0.1 Theoretical Analysis of Proposed Approach

In this section, we present preliminaries of Markov chain and the maximum likelihood estimator of the transition probabilities, and we describe the theoretical properties of our proposed synchronization operator and its relation with the maximum likelihood estimator.

Preliminaries

In this section, we first present some definitions related to Markov chain theory, where the theoretical definitions presented are based on the work described in [2, 3, 1, 5].

Definition 1. *Let* $\{s_0, s_1, \dots s_n\}$ *be a sequence of random variables as* ***Markov chain****, where* s_i *belongs to a finite state space* $\mathbf{S} = \{\mathbf{1}, \dots \mathbf{m}\}$ *and represents the observed state of the chain at time* i*. Let the transition probabilities of the Markov chain* $p_{ij}(t+1)$ *such that* $i, j \in S$ *and* $t = 0, \dots, n$*, where* $p_{ij}(t+1)$ *is the probability of the state* j *at time* $t+1$*, given state* i *at time* t*, where the sequence* $\{s_0, s_1, \dots s_n\}$ *satisfies the* ***Markov property***

$$P(s_{t+1} = j | s_t = i, s_{t-1} = i_{t-1}, \dots, s_0 = i_0) = P(s_{t+1} = j | s_t = i) \quad \forall i, j, i_{t-1}, i_0 \in S \tag{0.1}$$

Thus, the probability of moving to a future state only depends on the current state (first-order Markov chain). While for higher order m *Markov chains the conditional probabilities can be modeled to be dependent on the last* m *states.*

When the conditional probabilities $P(s_{t+1} = j | s_t = i)$ *are independent of the time* t*, the Markov chain is called* ***homogeneous*** *such that* $p_{ij} := P(s_{t+1} = j | s_t = i)$*.*

The transition probabilities of the Markov chain are represented by a $m \times m$ matrix that called **transition probability matrix** $\mathbf{\Pi}$ with p_{ij} elements

$$\mathbf{\Pi} = \begin{pmatrix} p_{1,1} & p_{1,2} & . & . & . & p_{1,m} \\ p_{2,1} & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ p_{m,1} & p_{m,1} & . & . & . & p_{m,m} \end{pmatrix} \tag{0.2}$$

where $0 \leq p_{i,j} \leq 1$ and the rows sum up to one

$$\sum_{j=1}^{m} p_{i,j} = 1 \qquad i = 1, 2 \dots m \tag{0.3}$$

Learning the Transition Probability Matrix. As mentioned in Section **??**, we rely on the transition probability matrix of $PMC_m^{\mathcal{P}}$ to build the predictions table. However, in practice the underlying transition probability matrix is unknown, and desirable to estimate or learn it form the observed sequence $\{s_0, s_1, \dots s_n\}$. The maximum likelihood estimator (MLE) is a common method to estimate the transition probability matrix [1].

Definition 2. *Let* $\mathbf{\Pi}$ *is the transition probability matrix of a single Markov chain with a set of states* S, $\pi_{i,j}$ *the transition probability from state* i *to state* j, $n_{i,j}$ *the number of observed transitions from state* i *to state* j, *then the maximum likelihood estimator finds* $\hat{\mathbf{\Pi}}$ *as an estimate for* $\mathbf{\Pi}$, *where its elements* $\hat{p}_{i,j}$ *are*

$$\hat{p}_{i,j} = \frac{n_{i,j}}{\sum_{l \in S} n_{i,l}} = \frac{n_{i,j}}{n_i} \tag{0.4}$$

The maximum likelihood estimates of transition probabilities of a single sequence $\{s_0, s_1, \dots s_n\}$ are obtained based on the observed transitions between the states of the chain. That is, the maximum likelihood estimates are basically the count of transitions from i to j divided by the total count of the chain being in state i.

