Extending Conformal Prediction to Hidden Markov Models with Exact Validity via de Finetti's Theorem for Markov Chains

Buddhika Nettasinghe [1] **Samrat Chatterjee** [2] **Ramakrishna Tipireddy** [3] **Mahantesh Halappanavar** [2]

Abstract

Conformal prediction is a widely used method to quantify the uncertainty of a classifier under the assumption of exchangeability (e.g., IID data). We generalize conformal prediction to the Hidden Markov Model (HMM) framework where the assumption of exchangeability is not valid. The key idea of the proposed method is to partition the non-exchangeable Markovian data from the HMM into exchangeable blocks by exploiting the *de Finetti's Theorem for Markov Chains* discovered by Diaconis and Freedman (1980). The permutations of the exchangeable blocks are viewed as randomizations of the observed Markovian data from the HMM. The proposed method provably retains all desirable theoretical guarantees offered by the classical conformal prediction framework in both exchangeable and Markovian settings. In particular, while the lack of exchangeability introduced by Markovian samples constitutes a violation of a crucial assumption for classical conformal prediction, the proposed method views it as an advantage that can be exploited to improve the performance further. Detailed numerical and empirical results that complement the theoretical conclusions are provided to illustrate the practical feasibility of the proposed method.

1. Introduction

This paper extends the conformal prediction framework proposed in (Vovk et al., 2005) to the Hidden Markov Model (HMM) framework in a manner that provably preserves all its desirable theoretical guarantees. In particular, we focus on the following problem.

Problem of quantifying the uncertainty of an unknown HMM: Let $\{X_t\}_{t \geq 1}$ be a Markov chain taking values in the space $\mathcal{X} = \{1, 2, \ldots, n\}$ that is observed only through the memoryless observations $\{Y_t\}_{t \geq 1}$ taking values in the space $\mathcal{Y} = \{1, 2, \ldots, m\}$. The transition and observation probability matrices given by,

$$
\begin{aligned}
P_{ij} &= \mathbb{P}\{X_{t+1} = j | X_t = i\}, \, i, j \in \mathcal{X} \\
B_{ij} &= \mathbb{P}\{Y_t = j | X_t = i\}, \, i \in \mathcal{X}, j \in \mathcal{Y},
\end{aligned}
\tag{1}
$$

are unknown and they are sampled from an unknown distribution μ. Given a fully observed sequence $\{(X_t, Y_t)\}_{t=1}^{T}$, a sequence of observations $\{Y_t\}_{t=T+1}^{T+T_1}$ and a miscoverage level $\alpha \in [0, 1]$, our aim is to construct a set of sequences $\mathcal{C}_{1-\alpha} \subseteq \mathcal{X}^{T_1}$ such that,

$$
\mathbb{P}\left\{ \{X_t\}_{t=T+1}^{T+T_1} \in \mathcal{C}_{1-\alpha} \right\} \geq 1 - \alpha.
\tag{2}
$$

Currently available solution and its limitations: The *conformal prediction* framework presented in (Vovk et al., 2005) provides an elegant solution to the above problem for the special case where the sequence of states $\{X_t\}_{t=1}^{T+T_1}$ is exchangeable.[1] More specifically, conformal prediction (reviewed in Sec. 2.1) exploits exchangeability of the distribution to permute the data and construct confidence sets. However, the conformal prediction framework is not directly applicable to our problem since Markov processes are not generally exchangeable. A general solution to the above problem that provably achieves the coverage guarantee (2) is not currently available to the best of our knowledge.

Main contributions: We generalize the original conformal prediction framework proposed in (Vovk et al., 2005) to the non-exchangeable Markovian setting in a principled manner that provably preserves all theoretical guarantees offered by the original framework. The key idea behind the proposed method is to view the process $\{(X_t, Y_t)\}_{t=1}^{T+T_1}$ as a mixture of Markov chains with respect to the mixing measure μ from which the parameters are sampled and

---

[1]Tippie College of Business, University of Iowa, Iowa City IA 52242, USA [2]Pacific Northwest National Laboratory, Richland WA 99354, USA [3]Palo Alto Networks, Santa Clara CA 95054, USA. Correspondence to: Buddhika Nettasinghe <buddhika-nettasinghe@uiowa.edu>.

[1]A sequence of random variables $X_1, X_2, \ldots$ is *exchangeable*, if for any finite permutation $\pi : \{1, \ldots, T\} \rightarrow \{1, \ldots, T\}$, $\mathbb{P}\{X_1 = x_1, X_2 = x_2, \ldots, X_T = x_T\} = \mathbb{P}\{X_{\pi(1)} = x_1, X_{\pi(2)} = x_2, \ldots, X_{\pi(T)} = x_T\}$ i.e., the joint probability distribution is invariant under finite permutations.

then use a block-wise permutation scheme under which a mixture of Markov chains is exchangeable. The proposed block-wise permutation method is inspired by the notion of *partial exchangeability* and the *de Finetti's Theorem for Markov chains* presented in (Diaconis & Freedman, 1980). The proposed method can be used as a wrapper for any algorithm based on the HMM framework (filtering, prediction, smoothing) and yields the coverage guarantee (2) for any finite length of the time series. The performance of the proposed method is illustrated with detailed numerical experiments as well as empirical results based on multiple real-world datasets.

Motivation: A solution to the above problem is useful in high-stakes settings where the hidden underlying states of a Markov process need to be inferred to a given confidence level using only the noisy measurements and a past calibration (i.e., training) sequence. Examples of such high-stakes settings include inferring life-threatening health events via noisy wearable sensors used for home-based monitoring of patients considered in (Uddin, 2019; Forkan & Khalil, 2017), predicting vehicle movements in automated navigation systems considered in (Yuan et al., 2018), human safety systems in hazardous work environments considered in (Petković et al., 2019; Rashid & Behzadan, 2018), and predicting movements in the stock market for making investment decisions considered in (Hassan & Nath, 2005). Further examples of application settings are discussed in (Gupta et al., 2016; Batu et al., 2004).

2. Preliminaries and Related Work

In this section, we discuss results from the literature that are closely related to our work. In particular, we briefly review the original conformal prediction framework for exchangeable processes proposed in (Vovk et al., 2005) and some of its recent extensions.

2.1. Conformal Prediction for Exchangeable Data

The classical conformal prediction framework proposed in (Vovk et al., 2005) applies to exchangeable data. In particular, conformal prediction has been widely utilized to quantify the uncertainty of classifiers that deal with independently and identically distributed (IID) processes. Let us first briefly review how the classical conformal prediction works in the exchangeable setting.[2]

Conformal prediction algorithm proposed in (Vovk et al., 2005): Assume that we are given an exchangeable (e.g., IID) sequence $\{(X_t, Y_t)\}_{t=1}^T$. For the next observation Y_{T+1}, we aim to generate a set $\mathcal{C}_{1-\alpha}$ which contains the unknown un-

---

[2]We refer the reader to (Shafer & Vovk, 2008; Angelopoulos et al., 2023) for detailed tutorial introductions to the classical conformal prediction framework.

derlying state X_{T+1} with a confidence $1 - \alpha$. In other words, we are considering the main problem stated in Sec. 1 when the process is exchangeable and $T_1 = 1$. In order to implement conformal prediction, we first need to identify a conformity score function $\sigma : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ which quantifies the agreement between the state $X \in \mathcal{X}$ and the observation $Y \in \mathcal{Y}$: larger $\sigma(X, Y)$ indicates disagreement while smaller $\sigma(X, Y)$ indicates agreement. The conformity score function $\sigma(\cdot, \cdot)$ could be based on a given pre-trained classifier (e.g., $1 - \sigma(X, Y)$ could be the Xth element of the softmax output of a neural network for observation Y), or it can also be based on any classifier derived from the calibration sequence $\{(X_t, Y_t)\}_{t=1}^T$. Next, at time $t = T + 1$, assume $X_{T+1} = i$ and calculate

$$\hat{q}(i) = \frac{1}{T+1} \sum_{t=1}^{T+1} \mathbb{1}\left(\sigma\left(X_t, Y_t\right) \leq \sigma\left(X_{T+1}, Y_{T+1}\right)\right) \quad (3)$$

which is the quantile of $\sigma(i, Y_{T+1})$ among the conformity scores of the observation sequence $\{(X_t, Y_t)\}_{t=1}^T$ together with (i, Y_{T+1}). After calculating $\hat{q}(i)$ for each state $i \in \mathcal{X}$, the confidence set for the unknown state X_{T+1} corresponding to the observation Y_{T+1} is constructed as,

$$\mathcal{C}_{1-\alpha} = \{i \in \mathcal{X} : (T+1)\hat{q}(i) \leq \lceil (1-\alpha)(T+1) \rceil \}. \quad (4)$$

