kT-\ell-1}=j, O_T^{(k')}=i, O_{T-1}^{(k')}=j\right), \\
&= \left(\boldsymbol{\pi}^\top\mathbf{D}_{j,i}\mathbf{1}_M\right)^2. \qquad (21)
\end{aligned}$$

Combining (18), (20), and (21) into (17), we get

$$\begin{aligned}
\mathbf{Var}\left([\hat{\mathbf{P}}_2^{(\ell)}]_{i,j} - [\hat{\mathbf{P}}_2^{(\perp)}]_{i,j}\right) &= \frac{2\boldsymbol{\pi}^\top\mathbf{D}_{j,i}\mathbf{1}_M}{\bar{N}} + \frac{1}{\bar{N}^2}\sum_{k=1}^{\bar{N}}\sum_{\substack{k'=1\\k'\neq k}}^{\bar{N}}\boldsymbol{\pi}^\top\mathbf{D}_{j,i}\mathbf{A}^{|k-k'|T-1}\mathbf{D}_{j,i}\mathbf{1}_M \\
&\quad - (\bar{N}+1)\frac{\left(\boldsymbol{\pi}^\top\mathbf{D}_{j,i}\mathbf{1}_M\right)^2}{\bar{N}}, \\
&= \frac{2\left(\boldsymbol{\pi}^\top\mathbf{D}_{j,i}\mathbf{1}_M - (\boldsymbol{\pi}^\top\mathbf{D}_{j,i}\mathbf{1}_M)^2\right)}{\bar{N}} \\
&\quad + \frac{1}{\bar{N}^2}\sum_{k=1}^{\bar{N}}\sum_{\substack{k'=1\\k'\neq k}}^{\bar{N}}\boldsymbol{\pi}^\top\mathbf{D}_{j,i}\left(\mathbf{A}^{|k-k'|T-1} - \mathbf{1}_M\boldsymbol{\pi}^\top\right)\mathbf{D}_{j,i}\mathbf{1}_M, \\
&\lesssim \frac{2\left(\boldsymbol{\pi}^\top\mathbf{D}_{j,i}\mathbf{1}_M - (\boldsymbol{\pi}^\top\mathbf{D}_{j,i}\mathbf{1}_M)^2\right)}{\bar{N}} + \frac{\|\boldsymbol{\pi}^\top\mathbf{D}_{j,i}\|\|\mathbf{D}_{j,i}\mathbf{1}_M\|\tau_{\mathrm{MC}}\rho_{\mathrm{MC}}^{T-1}}{\bar{N}(1-\rho_{\mathrm{MC}}^T)}, \\
&\lesssim \frac{\boldsymbol{\pi}^\top\mathbf{D}_{j,i}\mathbf{1}_M - (\boldsymbol{\pi}^\top\mathbf{D}_{j,i}\mathbf{1}_M)^2}{\bar{N}}, \qquad (22)
\end{aligned}$$

where we get the last inequality by choosing,

$$T \gtrsim 1 + \log\left(\frac{\|\boldsymbol{\pi}^\top\mathbf{D}_{j,i}\|\|\mathbf{D}_{j,i}\mathbf{1}_M\|\tau_{\mathrm{MC}}}{(\boldsymbol{\pi}^\top\mathbf{D}_{j,i}\mathbf{1}_M - (\boldsymbol{\pi}^\top\mathbf{D}_{j,i}\mathbf{1}_M)^2)\,(1-\rho_{\mathrm{MC}}^T)}\right)/(1-\rho). \qquad (23)$$

Hence, using similar line of reasoning as we did in the case of P_1, with probability at least $1-\delta$, we have

$$\|[\hat{\mathbf{P}}_2^{(\ell)} - \hat{\mathbf{P}}_2^{(\perp)}\| \lesssim \sqrt{\frac{\sum_{i,j=1}^{L}\left(\boldsymbol{\pi}^\top\mathbf{D}_{j,i}\mathbf{1}_M - (\boldsymbol{\pi}^\top\mathbf{D}_{j,i}\mathbf{1}_M)^2\right)}{\bar{N}}\log\left(\frac{L^2T}{\delta}\right)}, \qquad (24)$$

given $T \gtrsim 1+\max_{i,j\in[L]}\left\{\log\left(\frac{\|\boldsymbol{\pi}^\top\mathbf{D}_{j,i}\|\|\mathbf{D}_{j,i}\mathbf{1}_M\|\tau_{\text{MC}}}{(\boldsymbol{\pi}^\top\mathbf{D}_{j,i}\mathbf{1}_M-(\boldsymbol{\pi}^\top\mathbf{D}_{j,i}\mathbf{1}_M)^2)(1-\rho_{\text{MC}}^T)}\right)\right\}/(1-\rho)$. This further implies that, with probability at least $1-\delta$, the same upper bound also holds for $\|[\hat{\mathbf{P}}_2-\hat{\mathbf{P}}_2^{(\perp)}\|$. Combining this with the triangle inequality and [3], with probability at least $1-\delta$, we have

$$\|\hat{\mathbf{P}}_2-\mathbf{P}_2\| \lesssim \sqrt{\frac{\log(1/\delta)}{\bar{N}}}+\sqrt{\frac{1}{\bar{N}}}+\sqrt{\frac{\sum_{i,j=1}^{L}(\boldsymbol{\pi}^\top\mathbf{D}_{j,i}\mathbf{1}_M-(\boldsymbol{\pi}^\top\mathbf{D}_{j,i}\mathbf{1}_M)^2)}{\bar{N}}\log\left(\frac{L^2T}{\delta}\right)}. \tag{25}$$