Anderson and Goodman [1] have shown that

$$\begin{aligned} \sqrt{n}\ (\hat{p}_{i,j} - p_{i,j}) \xrightarrow{d} \mathcal{N}(\mu,\, \sigma^2_{mle_n}) \\ as\ n \to \infty \end{aligned} \tag{0.5}$$

Thus, the random variable $\sqrt{n}\ (\hat{p}_{i,j} - p_{i,j})$ has asymptotically normal distribution with mean $\mu = 0$. Therefore, the MLE is an asymptotically normal. While the variance $\sigma^2_{mle_n}$ is given by

$$\begin{aligned} \sigma^2_{mle_n} = \mathrm{Var}(\sqrt{n}\ (\hat{p}_{i,j} - p_{i,j})) = \frac{p_{i,j}\ (1 - p_{i,j})}{\phi_i} \\ \text{s.t. } \phi_i = \sum_{l=1}^{m} \sum_{t=1}^{n} \eta_l\ p_{l,j}^{t-1} \end{aligned} \tag{0.6}$$

Where $p_{l,j}^{t-1}$ is the probability of state j at time $t-1$ given that the state l at time 0 [1]. We are interested in the variances of $(\hat{p}_{i,j} - p_{i,j})$ that represents the

error in estimating $p_{i,j}$ by MLE, which is given by:

$$\mathrm{Var}(\hat{p}_{i,j} - p_{i,j}) = \frac{\sigma^2_{mle_n}}{n} \tag{0.7}$$

It is clearly seen that variances are dropping as the sample size n grows large. In next, we will show that our proposed approach of synchronizing the maximum likelihood estimators over k chains is preserving a similar asymptotic behavior.

0.1.1 Properties of Proposed Approach

The proposed synchronization operator is basically aggregating the maximum likelihood estimates over k observed sequences (i.e., sequences of the DFA states based on the consumed event streams), the operator estimates the maximum likelihood of the probabilities for a set of k sequences, which are arranged in serial order as one large chain with length $N = kn$ where we assume that all k sequences have n observations. For the sake of simplicity, we assume that the synchronization phase happens on batch size equals n (i.e., $b = n$) the, then it follows that

$$\begin{aligned} \hat{\pi}_{i,j} = \frac{\sum_{k \in K} n_{k,i,j}}{\sum_{k \in K} \sum_{l \in L} n_{k,i,l}} = \hat{p}_{i,j}(N) \\ where\ N = kn. \end{aligned} \tag{0.8}$$

Thus, this operation it allows to observe more samples, which is naturally producing a better estimates of the transition probabilities. In addition, our proposed synchronization operation of the k transition matrices has the same proprieties as the maximum likelihood estimator over a serial sequence of all k sequences, but with skipping $k-1$ transitions between each two consecutive sequences, which is in practice a small number that can be neglected comparing to the total transitions count kn. As result, the probabilities estimates of our estimator (i.e.,global) based on the proposed operation within the distributed online learning protocol have the same properties as maximum likelihood estimates, in particular, the the random variable $\sqrt{N}\ (\hat{\pi}_{i,j} - p_{i,j})$ has asymptotically normal distribution with mean $\mu = 0$ following Equation 0.5 we have

$$\begin{aligned} \sqrt{N}\ (\hat{\pi}_{i,j} - p_{i,j}) \xrightarrow{d} \mathcal{N}(0,\ \sigma^2_{mle_N}) \\ as\ N \to \infty \\ where\ N = nk. \end{aligned} \tag{0.9}$$

So,

$$\mathrm{Var}(\hat{\pi}_{i,j} - p_{i,j}) = \frac{\sigma^2_{mle_N}}{N} = \frac{\sigma^2_{mle_n}}{kn} \tag{0.10}$$