In other words, each state $i \in \mathcal{X}$ for which the conformity score $\sigma(i, Y_{T+1})$ is within the smallest $\lceil (1-\alpha)(T+1) \rceil$ among the $\{\sigma(X_t, Y_t)\}_{t=1}^T$ are included in the set $\mathcal{C}_{1-\alpha}$. The constructed confidence set $\mathcal{C}_{1-\alpha}$ is guaranteed to contain the true unknown underlying state X_{T+1} with probability $1 - \alpha$ i.e.,

$$\mathbb{P}\{X_{T+1} \in \mathcal{C}_{1-\alpha}\} \geq 1 - \alpha. \quad (5)$$

The role of exchangeability in conformal prediction: The confidence bound (5) is guaranteed to be satisfied by the classical conformal prediction framework due to the exchangeability of the sequence $\{(X_t, Y_t)\}_{t=1}^{T+1}$. To understand this, observe that,

$$\mathbb{P}\{X_{T+1} \in \mathcal{C}_{1-\alpha}\} = \mathbb{P}\left\{\hat{q}(X_{T+1}) \leq \frac{\lceil (1-\alpha)(T+1) \rceil}{T+1}\right\}$$

$$= \mathbb{P}\left\{\sum_{t=1}^{T+1} \mathbb{1}\left(\sigma\left(X_t, Y_t\right) \leq \sigma\left(X_{T+1}, Y_{T+1}\right)\right) \leq \lceil (1-\alpha)(T+1) \rceil \right\} \quad (6)$$

which follow from (3), (4). Thus, $\mathbb{P}\{X_{T+1} \in \mathcal{C}_{1-\alpha}\}$ is equal to the probability that the rank of $\sigma(X_{T+1}, Y_{T+1})$ (among $\sigma(X_t, Y_t), t = 1, 2, \ldots, T$) is less than or equal to $\lceil (1-\alpha)(T+1) \rceil$. Due to the exchangeability of the sequence $\{(X_t, Y_t)\}_{t=1}^{T+1}$, the rank of $\sigma(X_{T+1}, Y_{T+1})$ could be any integer from 1 to $T + 1$ with equal probability, im-

plying that, for $k = 1 \ldots, T + 1$,

$$\mathbb{P}\left\{\sum_{t=1}^{T+1} \mathbb{1}\left(\sigma\left(X_t, Y_t\right) \leq \sigma\left(X_{T+1}, Y_{T+1}\right)\right) \leq k\right\} = \frac{k}{T+1}. \tag{7}$$

Then (7) and (6) yield the coverage guarantee in (5).

Therefore, exchangeability is the crucial assumption in the classical conformal prediction framework. As such, the classical conformal prediction framework does not guarantee the desired coverage (5) in non-exchangeable settings such as the HMM setting that we are dealing with.

2.2. Relaxing the Assumption of Exchangeability in Conformal Prediction

Several works in the literature aimed to generalize the classical conformal prediction framework summarized in Sec. 2.1 to non-exchangeable settings. We briefly discuss some of those works that are most relevant to our work below.[3]

Our work is in particular motivated by the approach presented in (Chernozhukov et al., 2018) that proposed a method to extend the conformal prediction framework to time series data via a randomization method which accounts for potential temporal dependencies in the data. The key idea is to construct an algebraic group of block-wise permutations (instead of the element-wise permutations used in classical conformal prediction) such that each permutation in that group is likely to preserve the potential temporal dependencies. Extending the work in (Chernozhukov et al., 2018) further, (Xu & Xie, 2021) proposed to derive prediction intervals using an ensemble of bootstrapped estimators to avoid having to split data into blocks. However, when exchangeability assumption fails, the approaches presented in (Chernozhukov et al., 2018; Xu & Xie, 2021) are only approximately valid (i.e., not guaranteed to satisfy the bound (2)). In contrast, the aim of our work is to devise a method that is guaranteed to satisfy the bound (2) in any unknown HMM. In particular, the approach we present is based on constructing a block-wise permutation (which exploits the *de Finetti's theorem for Markov chains* reviewed in Sec. 3.1) that adapts to the observed sequence $\{(X_t, Y_t)\}_{t=1}^T$ in a manner that guarantees the exact exchangeability of the blocks. As a consequence, the method that we propose is exactly valid in the sense that it is guaranteed to achieve the bound (2).

In another direction, (Cherubin & Nouretdinov, 2016; Stankeviciute et al., 2021) proposed to apply the classical conformal prediction algorithm as a solution when multiple independent calibration sequences are available. However,

---

[3]We refer the reader to (Fontana et al., 2023) for a comprehensive review of more versions of the classical conformal prediction framework.

our problem (stated in Sec. 1) assumes that only one realized sequence is available and since a Markov chain is not exchangeable in general, the classical conformal prediction framework is not applicable to our setting.

Making conformal prediction robust to changes in the underlying distributions has also received significant attention in the literature that focuses on generalizing conformal prediction. When the distributions of calibration states and test states are both exchangeable but different from each other, (Tibshirani et al., 2019) proposed an approach based on re-weighting the calibration data with a likelihood ratio (of test and training distributions). (Cauchois et al., 2020; Gibbs & Candes, 2021; Barber et al., 2022) proposed methods for even more general settings such as arbitrary number of changes in both the state and observation distributions, etc. In contrast to (Tibshirani et al., 2019; Cauchois et al., 2020; Gibbs & Candes, 2021; Barber et al., 2022), our aim is to extend the conformal prediction framework specifically to HMMs with exact validity (instead of approximate validity) for any sequence length. Additionally, our approach is based on finding exchangeable blocks in the Markovian data whereas (Tibshirani et al., 2019; Cauchois et al., 2020; Gibbs & Candes, 2021; Barber et al., 2022) utilize methods such as online updates to reflect the difference between empirically achieved confidence and target confidence, non-uniformly weighting calibration data, etc.

3. Quantifying the Uncertainty in Hidden Markov Models via Conformal Prediction

In this section, we first review the notions of mixtures of Markov chains and partial exchangeability. We then exploit a characterization of partial exchangeability in terms of mixtures of Markov chains presented in (Diaconis & Freedman, 1980) to extend the classical conformal prediction framework to Hidden Markov Models in a manner that provably preserves all its key theoretical guarantees.

3.1. Mixtures of Markov Chains and Partial Exchangeability

This subsection provides a brief review of the main result of (Diaconis & Freedman, 1980) which characterizes a mixture of Markov chains as a partially exchangeable process.

Formally, a process $X_1, X_2, \ldots$ is a *mixture of Markov chains*, if there exists a probability measure μ on the space of all $n \times n$ stochastic matrices $\mathcal{P}$ (for the state space $\mathcal{X} = \{1, 2, \ldots, n\}$) such that,

$$\mathbb{P}\{X_t = x_t \text{ for } 1 \leq t \leq T\} = \int_{\mathcal{P}} \prod_{t=1}^{T-1} P_{x_t x_{t+1}} \mu(dP) \tag{8}$$

for any sequence of states $x_1, x_2, \ldots, x_T \in \mathcal{X}$. (Di-

aconis & Freedman, 1980) characterized a mixture of Markov chains of the form (8) using a concept called partial exchangeability which is defined as follows. A distribution $\mathbb{P}$ is *partially exchangeable*, if for any pair of finite sequences $x_1, x_2, \ldots, x_T$ and $x_1', x_2', \ldots, x_T'$ that start at the same state (i.e., $x_1 = x_1'$) and undergo the same number of transitions from i to j for all $i, j \in \mathcal{X}$ (i.e., $\sum_{t=1}^{T-1} \mathbb{1}(x_t = i, x_{t+1} = j) = \sum_{t=1}^{T-1} \mathbb{1}(x_t' = i, x_{t+1}' = j)$), we have,

$$\mathbb{P}\left\{X_t = x_t \text{ for } 1 \leq t \leq T\right\} = \mathbb{P}\left\{X_t = x_t' \text{ for } 1 \leq t \leq T\right\}. \quad (9)$$

In other words, a distribution is partially exchangeable if it assigns the same probability to all finite sequences that start at the same state and undergo the same number of transitions from one state to another. For example,

$$\mathbb{P}\left\{X_1 = 1, X_1 = 1, X_2 = 5, X_3 = 1, X_4 = 1, X_5 = 7\right\}$$
$$= \mathbb{P}\left\{X_1 = 1, X_1 = 1, X_2 = 1, X_3 = 5, X_4 = 1, X_5 = 7\right\}$$

is a necessary condition for partial exchangeability since the two sequences 115117 and 111517 both start at the same state (i.e., 1) and undergo the same number of transitions from one state to another (i.e., twice from 1 to 1, once from 1 to 5, once from 5 to 1, and once from 1 to 7). The main result of (Diaconis & Freedman, 1980) stated below says that partial exchangeability of the distribution is a characterization of a mixture of Markov chains (for recurrent processes).

Theorem 3.1 (adapted from (Diaconis & Freedman, 1980)). *Suppose $X_1, X_2, \ldots$ is a recurrent process taking values in the finite state space $\mathcal{X} = \{1, 2, \ldots, n\}$ i.e.,*

$$\mathbb{P}\left\{X_t = i \text{ for infinitely many } t \mid X_1 = i\right\} = 1 \quad (10)$$

for all $i \in \mathcal{X}$. Then, $X_1, X_2, \ldots$ is a mixture of Markov chains in the sense of (8) if and only if it is partially exchangeable in the sense of (9).

Theorem 3.1 generalizes the well-known de Finetti's theorem which states that a sequence of random variables is exchangeable if and only if their joint distribution is a mixture of IID random variables. Thus, Theorem 3.1 is referred to as the *de Finetti's theorem for Markov chains*.