• **Estimation of $\mathbf{P}_3$:** Here, we follow a similar line of reasoning as above. We begin with defining the three estimators of $\mathbf{P}_3$ as follows,

$$[\hat{\mathbf{P}}_{3,n}]_{i,j}=\frac{\sum_{k=1}^{N}\mathbf{1}_{\{o_k=i,o_{k-1}=n,o_{k-2}=j\}}}{N} \qquad [\hat{\mathbf{P}}_{3,n}^{(\ell)}]_{i,j}=\frac{\sum_{k=1}^{\bar{N}}\mathbf{1}_{\{o_{kT-\ell}=i,o_{kT-\ell-1}=n,o_{kT-\ell-2}=j\}}}{\bar{N}},$$
$$[\hat{\mathbf{P}}_{3,n}^{(\perp)}]_{i,j}=\frac{\sum_{k=1}^{\bar{N}}\mathbf{1}_{\{o_T^{(k)}=i,o_{T-1}^{(k)}=no_{T-2}^{(k)}=j\}}}{\bar{N}} \tag{26}$$

Following the same line of reasoning as we did in the case of $\mathbf{P}_2$, with probability at least $1-\delta$, we have

$$\|\hat{\mathbf{P}}_{3,n}-\mathbf{P}_{3,n}\| \lesssim \sqrt{\frac{\log(1/\delta)}{\bar{N}}}+\sqrt{\frac{1}{\bar{N}}}+\sqrt{\frac{\sum_{i,j,n=1}^{L}(\boldsymbol{\pi}^\top\mathbf{D}_{j,n,i}\mathbf{1}_M-(\boldsymbol{\pi}^\top\mathbf{D}_{j,n,i}\mathbf{1}_M)^2)}{\bar{N}}\log\left(\frac{L^3T}{\delta}\right)}, \tag{27}$$

provided that,

$$T \gtrsim 2+\max_{i,j,n\in[L]}\left\{\log\left(\frac{\|\boldsymbol{\pi}^\top\mathbf{D}_{j,n,i}\|\|\mathbf{D}_{j,n,i}\mathbf{1}_M\|\tau_{\text{MC}}}{(\boldsymbol{\pi}^\top\mathbf{D}_{j,n,i}\mathbf{1}_M-(\boldsymbol{\pi}^\top\mathbf{D}_{j,n,i}\mathbf{1}_M)^2)\,(1-\rho_{\text{MC}}^T)}\right)\right\}/(1-\rho), \tag{28}$$

where, given the i-th column $\mathbf{b}_i$, the j-th column $\mathbf{b}_j$ and the n-th column $\mathbf{b}_n$ of $\mathbf{B}$, we define

$$\mathbf{D}_{j,n,i}:=\textbf{diag}\,(\mathbf{b}_j)\,\mathbf{A}\,\textbf{diag}\,(\mathbf{b}_n)\,\mathbf{A}\,\textbf{diag}\,(\mathbf{b}_i) \tag{29}$$

• **Finalizing the proof:** Theorem 1 follows by repeating the proof of Theorem 7 in [3], with the i.i.d. estimators replaced by the single trajectory estimators, and the values of $\epsilon_1, \epsilon_{2,1}$ and $\epsilon_{3,x,1}$ replaced by,

$$\epsilon_1 \lesssim \sqrt{\frac{\log(1/\delta)}{\bar{N}}}+\sqrt{\frac{1}{\bar{N}}}+\sqrt{\frac{\mathbf{1}_L^\top\mathbf{B}^\top\boldsymbol{\pi}-\|\mathbf{B}^\top\boldsymbol{\pi}\|^2}{\bar{N}}\log\left(\frac{LT}{\delta}\right)},$$
$$\epsilon_{2,1} \lesssim \sqrt{\frac{\log(1/\delta)}{\bar{N}}}+\sqrt{\frac{1}{\bar{N}}}+\sqrt{\frac{\sum_{i,j=1}^{L}(\boldsymbol{\pi}^\top\mathbf{D}_{j,i}\mathbf{1}_M-(\boldsymbol{\pi}^\top\mathbf{D}_{j,i}\mathbf{1}_M)^2)}{\bar{N}}\log\left(\frac{L^2T}{\delta}\right)},$$
$$\epsilon_{3,x,1} \lesssim \sqrt{\frac{\log(1/\delta)}{\bar{N}}}+\sqrt{\frac{1}{\bar{N}}}+\sqrt{\frac{\sum_{i,j,n=1}^{L}(\boldsymbol{\pi}^\top\mathbf{D}_{j,n,i}\mathbf{1}_M-(\boldsymbol{\pi}^\top\mathbf{D}_{j,n,i}\mathbf{1}_M)^2)}{\bar{N}}\log\left(\frac{L^3T}{\delta}\right)},$$

where $\bar{N}=\lfloor\frac{N}{T}\rfloor=\mathcal{O}\left(N(1-\lambda_2(\mathbf{A}))\right)$. The proof is completed by upper bounding the Hellinger-distance in terms of KL-distance.