That is, since $N > n$ combining k sequences, the variances of our method estimates $\mathrm{Var}(\hat{\pi}_{i,j} - p_{i,j})$ are smaller than the estimates of MLE over an isolated sequence $\mathrm{Var}(\hat{p}_{i,j} - p_{i,j})$. Thus, it follows from the Chebyshev's inequality [4] that we have for the random variable $\hat{p}_{i,j} - p_{i,j}$, for any constant $c > 0$

$$\Pr\left(|(\hat{p}_{i,j} - p_{i,j}) - \mu| \geq c\right) \leq \frac{\mathrm{Var}(\hat{p}_{i,j} - p_{i,j})}{c^2}$$

where the mean $\mu = 0$ is zero and the $\mathrm{Var}(\hat{p}_{i,j} - p_{i,j})$ equals $\frac{\sigma^2_{mle_n}}{n}$, and therefore

$$\Pr\left(|\hat{p}_{i,j} - p_{i,j}| \geq c\right) \leq \frac{\sigma^2_{mle_n}}{c^2 n}$$

$\hat{p}_{i,j} - p_{i,j}$ represents the deviation/error between the estimates of MLE over a single (i.e., isolated) sequence and the true probabilities. On the other hand, we can obtain, in the same way, the probability bound of deviations for our synchronization operator estimates as follows:

$$\Pr\left(|\hat{\pi}_{i,j} - p_{i,j}| \geq c\right) \leq \frac{\sigma^2_{mle_n}}{c^2 nk}$$

Using Equation 0.10 we obtained the value $\mathrm{Var}(\hat{\pi}_{i,j} - p_{i,j})$. Since $k \geq 1$ we have that the variance of $(\hat{\pi}_{i,j} - p_{i,j})$ is less than or equal to the variance of $\hat{p}_{i,j} - p_{i,j}$

$$\frac{\sigma^2_{mle_n}}{c^2 nk} \leq \frac{\sigma^2_{mle_n}}{c^2 n}$$

This is equivalent to,for any constant $c > 0$ and $k \geq 1$ we have

$$\Pr\left(|\hat{\pi}_{i,j} - p_{i,j}| \geq c\right) \leq \Pr\left(|\hat{p}_{i,j} - p_{i,j}| \geq c\right)$$

To summarize, our approach is based aggregating the MLE estimates over k sequences, which speeds up the convergence to reach the true transition probabilities as result of the smaller variances.

0.1.2 Computing the Transition Matrix From the Matrix of $PMC^{\mathcal{P}}_m$

In order to empirically study the asymptotic behavior of our proposed synchronization operator, we need to compute the transition probability matrix of the underlying Markov chain that the events belong to, and we introduce to calculate

it based on the transition matrix (Π) of $PMC_m^{\mathcal{P}}$ that describes the Markov chain of the pattern.

Nuel [6] showed in **Theorem 3** the relation between the elements of Π and the conditional probabilities of the $m-$order Markov chain $X = \{X_1, X_2, \dots X_n\}$ described by

$$\Pi(p,q) = \begin{cases} P(X_{m+1} = b | X_1 \dots X_m = \delta^{-m}(p)) & \text{if } \delta(p,q) = b \\ 0 & \text{if } p \notin \delta(p, X) \end{cases}$$

Using this theorem, we can compute the transition probabilities of the Markov chain X.

Bibliography

[1] Theodore W Anderson and Leo A Goodman. Statistical inference about markov chains. *The Annals of Mathematical Statistics*, pages 89–110, 1957.

[2] Dimitri P Bertsekas and John N Tsitsiklis. *Introduction to probability*, volume 1. Athena Scientific Belmont, MA, 2002.

[3] P Billingsley. Statistical Methods in Markov Chains. *The Annals of Mathematical Statistics*, 32(1):12–40, 1961. ISSN 00034851. doi: 10.1214/aoms/1177705148. URL `http://www.jstor.org/stable/2238700`.

[4] William Feller. *An introduction to probability theory and its applications: volume I*, volume 3. John Wiley & Sons New York, 1968.

[5] Ronald A Howard. *Dynamic probabilistic systems: Markov models*, volume 1. Courier Corporation, 2012.

[6] Grégory Nuel. Pattern Markov Chains: Optimal Markov Chain Embedding through Deterministic Finite Automata. *Journal of Applied Probability*, 2008.