To see how Theorem 3.1 is applicable to the problem of quantifying the uncertainty of an unknown HMM (stated in Sec. 1), let us first consider a fully observed recurrent Markov chain (instead of an HMM) whose transition probability matrix is unknown and is assumed to be sampled from some prior distribution μ. Thus, an observed sequence is a mixture of Markov chains in the sense of (8) with respect to the prior distribution μ, and Theorem 3.1 implies that the observed sequence is partially exchangeable. Due to the partial exchangeability, we can permute the elements of

the observed sequence in a manner that preserves the initial state and the number of transitions between all pairs of states and then view the permuted sequences as randomizations of the observed sequence since they have the same joint probability. To construct such a group of permutations, let us define an $i-block$ as a finite string of states that begins with the state $i \in \mathcal{X}$ and contains no further $i's$. For example, the sequence of states 7521781663513421 can be partitioned into $1-$blocks as follows,

$$752 \quad \mathbf{178} \quad \mathbf{16635} \quad \mathbf{1342} \quad 1, \quad (11)$$

where the blocks indicated in bold red font are the $1-$blocks. Note that permuting the $i-$blocks changes neither the initial state nor the number of transitions from u to v for any $u, v \in \mathcal{X}$. Also, the recurrence condition (10) implies that the $i-$blocks are almost surely well-defined for any $i \in \mathcal{X}$. Thus, $i-$blocks are exchangeable and sequences obtained by permuting the $i-$blocks have the same probability as the observed sequence. Consequently, we can randomize the single observed sequence by permuting the exchangeable $i-$blocks while still preserving the temporal dependencies.

Therefore, although an observed sequence of a Markov chain with unknown parameters is not exchangeable in general, the set of $i-$blocks (for any fixed $i \in \mathcal{X}$) of that sequence is always exchangeable. As we will see next, this observation of the exchangeability of the $i-$blocks of a Markov chain can be exploited to extend the conformal prediction framework to the HMM setting.

3.2. Conformal Prediction for a Hidden Markov Model with Unknown Parameters

The following well-known lemma allows us to extend the previously outlined randomization method for a Markov chain based on Theorem 3.1 to the HMM setting (which is the context that we are dealing with in the main problem stated in Sec. 1).

Lemma 3.2. *If $X_1, X_2, \ldots$ is a Markov chain and $Y_1, Y_2, \ldots$ are its discrete memoryless observations, then the augmented process $(X_1, Y_1), (X_2, Y_2), \ldots$ is also a Markov chain.*

According to Lemma 3.2, the augmented process $(X_1, Y_1), (X_2, Y_2), \ldots$ is a Markov chain for any HMM. When the parameters of the HMM (i.e., transition probability matrix P and observation probability matrix B) are unknown, the augmented process $(X_1, Y_1), (X_2, Y_2), \ldots$ can be viewed as a mixture of Markov chains. Then, we define an $(i, j)-$block as a finite string of augmented states that begins with the state i and observation j and contains no more instances where $X_t = i, Y_t = j$. Since the augmented process is a mixture of Markov processes, it is partially exchangeable according to Theorem 3.1. Therefore, we can randomize the augmented process by permuting the

$(i, j)-$blocks while preserving the temporal dependencies by exploiting Theorem 3.1. This approach is formalized in Algorithm 1, and the theoretical guarantees that it offers and more details on practical implementation are discussed in the next two subsections.

Algorithm 1 Conformal Prediction for Hidden Markov Models

Input: Calibration data $\{(X_t, Y_t)\}_{t=1}^T$, Test observations $\{Y_t\}_{t=T+1}^{T+T_1}$, Miscoverage level $\alpha \in (0, 1)$

Output: A set of sequences $\mathcal{C}_{1-\alpha} \subseteq \mathcal{X}^{T_1}$

for $x = (x_{T+1}, \ldots, x_{T+T_1}) \in \mathcal{X}^{T_1}$ **do**

Step 1: Let $X_{T+1} = x_{T+1}, \ldots, X_{T+T_1} = x_{T+T_1}$

Step 2: Using $\{(X_t, Y_t)\}_{t=1}^{T+T_1}$, estimate the transition probability matrix P and the observation probability matrix B as:

$$\hat{P}_{ij} = \frac{\sum_{t=1}^{T+T_1-1} \mathbb{1}(X_t = i \wedge X_{t+1} = j)}{\sum_{t=1}^{T+T_1-1} \mathbb{1}(X_t = i)}$$
$$\hat{B}_{ij} = \frac{\sum_{t=1}^{T+T_1} \mathbb{1}(X_t = i \wedge Y_t = j)}{\sum_{t=1}^{T+T_1} \mathbb{1}(X_t = i)} \quad (12)$$

Step 3: Find all (i, j)-blocks of $\{(X_t, Y_t)\}_{t=1}^{T+T_1}$ where $X_{T+T_1} = i$ and $Y_{T+T_1} = j$. Let d be the number of (i, j)-blocks.

Step 4: Let Π be any permutation group of degree d.

for $\pi \in \Pi$ **do**

Let $\{(X_t^{(\pi)}, Y_t^{(\pi)})\}_{t=1}^{T+T_1}$ be the sequence where $(i, j)-$blocks obtained in Step 3 are permuted according to $\pi \in \Pi$ and calculate,

$$S(\pi) =$$
$$1 - \frac{\sum_{k=1}^{T_1} \mathbb{P}_{\hat{P}, \hat{B}}\left(X_{T+k}^{(\pi)} \middle| X_T^{(\pi)}; Y_{T+1}^{(\pi)}, \ldots, Y_{T+k}^{(\pi)}\right)}{T_1}$$

(13)

via an HMM filter recursion (see Appendix A).

end

Step 5: Calculate,

$$\hat{q}(x) = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} \mathbb{1}(S(\pi) \geq S(\mathbf{I})), \quad (14)$$

where $\mathbf{I}$ is the identity permutation.

end

return $\mathcal{C}_{1-\alpha} = \{x \in \mathcal{X}^{T_1} : \hat{q}(x) > \alpha\}$

For each possible candidate sequence $x = (x_{T+1}, \ldots, x_{T+T_1}) \in \mathcal{X}^{T_1}$, five steps are followed. In the first step, an augmented sequence $\{(X_t, Y_t)\}_{t=1}^{t=T+T_1}$ is generated by assuming $X_{T+1} = x_{T+1}, \ldots, X_{T+T_1} = x_{T+T_1}$. In the step 2, the augmented sequence $\{(X_t, Y_t)\}_{t=1}^{t=T+T_1}$ is used to estimate the transition probability matrix and the observation probability matrix. Step 3 partitions the augmented sequence into exchangeable (i, j)-blocks. In Step 4, the permutations of the (i, j)-blocks obtained by applying the set of permutations Π are viewed as randomizations of the augmented sequence $\{(X_t, Y_t)\}_{t=1}^{t=T+T_1}$ according to Theorem 3.1. For each permutation $\pi \in \Pi$, a conformity score $S(\pi)$ is calculated using the block-wise permuted sequence $\{(X_t^{(\pi)}, Y_t^{(\pi)})\}_{t=1}^{t=T+T_1}$. The conformity score $S(\pi)$ is based on the values $\mathbb{P}_{\hat{P}, \hat{B}}\left(X_{T+k}^{(\pi)} \middle| X_T^{(\pi)}; Y_{T+1}^{(\pi)}, \ldots, Y_{T+k}^{(\pi)}\right)$, $k = 1, \ldots, T_1$, where $\mathbb{P}_{\hat{P}, \hat{B}}\left(X_{T+k}^{(\pi)} \middle| X_T^{(\pi)}; Y_{T+1}^{(\pi)}, \ldots, Y_{T+k}^{(\pi)}\right)$ can be interpreted as the conditional probability that the underlying state of an HMM at time $T + k$ being $X_{T+k}^{(\pi)}$ given that the state at time T is $X_T^{(\pi)}$, observations at time instants $T + 1, \ldots, T + k$ are $Y_{T+1}^{(\pi)}, \ldots, Y_{T+k}^{(\pi)}$, transition probability matrix is $\hat{P}$ and observation probability matrix is $\hat{B}$. Hence, $\mathbb{P}_{\hat{P}, \hat{B}}\left(X_{T+k}^{(\pi)} \middle| X_T^{(\pi)}; Y_{T+1}^{(\pi)}, \ldots, Y_{T+k}^{(\pi)}\right)$, $k = 1, \ldots, T_1$ can be calculated recursively using the HMM filtering algorithm (given in Appendix A). For each candidate sequence $x = (x_{T+1}, \ldots, x_{T+T_1}) \in \mathcal{X}^{T_1}$, the fraction of permutations for which the conformity score exceeds the conformity score of the unpermuted sequence is denoted as $\hat{q}(x)$. Finally, the confidence set $\mathcal{C}_{1-\alpha}$ is constructed by including all candidate sequences $x = (x_{T+1}, \ldots, x_{T+T_1}) \in \mathcal{X}^{T_1}$ for which $\hat{q}(x)$ is larger than the miscoverage level α.

3.3. Theoretical Analysis of the Algorithm 1

The following result establishes the exact validity of the Algorithm 1.

Theorem 3.3 (exact validity of Algorithm 1). *Consider the Algorithm 1. If the joint distribution of $\{(X_t, Y_t)\}_{t=1}^{T+T_1}$ is a mixture of Markov Chains in the sense of (8), then the output $\mathcal{C}_{1-\alpha}$ satisfies*

$$1 - \alpha \leq \mathbb{P}\left\{\{X_t\}_{t=T+1}^{T+T_1} \in \mathcal{C}_{1-\alpha}\right\} \leq 1 - \alpha + \frac{1}{|\tilde{\Pi}|}, \quad (15)$$

where $\tilde{\Pi}$ is the permutation group (in Step 4) when the candidate sequence is the true sequence of hidden states.