G Additional Real World Experiments

We design an additional experiment using real-world datasets to validate our findings. We artificially simulate different emission entropy levels for the same underlying hidden transition process by controlling the amount of information included in the observation sequence. Using complete information corresponds to low emission entropy, while limiting information artificially increases emission entropy.

We use the IBL decision-making mice dataset [4]. In our LLM in-context learning experiment, we implement four ablation conditions that vary the information presented in each trial: (i) "choice only"; (ii) "choice reward"; (iii) "stimulus choice"; (iv) "stimulus choice reward". Note that the baseline GLM-HMM uses all available information as in condition (iv). These ablations describe the same underlying mouse decision-making sequences but with varying levels of environmental state detail.

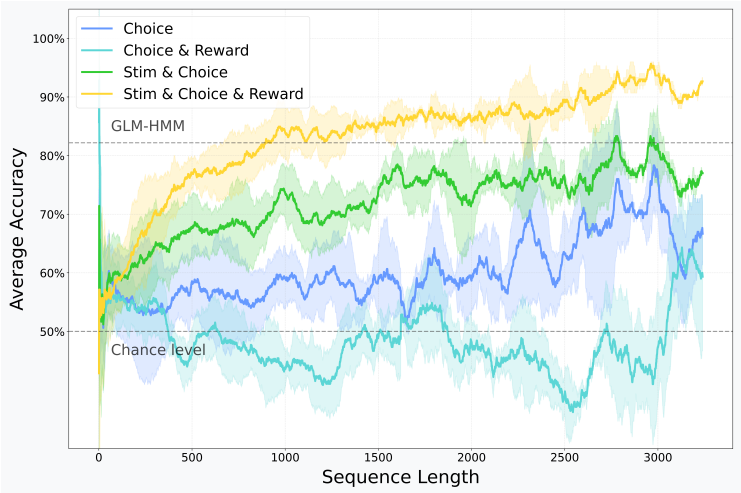

Figure 16: LLM in-context learning prediction accuracy for mice decision-making task with varying types of information in the observed sequences. Each line is averaged over 7 mice, with 1-σ error bar. The model we use is Qwen2.5-7B.

The results shown in Figure 16 reveal significant differences across ablation conditions: while "stimulus choice reward" achieves performance exceeding GLM-HMM, "choice reward" is merely at chance level with its convergence trend similar to the synthetic experiments when the transitions or emissions are near random. This demonstrates that accurately modeling mouse decision-making in this task requires both stimulus and reward information.

These findings highlight a broader principle: obtaining appropriate information (corresponding to low emission entropy) is essential for successful task modeling. This experiment parallels real-world experimental design, where scientists must choose which signals to collect when studying task structure. When researchers omit critical information needed to describe a sequence, it can easily lead to incorrect conclusions about the underlying process.

References

[1] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. ISBN 0471241954.

[2] Robert G Gallager. Discrete stochastic processes. *Journal of the Operational Research Society*, 48(1):103–103, 1997.

[3] Daniel Hsu, Sham M Kakade, and Tong Zhang. A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.

[4] The International Brain Laboratory, Valeria Aguillon-Rodriguez, Dora Angelaki, Hannah Bayer, Niccolo Bonacchi, Matteo Carandini, Fanny Cazettes, Gaelle Chapuis, Anne K Churchland, Yang Dan, Eric Dewitt, Mayo Faulkner, Hamish Forrest, Laura Haetzel, Michael Häusser, Sonja B Hofer, Fei Hu, Anup Khanal, Christopher Krasniak, Ines Laranjeira, Zachary F Mainen, Guido Meijer, Nathaniel J Miska, Thomas D Mrsic-Flogel, Masayoshi Murakami, Jean-Paul Noel, Alejandro Pan-Vazquez, Cyrille Rossant, Joshua Sanders, Karolina Socha, Rebecca Terry, Anne E Urai, Hernando Vergara, Miles Wells, Christian J Wilson, Ilana B Witten, Lauren E Wool, and Anthony M Zador. Standardized and reproducible measurement of decision-making in mice. *eLife*, 10:e63711, may 2021. ISSN 2050-084X. doi: 10.7554/eLife.63711. URL https://doi.org/10.7554/eLife.63711.

[5] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

[6] Florence Merlevède, Magda Peligrad, and Emmanuel Rio. Bernstein inequality and moderate deviations under strong mixing conditions. In *High dimensional probability V: the Luminy volume*, volume 5, pages 273–293. Institute of Mathematical Statistics, 2009.

[7] Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.