The complete proof of Theorem 3.3 is given in Appendix B and its key idea is to invoke the exchangeability of $(i, j)-$blocks as outlined in Sec. 3.2. To intuitively understand the proof, let us consider the case where the candidate

sequence x in Algorithm 1 is the true hidden sequence. Then, $\{\hat{q}(x) > \alpha\}$ is the desired event where the true hidden sequence x is included in the constructed confidence set $\mathcal{C}_{1-\alpha}$. According to (14), the event $\{\hat{q}(x) > \alpha\}$ is the same as the event where the conformity score of the unpermuted augmented sequence $S(\mathbb{I})$ ranks at or below $\lceil(1-\alpha)|\tilde{\Pi}|\rceil$ among all conformity scores $S(\pi), \pi \in \tilde{\Pi}$ sorted in the ascending order (where $\tilde{\Pi}$ is the permutation group in Step 4 when the candidate sequence x is the true sequence of hidden states). Since the $(i,j)-$blocks created in Step 3 are exchangeable when x is the true hidden sequence, $S(\mathbb{I})$ is equally likely to take any rank among $1, 2, \ldots, |\tilde{\Pi}|$. Therefore, we have $\mathbb{P}(\{\hat{q}(x) > \alpha\}) = \frac{\lceil(1-\alpha)|\tilde{\Pi}|\rceil}{|\tilde{\Pi}|}$ and (15) follows by noting that $1 - \alpha \le \frac{\lceil(1-\alpha)|\tilde{\Pi}|\rceil}{|\tilde{\Pi}|} \le 1 - \alpha + \frac{1}{|\tilde{\Pi}|}$. The lower bound in (15) implies that the Algorithm 1 yields the confidence guarantee (2) that we wanted to achieve i.e., the set $\mathcal{C}_{1-\alpha}$ contains the true state sequence $\{X_t\}_{t=T+1}^{T+T_1}$ with a probability greater than $1 - \alpha$. The upper bound in (15) implies that for sequences containing large number of $(i,j)-$blocks, $\mathbb{P}\left\{\{X_t\}_{t=T+1}^{T+T_1} \in \mathcal{C}_{1-\alpha}\right\}$ is approximately equal to $1 - \alpha$. Hence, as the calibration sequence length T increases, the set $\mathcal{C}_{1-\alpha}$ will become smaller to contain only enough candidate sequences for achieving the $1 - \alpha$ coverage.

3.4. Practical Considerations and Alternative Implementations of Algorithm 1

Let us briefly discuss some practical aspects of implementing Algorithm 1 and additional settings where it could be utilized.

Computational complexity of Algorithm 1: Computing each of the T_1 summands in (13) in Step 4 using an HMM filter (given in Appendix A) needs $O(|\mathcal{X}|^2)$ computations. Thus, for each candidate sequence, Algorithm 1 requires $O(|\Pi| T_1 |\mathcal{X}|^2)$ computations, where $|\Pi|$ is the number of permutations of $(i,j)-$blocks (i.e., the cardinality of the permutation group), T_1 is the length of the sequence to be predicted and $|\mathcal{X}|$ is the cardinality of the state space.

Alternative choices for the group of permutations Π: In general, the group of permutations Π could even be the set of all permutations of the (i,j)-blocks. However, letting Π be the set of all possible permutations is computationally expensive and might be unnecessary in most practical cases since the conformity score that we use in Step 4 is dependent only on the last $T_1 + 1$ elements of the permuted sequence. Consequently, most permutations may result in the same conformity score. As such, considering a subgroup of permutations which includes the permutations corresponding to all possible variations of the last $T_1 + 1$ elements is computationally more efficient. Such an approach is used in our numerical experiments in Sec. 4.

Alternative choices for the conformity score function: The conformity score $S(\cdot)$ for each permuted sequence calculated in Step 4 is based on $\mathbb{P}_{\hat{P},\hat{B}}\left(X_{T+k}^{(\pi)} \mid X_T^{(\pi)}; Y_{T+1}^{(\pi)}, \ldots, Y_{T+k}^{(\pi)}\right)$, $k = 1, \ldots, T_1$ that can be recursively calculated using an HMM filter. Alternatively, one could use $\mathbb{P}_{\hat{P},\hat{B}}\left(X_{T+k}^{(\pi)} \mid X_T^{(\pi)}; Y_{T+1}^{(\pi)}, \ldots, Y_{T+T_1}^{(\pi)}\right)$, $k = 1, \ldots, T_1$, which are the smoothed probabilities that can be recursively calculated using an HMM smoother. In particular, HMM smoothing calculates the likelihood of the state $X_{T+k}^{(\pi)}$ conditional on all observations from $t = T + 1$ to $t = T + T_1$ whereas HMM filtering conditions only on the past observations from $t = T + 1$ to $t = T + k$. Compared to HMM filtering, HMM smoothing is computationally more expensive due to the use of forward-backward algorithm and it is also non-causal due to the conditioning on both past and future observations.[4] However, HMM smoothing may produce smaller confidence sets in cases where the observations are more informative and the number of available observations T_1 is relatively large. Going even further, the smoothed joint likelihoods of the form $\mathbb{P}_{\hat{P},\hat{B}}\left(X_{T+k}^{(\pi)}, \ldots, X_{T+k+m}^{(\pi)} \mid X_T^{(\pi)}; Y_{T+1}^{(\pi)}, \ldots, Y_{T+T_1}^{(\pi)}\right)$, $k = 1, \ldots, T_1 - m$ (where $m < T_1$) can also be used. Conformity scores based on such joint likelihoods will better capture the sequential correlations among the hidden states, and thus lead to smaller confidence sets when the states are highly Markovian.

Quantifying the uncertainty of a pre-trained predictor: If we are given a pre-trained predictor (e.g., previously obtained estimates of the transition probability matrix P and observation probability matrix B from a different dataset, or a neural network based predictor), Step 2 can be omitted and the likelihoods in Step 4 can be calculated using the given pre-trained predictor. In this setting, Algorithm 1 can be used to calibrate the given predictor to a desired confidence level $1 - \alpha$ and then quantify its uncertainty via the cardinality of confidence sets. For example, after calibrating to a miscoverage level $\alpha = 0.2$ (i.e., confidence level 0.8), if a given pre-trained black-box predictor generates confidence sets that are relatively large (e.g., $\mathbb{E}\{|\mathcal{C}_{1-\alpha}|\} \approx |\mathcal{X}^{T_1}|$), then we can infer that the given predictor is not suitable for high stakes prediction tasks. On the other hand, if the given pre-trained predictor generates confidence sets that are relatively small (e.g., $\mathbb{E}\{|\mathcal{C}_{1-\alpha}|\} \ll |\mathcal{X}^{T_1}|$), then it is suitable for high-stakes prediction tasks. Hence, Algorithm 1 can be adapted to manage the risk of utilizing black-box pre-trained predictors in high-stakes prediction tasks.

---

[4]See (Krishnamurthy, 2016) for a detailed exposition of HMM filtering and smoothing.

4. Numerical Experiments and Empirical Results

This section provides numerical experiments and empirical results that verify and complement the theoretical results in order to highlight the practical applicability of the Algorithm 1.

4.1. Numerical Experiments

Simulation setup: We use an HMM with the state space $\mathcal{X} = \{0, 1\}$, observation space $\mathcal{Y} = \{0, 1\}$, initialized at time $t = 1$ with X_1 chosen uniformly from $\mathcal{X}$. The parameterized transition probability matrix P and observation probability matrix B of the HMM are,

$$P = \begin{bmatrix} p & 1-p \\ 1-p & p \end{bmatrix} \quad B = \begin{bmatrix} b & 1-b \\ 1-b & b \end{bmatrix}, \quad (16)$$

where $p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and $b \in \{0.5, 0.75, 0.9\}$. These parameter configurations cover several important scenarios. In particular, setting petting $p = 0.5$ leads to IID states whereas $p = 0.1, 0.9$ leads to a strongly Markov process (with states $X_1, X_2, \ldots$ being heavily correlated). Setting $b = 0.5$ leads to meaningless observations (i.e., the observation Y_t and the underlying state X_t are statistically independent at any time instant t) and $b = 0.9$ leads to relatively more accurate observations (i.e., 0.9 probability for the event $Y_t = X_t$). We consider three calibration sequence lengths $T \in \{50, 100, 200\}$ and three lengths of the sequence to be predicted $T_1 \in \{1, 2, 3\}$. The miscoverage level is set to $\alpha = 0.2$ (i.e., desired confidence level is 80%). The permutation group Π (in Step 4 of Algorithm 1) is chosen so that for any d number of $(i, j)-$blocks, each $\pi \in \Pi$ corresponds to one of the $d!/(d - T_1 - 1)!$ unique ways to arrange the last $T_1 + 1$ $(i, j)-$blocks.[5] Algorithm 1 is then used to generate the 80% confidence set $\mathcal{C}_{0.8}$ for $(X_{T+1}, \ldots, X_{T+T_1}) \subset \mathcal{X}^{T_1}$. The probability of the confidence set containing the true sequence $\mathbb{P}\{(X_{T+1}, \ldots, X_{T+T_1}) \in \mathcal{C}_{0.8}\}$ (i.e., the empirical coverage) and the expected cardinality of the confidence set $\mathbb{E}\{|\mathcal{C}_{0.8}|\}$ are empirically estimated using a Monte-Carlo average over 500 independent iterations of this process. The empirical coverage and the empirically estimated expected cardinality of the prediction sets (scaled by the largest possible cardinality $|\mathcal{X}^{T_1}|$) obtained using this simulation setup are shown in Fig. 1. Additional numerical results obtained using a simulation setup with 3 states and 3 observations
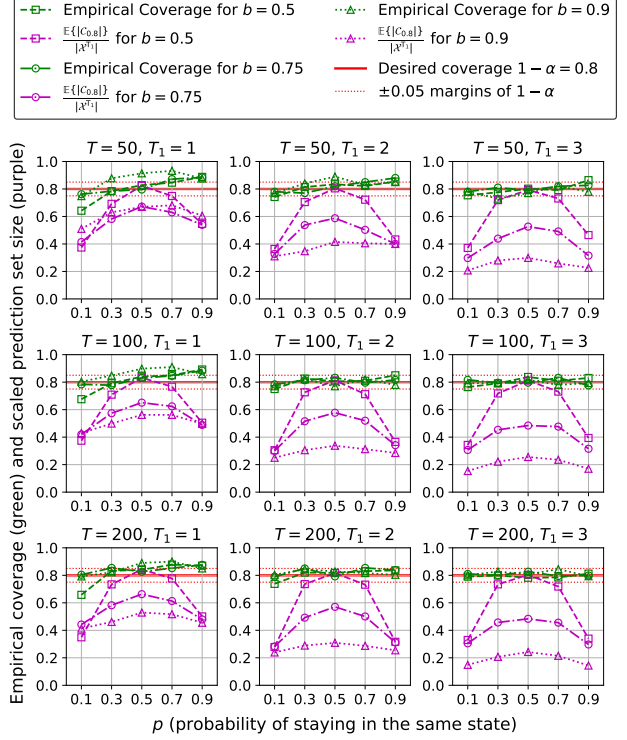

Figure 1. The coverage $\mathbb{P}\{(X_{T+1}, \ldots, X_{T+T_1}) \in \mathcal{C}_{0.8}\}$ (green lines) and the scaled expected cardinality of prediction sets $\frac{\mathbb{E}\{|\mathcal{C}_{1-\alpha}|\}}{|\mathcal{X}^{T_1}|}$ (purple lines) of Algorithm 1 estimated via the simulation setup discussed in Sec. 4.1. The numerical results show that the Algorithm 1 yields the coverage guarantee promised by Theorem 3.3 under both exchangeable (i.e., $p = 0.5$) and Markovian (i.e., $p \neq 0.5$) settings. In particular, the proposed Algorithm 1 is able to exploit the low aleatoric uncertainty in the strongly Markovian regime (i.e., $p \gg 0.5$ and $p \ll 0.5$) to make the confidence sets smaller whereas the classical conformal prediction is not applicable to strongly Markovian processes due to the lack of exchangeability. Therefore, Algorithm 1 successfully exploits the *de Finetti's Theorem for Markov Chains* (Theorem 3.1) to overcome the barrier of lack of exchangeability and helps unleash the powerful potential of the conformal prediction framework in the Markovian setting.

are given in Appendix C.

Insights from the numerical results: Fig. 1 yields insights that complement the Theorem 3.3 (which established the exact validity of the Algorithm 1) as we discuss below.

Empirical validity of the Algorithm 1: Fig. 1 shows that the empirical coverage (indicated by green lines) is approximately equal to (above 0.78) or exceeds the desired coverage ($1 - \alpha = 0.8$) in almost all considered parameter configurations. In particular, the empirical coverage remains above 0.78 when the states are strongly Markovian with low aleatoric uncertainty (i.e., $p \gg 0.5$ and $p \ll 0.5$) as well as

---

[5]The rationale for choosing $T_1 + 1$ here is as follows. This choice of Π would include all possible permutations of the $(i, j)-$blocks that constitute unique permutations of the last $T_1 + 1$ elements. Since our choice of the conformity score $S(\cdot)$ given in (13) depends only on the last $T_1 + 1$ elements of the permuted sequence, the permutation group Π constructed in this manner would yield all the unique conformity score values that can be obtained by permuting the $(i, j)-$blocks.

Table 1. Empirical results obtained by evaluating Algorithm 1 on four real-world datasets

Dataset Description	Iterations (time period used for testing)	Number of States $|\mathcal{X}|$	Calibration Sequence Length T	Empirical Coverage (for $1 - \alpha = 0.8$)	Scaled Prediction Set Size
Electricity Consumption in Delhi	170 (6 months)	5	300	0.81	0.25
Electricity Consumption in Punjab	170 (6 months)	5	300	0.8	0.28
Household Energy Consumption	53 (2 months)	7	300	0.78	0.22
SP500 Index	20 (1 month)	2	100	0.8	0.65

when the states are IID (i.e., $p = 0.5$). In contrast, the classical conformal prediction guarantees the desired coverage only in the IID setting (i.e., $p = 0.5$), and it is not applicable to the Markovian regime (i.e., $p \neq 0.5$) due to the lack of exchangeability. The only parameter configuration where the empirical coverage drops below the desired coverage (to approximately 0.7) is when $T_1 = 1, p = 0.1, b = 0.5$. The reason behind this observation could be that the available set of candidate sequences to choose from is only 2 when $T_1 = 1$, and this low resolution in the candidate sequences amplifies effects of the meaningless observations ($b = 0.5$) and the variance stemming from estimating a transition probability matrix with $p = 0.1$. In all other combinations of parameters p, b, T_1 and T, the empirical coverage is approximately equal or larger than the desired coverage. Hence, the numerical experiments confirm that the coverage guarantee given in Theorem 3.3 is in fact achieved in practice in both exchangeable and Markovian settings. Fig. 2 (in Appendix C) obtained using a larger state space further supports these conclusions.

Cardinality of the prediction sets: Fig. 1 shows that the scaled expected prediction set size $\frac{\mathbb{E}\{|\mathcal{C}_{1-\alpha}|\}}{|\mathcal{X}^{T_1}|}$ (indicated by purple lines) is smaller compared to the desired coverage $1 - \alpha$ in each considered case (except when $b = p = 0.5$ where the prediction set size needs to be 0.8 as there is no information in the state transitions or observations). Stated differently, Algorithm 1 yields smaller confidence sets compared to randomly choosing $1 - \alpha$ fraction of all possible candidate sequences $\mathcal{X}^{T_1}$. In particular, Fig. 1 shows that even with meaningless observations (i.e., $b = 0.5$), Algorithm 1 achieves a 0.8 confidence with a scaled prediction set size smaller than 0.8 by exploiting the information in the state transitions (when $p \neq 0.5$). The average prediction set size further decreases as the accuracy of observations and the length of the calibration sequence T increase. Fig. 2 (in Appendix C) also shows how the confidence sets remain relatively smaller even with a larger state space.

Performance difference between exchangeable and HMM settings: Fig. 1 shows that the Algorithm 1 achieves the desired coverage in both IID ($p = 0.5$) and HMM settings ($p \neq 0.5$). However, performance of the Algorithm 1 is better in the HMM setting in terms of the prediction set size i.e., $\frac{\mathbb{E}\{|\mathcal{C}_{1-\alpha}|\}}{|\mathcal{X}^{T_1}|}$ is smaller for the HMM setting compared to the IID setting. This is due to the fact that the Markovian structure of the states is exploited by the HMM filter to make the confidence sets smaller. Therefore, the proposed method views the low aleatoric uncertainty in Markovian regime as an advantage that can be exploited to make the confidence sets smaller whereas it is a violation of a crucial assumption for classical conformal prediction. For example, when $b = 0.5$ (i.e., meaningless observations) the only way for any algorithm to achieve the desired confidence $1 - \alpha = 0.8$ in the IID setting is to randomly choose 0.8 fraction of all possible candidate sequences. In the HMM setting with $p = 0.9$, the same confidence is achieved with approximately 0.4 fraction of all possible candidate sequences (e.g., in the case $T = 200, T_1 = 3$). Remarkably, Algorithm 1 is not explicitly aware whether the states are Markovian or the measurements are accurate in order to ensure the desired coverage. That information is only available via the $T-$length calibration sequence given as an input.

To summarize, the numerical results verify and complement the Theorem 3.3 which established the validity of the Algorithm 1. In particular, numerical results show that empirical coverage is approximately equal to the desired coverage (with the difference attributed to the randomness in the Monte Carlo averaging) in both HMM and exchangeable settings. Further, the prediction sets are relatively smaller in the HMM setting compared to exchangeable setting. Thus, by exploiting the de Finetti's Theorem for Markov Chains, the proposed Algorithm 1 overcomes the barrier of lack of exchangeability in Markovian settings and views it as an advantage instead of a violation of an assumption.

4.2. Empirical Results Using Real-World Datasets

To illustrate the practical applicability of Algorithm 1 in real-world settings, this section presents empirical results obtained by applying Algorithm 1 to the four publicly available real world datasets listed in Table 1: daily electricity consumption in two Indian states (Delhi, Punjab) (Khanna & Mundhra, 2020), daily energy consumption of an apartment in San Jose (Gopinadhan, 2020) and daily values of the SP500 stock market index (Yahoo Finance, 2023).

Experimental setup: For the first three datasets in Table 1, the continuous valued energy consumption data is converted to discrete states based on how many standard deviations away each value is from the average value (over all time steps). This yielded state space cardinalities 5, 5 and 7 for the first three datasets. For the last dataset (SP500 index), the binary state indicates whether the index finished with a gain or a loss for each day. Then, Algorithm 1 was used to obtain one step ahead prediction sets (i.e., $T_1 = 1$) by viewing the current state as the measurement for the next state. The empirical coverage and average prediction set size for each dataset are shown in Table 1. A time-series visualization of the true state with the predicted set of states and more details on the datasets are given in Fig. 3 in Appendix D.

Insights from the empirical results: Table 1 shows that the empirical coverage for each dataset is approximately equal to the desired coverage. For the first three datasets (energy related applications), the scaled prediction set size indicates that the 0.8 confidence is achieved with a smaller (0.22 to 0.28) scaled confidence set size. In other words, a high coverage can be achieved through a smaller confidence set due to the strongly Markovian nature of energy consumption. For the last dataset (SP500 index), achieving the desired confidence of 0.8 needs a scaled prediction set size of 0.65, indicating that one cannot confidently predict the state of SP500 assuming only a simple Markov model. Thus, the empirical results verify the validity of Theorem 3.3 in real-world settings and illustrate how Algorithm 1 is useful for uncertainty quantification in an HMM framework.

5. Discussion and Conclusion

This paper presented a generalized conformal prediction algorithm for the Hidden Markov Model framework. In particular, given a sequence of states and their corresponding noisy observations (for time $t = 1, \ldots, T$) from a Hidden Markov Model (HMM), our aim is to generate a $1 - \alpha$ confidence set for the hidden state sequence corresponding to new observations (from time $t = T + 1, \ldots, T + T_1$). When the underlying process is exchangeable (e.g., IID), *conformal prediction algorithm* (proposed in (Vovk et al., 2005)) can be used to generate the $1 - \alpha$ confidence set. However, as Markov processes in general are not exchangeable, the classical conformal prediction cannot be applied in Markovian frameworks. As a solution, the algorithm proposed in this paper exploits the *de Finetti's Theorem for Markov chains* (presented in (Diaconis & Freedman, 1980)) to partition the data from the HMM into blocks that are guaranteed to be exchangeable. The permutations of the constructed blocks are viewed as randomizations of the observed sequence from the HMM. The resulting algorithm provably yields the desired $1 - \alpha$ coverage guarantee in both the exchangeable setting and the HMM setting. In particular,

the proposed algorithm utilizes the lack of exchangeability in Markov processes to make the confidence sets smaller compared to the exchangeable setting. Further, only one (finite length) sample path from the HMM is needed as calibration (i.e., training) data and no prior knowledge of the parameters of the underlying HMM (i.e., transition and observation probabilities) is assumed. As such, the generalized algorithm proposed in this paper helps unleash the powerful potential of the conformal prediction framework in Markovian settings without any trade off in the confidence guarantees.

References

Angelopoulos, A. N., Bates, S., et al. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. Conformal prediction beyond exchangeability. *arXiv preprint arXiv:2202.13415*, 2022.

Batu, T., Guha, S., and Kannan, S. Inferring mixtures of Markov chains. In *Conference on Learning Theory*, pp. 186–199. Springer, 2004.

Cauchois, M., Gupta, S., Ali, A., and Duchi, J. C. Robust validation: Confident predictions even when distributions shift. *arXiv preprint arXiv:2008.04267*, 2020.

Chernozhukov, V., Wüthrich, K., and Yinchu, Z. Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Conference On Learning Theory*, pp. 732–749. PMLR, 2018.

Cherubin, G. and Nouretdinov, I. Hidden Markov models with confidence. In *Symposium on Conformal and Probabilistic Prediction with Applications*, pp. 128–144. Springer, 2016.

Diaconis, P. and Freedman, D. de Finetti's theorem for Markov chains. *The Annals of Probability*, pp. 115–130, 1980.

Fontana, M., Zeni, G., and Vantini, S. Conformal prediction: a unified review of theory and new challenges. *Bernoulli*, 29(1):1–23, 2023.

Forkan, A. R. M. and Khalil, I. Peace-home: Probabilistic estimation of abnormal clinical events using vital sign correlations for reliable home-based monitoring. *Pervasive and Mobile Computing*, 38:296–311, 2017.

Gibbs, I. and Candes, E. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34, 2021.

Gopinadhan, J. House hold energy data, 2020. URL https://www.kaggle.com/datasets/jaganadhg/house-hold-energy-data.

Gupta, R., Kumar, R., and Vassilvitskii, S. On mixtures of markov chains. *Advances in neural information processing systems*, 29, 2016.

Hassan, M. R. and Nath, B. Stock market forecasting using hidden Markov model: a new approach. In *5th International Conference on Intelligent Systems Design and Applications (ISDA'05)*, pp. 192–196. IEEE, 2005.

Khanna, T. and Mundhra, N. Power consumption in india (2019-2020), 2020. URL https://www.kaggle.com/datasets/twinkle0705/state-wise-power-consumption-in-india.

Krishnamurthy, V. *Partially observed Markov decision processes*. Cambridge university press, 2016.

Petković, T., Puljiz, D., Marković, I., and Hein, B. Human intention estimation based on hidden Markov model motion validation for safe flexible robotized warehouses. *Robotics and Computer-Integrated Manufacturing*, 57: 182–196, 2019.

Rashid, K. M. and Behzadan, A. H. Risk behavior-based trajectory prediction for construction site safety monitoring. *Journal of construction engineering and management*, 144(2):04017106, 2018.

Shafer, G. and Vovk, V. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Stankeviciute, K., M Alaa, A., and van der Schaar, M. Conformal time-series forecasting. *Advances in Neural Information Processing Systems*, 34, 2021.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Uddin, M. Z. A wearable sensor-based activity prediction system to facilitate edge computing in smart healthcare system. *Journal of Parallel and Distributed Computing*, 123:46–53, 2019.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Xu, C. and Xie, Y. Conformal prediction interval for dynamic time-series. In *International Conference on Machine Learning*, pp. 11559–11569. PMLR, 2021.

Yahoo Finance. SP500 Daily Data. https://finance.yahoo.com/quote/%5EGSPC?p=%5EGSPC, 2023. [Online; accessed 19-Jan-2023].

Yuan, W., Li, Z., and Wang, C. Lane-change prediction method for adaptive cruise control system with hidden Markov model. *Advances in Mechanical Engineering*, 10 (9):1687814018802932, 2018.

A. HMM Filtering Algorithm

The HMM filter subroutine used in Step 4 of the Algorithm 1 (for each permutation $\pi \in \Pi$) is as follows.

Algorithm 2 HMM Filter Subroutine in Algorithm 1

Input: Estimated transition probability matrix $\hat{P}$, estimated observation probability matrix $\hat{B}$, the state $X_T^{(\pi)}$ at time T and the last T_1 elements of the permuted sequence $\{(X_t^{(\pi)}, Y_t^{(\pi)})\}_{t=1}^{T+T_1}$

for $k = 1, \ldots, T_1$ **do**

    Calculate $p_{T+k} = \dfrac{\hat{B}_{Y_{T+k}} \hat{P}' p_{T+k-1}}{\mathbb{1}^T \hat{B}_{Y_{T+k}} \hat{P}' p_{T+k-1}}$, where $\hat{B}_i$ denotes a diagonal matrix with column-i of $\hat{B}$ as its diagonal, p_{T+k}

    is an $n-$dimensional probability vector with $p_{T+k}(i) = \mathbb{P}_{\hat{P}, \hat{B}}\left(X_{T+k}^{(\pi)} = i \,\middle|\, X_T^{(\pi)}; Y_{T+1}^{(\pi)}, \ldots, Y_{T+k}^{(\pi)} \right)$ and $\hat{P}'$ denotes

    the transpose of $\hat{P}$.

end

return $\mathbb{P}_{\hat{P}, \hat{B}}\left(X_{T+k}^{(\pi)} \,\middle|\, X_T^{(\pi)}; Y_{T+1}^{(\pi)}, \ldots, Y_{T+k}^{(\pi)} \right), \quad k = 1, \ldots, T_1$

Note that the inputs to the HMM filter are dependent only on the last $T_1 + 1$ elements of the permuted sequence. Hence, it suffices to use a permutation group Π containing permutations of the $(i, j)-$blocks corresponding to all possible ways to arrange the last $T_1 + 1$ elements.

B. Proof of Theorem 3.3 (Exact validity of Algorithm 1)

Note that the true sequence $\{X_t\}_{t=T+1}^{T+T_1}$ will not be included in the confidence set $\mathcal{C}_{1-\alpha}$ if and only if $\hat{q}(X_{T+1}, \ldots, X_{T+T_1}) \leq \alpha$ i.e.,

$$\left\{ \{X_t\}_{t=T+1}^{T+T_1} \notin \mathcal{C}_{1-\alpha} \right\} = \{\hat{q}(X_{T+1}, \ldots, X_{T+T_1}) \leq \alpha\} \tag{17}$$

We therefore focus on the event $\{\hat{q}(X_{T+1}, \ldots, X_{T+T_1}) \leq \alpha\}$. Next, let

$$S^{(1)}\left(\{(X_t, Y_t)\}_{t=1}^{T+T_1}\right) < S^{(2)}\left(\{(X_t, Y_t)\}_{t=1}^{T+T_1}\right) < \cdots < S^{(|\Pi|)}\left(\{(X_t, Y_t)\}_{t=1}^{T+T_1}\right)$$

denote the conformity scores $S(\pi) = S\left(\{(X_t^{(\pi)}, Y_t^{(\pi)})\}_{t=1}^{T+T_1}\right), \pi \in \Pi$ calculated in the Step 4 of Algorithm 1 sorted in the ascending order i.e., they are the sorted conformity scores of all sequences obtained by applying the transformations $\pi \in \Pi$ to the original sequence,

$$\{(X_t, Y_t)\}_{t=1}^{T+T_1} = \{(X_t^{(\mathbb{I})}, Y_t^{(\mathbb{I})})\}_{t=1}^{T+T_1},$$

where $\mathbb{I} \in \Pi$ is the identity transformation. Note that Π is an algebraic group, implying that,

$$S^{(j)}\left(\{(X_t, Y_t)\}_{t=1}^{T+T_1}\right) = S^{(j)}\left(\{(X_t^{(\pi)}, Y_t^{(\pi)})\}_{t=1}^{T+T_1}\right), \forall \pi \in \Pi, j = 1, 2, \ldots, |\Pi|, \tag{18}$$

i.e., the sorted conformity scores will be the same irrespective of which transformation is considered the original one.

Then, from (14) we get,

$$\{\hat{q}(X_{T+1}, \ldots, X_{T+T_1}) \leq \alpha\} = \left\{ S\left(\{(X_t, Y_t)\}_{t=1}^{T+T_1}\right) > S^{(\lceil |\Pi|(1-\alpha) \rceil)}\left(\{(X_t, Y_t)\}_{t=1}^{T+T_1}\right) \right\}.$$

Therefore, we have,

$$\mathbb{P}\{\hat{q}\left(X_{T+1},\ldots,X_{T+T_1}\right) \le \alpha\} = \mathbb{E}\{\mathbb{1}\{\hat{q}\left(X_{T+1},\ldots,X_{T+T_1}\right) \le \alpha\}\}$$

$$= \mathbb{E}\left\{\mathbb{1}\left\{S\left(\{(X_t,Y_t)\}_{t=1}^{T+T_1}\right) > S^{(\lceil|\Pi|(1-\alpha)\rceil)}\left(\{(X_t,Y_t)\}_{t=1}^{T+T_1}\right)\right\}\right\}$$

$$= \mathbb{E}\left\{\frac{1}{|\Pi|}\sum_{\pi\in\Pi}\mathbb{1}\left(S\left(\{(X_t^{(\pi)},Y_t^{(\pi)})\}_{t=1}^{T+T_1}\right) > S^{(\lceil|\Pi|(1-\alpha)\rceil)}\left(\{(X_t^{(\pi)},Y_t^{(\pi)})\}_{t=1}^{T+T_1}\right)\right)\right\}$$

(Since $\{(X_t,Y_t)\}_{t=1}^{T+T_1}$ is a mixture of Markov chains, it is partially exchangeable according to Theorem 3.1.

The partial exchangeability of $\{(X_t,Y_t)\}_{t=1}^{T+T_1}$ implies that $\{(X_t^{(\pi)},Y_t^{(\pi)})\}_{t=1}^{T+T_1} \stackrel{d}{=} \{(X_t,Y_t)\}_{t=1}^{T+T_1}, \forall\pi\in\Pi.$)

$$= \mathbb{E}\left\{\frac{1}{|\Pi|}\sum_{\pi\in\Pi}\mathbb{1}\left(S\left(\{(X_t^{(\pi)},Y_t^{(\pi)})\}_{t=1}^{T+T_1}\right) > S^{(\lceil|\Pi|(1-\alpha)\rceil)}\left(\{(X_t,Y_t)\}_{t=1}^{T+T_1}\right)\right)\right\} \text{ (from (18))}$$

$$= \frac{|\Pi| - \lceil|\Pi|(1-\alpha)\rceil}{|\Pi|} \le \alpha$$

$$\implies \mathbb{P}\left\{\{X_t\}_{t=T+1}^{T+T_1}\in\mathcal{C}_{1-\alpha}\right\} = \mathbb{P}\{\hat{q}\left(X_{T+1},\ldots,X_{T+T_1}\right) > \alpha\} > 1-\alpha$$

which yields the lower bound in Theorem 3.3. The upper bound follows by noting that $\alpha - \frac{1}{|\Pi|} \le \frac{|\Pi|-\lceil|\Pi|(1-\alpha)\rceil}{|\Pi|} \le \alpha$. This completes the proof.

C. Additional Numerical Results

Simulation setup for the case where the number of states $|\mathcal{X}| = 3$: The second setup uses the transition probability matrix P and the parameterized observation probability matrix B,

$$P_2 = \begin{bmatrix} 0.1 & 0.6 & 0.3 \\ 0.3 & 0.1 & 0.6 \\ 0.6 & 0.3 & 0.1 \end{bmatrix} \qquad B_2 = \begin{bmatrix} b & \frac{1-b}{2} & \frac{1-b}{2} \\ \frac{1-b}{2} & b & \frac{1-b}{2} \\ \frac{1-b}{2} & \frac{1-b}{2} & b \end{bmatrix}, \tag{19}$$

where $b \in \left\{\frac{1}{3}, 0.6, 0.9\right\}$. Analogous to simulation setup in Sec. 4.1, setting $b = \frac{1}{3}$ leads to meaningless observations and $b = 0.9$ leads to more accurate measurements. We consider $T = 60, 90, 120, 150, 180$ for the length of the calibration sequence and the length of the sequence to be predicted is set to $T_1 = 3$. The other steps are the same as outlined in the simulation setup in Sec. 4.1. The same experiment is then performed for the IID case (i.e., each element in the transition probability matrix P_2 is set to $\frac{1}{3}$). The empirical coverage and the empirically estimated expected cardinality of the prediction sets (scaled by the largest possible cardinality $|\mathcal{X}^{T_1}|$) obtained using this simulation setup are shown in Fig. 2.

Insights from the numerical results for the case where the number of states $|\mathcal{X}| = 3$: Fig. 2 yields insights that complement and support the Theorem 3.1 (which established the validity of the Algorithm 1) and the numerical results provided in Sec. 4.1 as discussed below.

Empirical validity of the Algorithm 1: Fig. 2 shows that the empirical coverage (indicated by green lines) is very close (within ± 0.03 range) to the desired coverage ($1-\alpha = 0.8$) in both the HMM setting and the exchangeable (IID) setting for all considered parameter values. Hence, the empirical results confirm that the validity of the Algorithm 1 established in Theorem 3.3 is in fact achieved in practice in both exchangeable and Markovian settings.

Cardinality of the prediction sets: Fig. 2(a) shows that the scaled expected prediction set size $\frac{\mathbb{E}\{|\mathcal{C}_{1-\alpha}|\}}{|\mathcal{X}^{T_1}|}$ (indicated by purple lines) is smaller compared to the desired coverage $1-\alpha$ in each considered case. Stated differently, Algorithm 1 yields smaller confidence sets compared to randomly choosing $1-\alpha$ fraction of all possible candidate sequences $\mathcal{X}^{T_1}$. In particular, Fig. 2(a) shows that even with meaningless observations (i.e. $b = 1/3$), Algorithm 1 achieves a 0.8 confidence with approximately 0.6 fraction of all candidate sequences. The average prediction set size further decreases as the accuracy of the observations and the length of the calibration sequence T increase. For example, when $b = 0.9$ and $T = 180$, Algorithm 1 achieves the 0.8 confidence with approximately only 0.15 fraction of all possible candidate sequences.
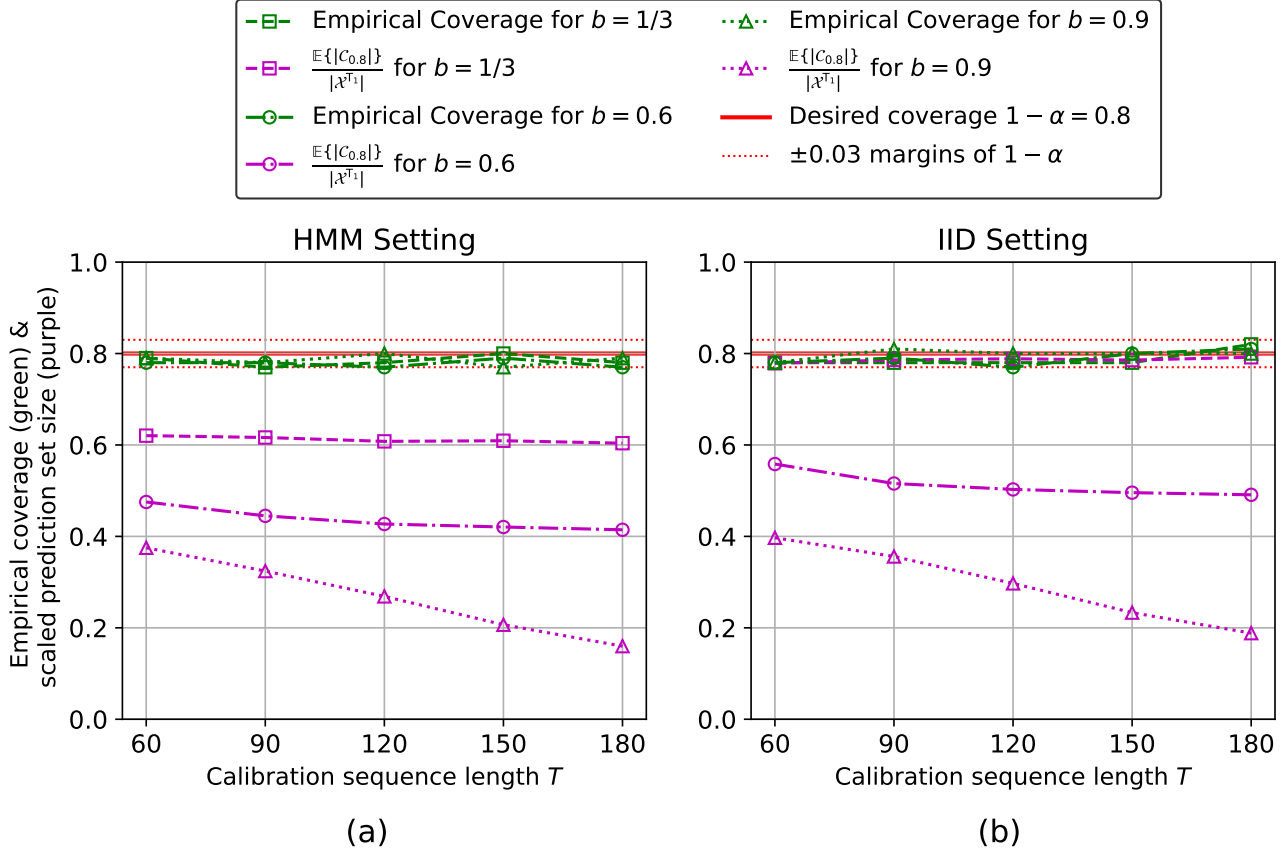
Figure 2. The empirically estimated coverage $\mathbb{P}\{(X_{T+1}, \ldots, X_{T+T_1}) \in \mathcal{C}_{0.8}\}$ (green lines) and the scaled expected cardinality of the prediction sets $\frac{\mathbb{E}\{|\mathcal{C}_{1-\alpha}|\}}{|\mathcal{X}^{T_1}|}$ (purple lines) for two cases: Fig. 2(a) for transition probability matrix P given in (19) and Fig. 2(b) for the IID case (i.e., each element of the transition probability matrix P is equal to $\frac{1}{3}$). The length of the sequence to be predicted is $T_1 = 3$ and different markers indicate different values of the parameter b of the observation probability matrix B in (19) where larger values of b correspond to more accurate observations. The desired coverage is $1 - \alpha = 0.8$. The coverage and the expected cardinality of the prediction set for each parameter configuration were estimated using a Monte-Carlo average over 500 iterations as discussed in Sec. 4. Fig. 2(a) indicates that the empirical coverage (green line) is within ± 0.03 margins of the desired coverage ($1 - \alpha = 0.8$) for each considered parameter configuration in the HMM setting. Hence the numerical results agree with the validity of the Algorithm 1 established in Theorem 3.3 for HMM models. Fig. 2(b) shows that the proposed algorithm yields the desired coverage in the exchangeable (IID) setting as well. Thus, the numerical results confirm that the proposed algorithm produces valid confidence sets in both HMM and exchangeable settings.

Performance difference between exchangeable and HMM settings: As discussed earlier, Algorithm 1 achieves the desired coverage in both IID (exchangeable) and HMM settings. However, performance of the Algorithm 1 is better in the HMM setting in terms of the prediction set size i.e., $\frac{\mathbb{E}\{|\mathcal{C}_{1-\alpha}|\}}{|\mathcal{X}^{T_1}|}$ is smaller for the HMM setting compared to the IID setting. This is due to the fact that the Markovian structure of the states is exploited by the HMM filter to make the confidence sets smaller. For example, when $b = 1/3$ (i.e., meaningless observations) the only way for any algorithm to achieve the desired confidence $1 - \alpha = 0.8$ in the IID setting is to randomly choose 0.8 fraction of all possible candidate sequences. In the HMM setting, the same confidence is achieved with approximately 0.6 fraction of all possible candidate sequences. Further, Algorithm 1 is not explicitly aware of whether the states are Markovian or the measurements are accurate. This information is only available via the $T-$length calibration sequence given as an input.

To summarize, the supplementary numerical results (Fig. 2) verify that the coverage guarantee (for Algorithm 1) provided by Theorem 3.3 is in fact valid in practice. In addition, Fig. 2 also illustrates that the insights obtained using the numerical experiments for the binary state space (Fig. 1) extend to larger state spaces.

D. Supplementary Details on the Empirical Results

Details on the four datasets used to obtain empirical results (summarized in Table 1)

Electricity Consumption in Delhi and Punjab: These two datasets were obtained from the publicly available repository (Khanna & Mundhra, 2020). It contains a time series for a period of 17 months beginning from 2-Jan-2019 till 23-May-2020 and has been scraped from the weekly energy reports of Power System Operation Corporation Limited (POSOCO). In experiments in Sec. 4.2, we set $T = 300$ and $T_1 = 1$ (i.e., predicting the state energy consumption next day). The predictions were carried out for last 6 months (170 days) in each of the two datasets.

Household Energy Consumption: This dataset was obtained from the publicly available repository (Gopinadhan, 2020). It contains household energy consumption of an apartment unit in San Jose for approximately a year. The data has been collected using the smart meters of the energy company. In experiments in Sec. 4.2, we set $T = 300$ and $T_1 = 1$ (i.e., predicting the energy consumption of the apartment on the next day). The predictions were carried out for last 2 months (53 days) in the dataset.

SP500 Index: This dataset was obtained from the Yahoo Finance (Yahoo Finance, 2023) on 19-Jan-2023. We set $T = 100$ (1 month) and $T_1 = 1$ (i.e., predicting whether the index will incur a loss or a gain on the next day). The predictions were carried out for last month (20 business days) in the dataset.

In addition to Table 1 which summarizes the empirical results, Fig. 3 presents a time series visualization of the predicted set of states $\mathcal{C}_{1-\alpha}$ together with the true state X_t for each of the above four datasets.
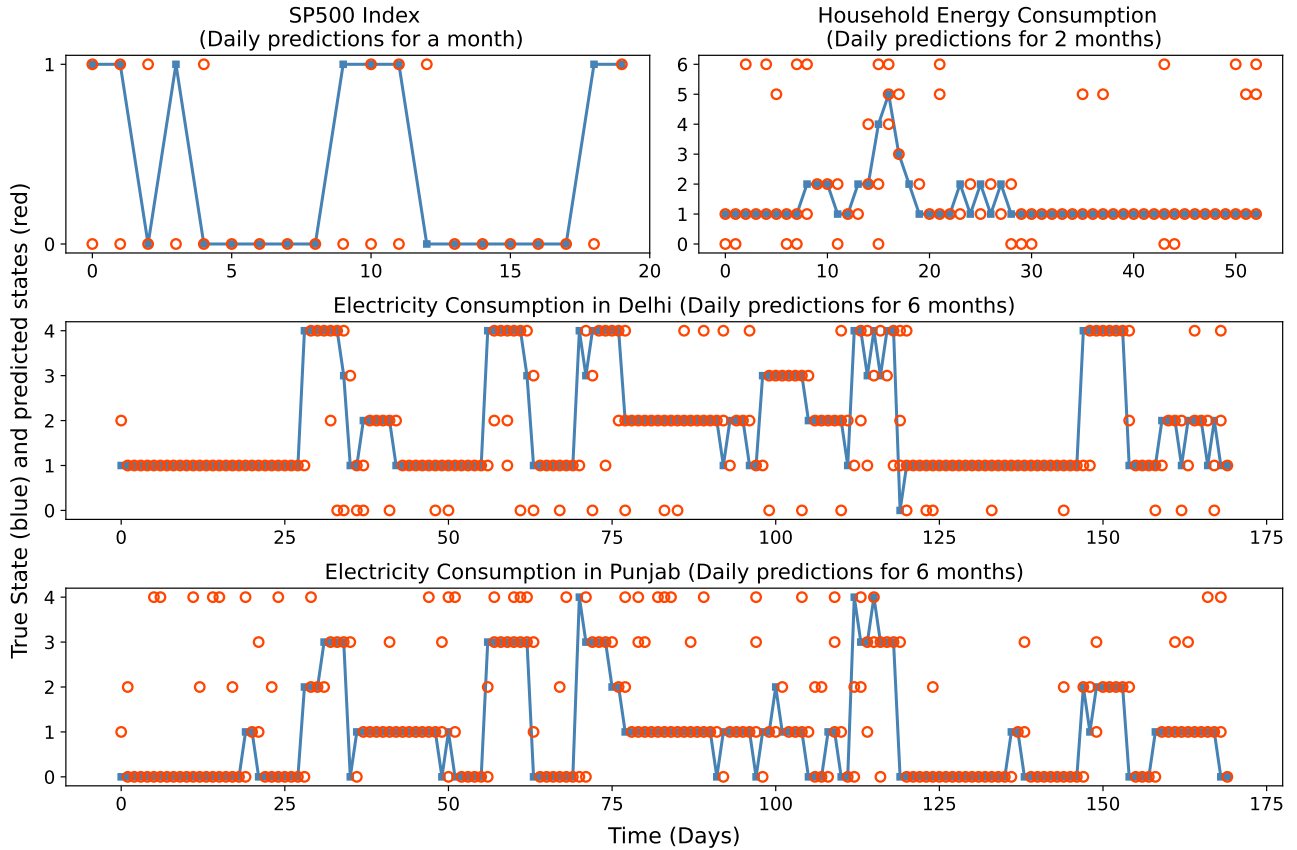

Figure 3. A time series visualization of the true state (blue) and the predicted set of states for each dataset in Table 1 used to obtain the empirical results in Sec. 4.2. For each dataset, the empirical coverage is approximately equal to the desired coverage 0.8. For predicting energy consumption (first three datasets in Table 1), the scaled average prediction set size is also smaller than 0.28. Hence, an 80% coverage is obtained with less than 28% of the states, indicating that the uncertainty is low. To obtain the same coverage when predicting the SP500 index (daily gain or loss), 65% of the states has been needed, indicating that the uncertainty is high. Thus, the proposed algorithm yields the desired performance in practice and helps deal with uncertainty in sequential prediction tasks.