A continuous-time model-based approach for activity recognition in pervasive environments: Supplemental Material

Marco Biagi, Laura Carnevali, Marco Paolieri, Fulvio Patara, and Enrico Vicario, IEEE member

I. STOCHASTIC STATE CLASSES

In the method of stochastic state classes, the stochastic process $\mathbb{X} := \{s(t), \; t \in \mathbb{R}_{\geq 0}\}$, where $s(t)$ is the state of the model at time $t$, is sampled after each observable or unobservable transition, representing the continuous sets of states that can be reached after any transition sequence. To support transient analysis of the model, an additional timer termed $\tau_{age}$ is maintained in each stochastic state class to keep track of the absolute time of each transition [32].

Definition 1. A stochastic state class (class for short) is a tuple $\Sigma = \langle \psi, D(\theta, t, \tau_{age}), f(\theta, t, \tau_{age}) \rangle$ where $\psi \in A_*$ is an activity, and $D(\theta, t, \tau_{age}) \subseteq \mathbb{R}_{\geq 0} \times \mathbb{R} \leq 0$ and $f(\theta, t, \tau_{age}) : D(\theta, t, \tau_{age}) \rightarrow [0, 1]$ are the support and the PDF, respectively, of the random vector $(\theta, t, \tau_{age})$ made of the remaining sojourn time $\theta$ in $\psi$, the remaining time $t$ to the next observable event in $\psi$, and the absolute elapsed time $\tau_{age}$.

As in [32], $\tau_{age}$ is the opposite of the absolute elapsed time and is decreased at each transition by the time elapsed in the previous class. Hence, $\theta$, $t$, and $\tau_{age}$ evolve with the same speed and slope, and their joint support is a Difference Bounds Matrix (DBM), i.e., a set of linear inequalities constraining the difference between pairs of timers, guaranteeing manipulation and encoding in polynomial time [33].

Definition 2. $\Sigma' = \langle \psi', D'(\theta, t', \tau_{age}'), f'(\theta, t', \tau_{age}') \rangle$, is the successor of $\Sigma = \langle \psi, D(\theta, t, \tau_{age}), f(\theta, t, \tau_{age}) \rangle$ through transition $t$ with probability $\mu$ if when the location is $\psi$ and $(\theta, t, \tau_{age})$ has support $D(\theta, t, \tau_{age})$ and PDF $f(\theta, t, \tau_{age})$, transition $t$ occurs with probability $\mu$, yielding a location $\psi'$ and a random vector $(\theta', t', \tau'_{age})$ with support $D'(\theta, t, \tau_{age})$ and PDF $f'(\theta, t, \tau_{age})$.

The calculus of stochastic state classes is tailored to support transient analysis of the model of Sect. 3.1. Specifically, the successor class $\Sigma'$ of a class $\Sigma$ through a transition $t$ with probability $\mu$ is derived distinguishing four cases depending on the occurred transition.

1) The unobservable transition $t = (a, \phi)$ can occur in $\Sigma$ iff: i) the current location is $a_*$ i.e., $\psi = a_*$; ii) domain $D(\theta, t, \tau_{age})$ conditioned to the assumption that $\theta$ is not larger than $t$ is not empty, i.e., $D_t = D(\theta, t, \tau_{age}) \cap \{ x_\theta \leq x_1 \} = \emptyset$; iii) the occurrence probability of $t$ is larger than zero, i.e., $\mu = \int_{D_t} f(\theta, t, \tau_{age})(x_\theta, x_1, x_{age})dx_\theta dx_1 dx_{age} > 0$. The occurrence of $t$ in $\Sigma$ yields the deadlock class $\Sigma' = \langle \Omega, \cdot, \cdot \rangle$, which has no successor.

2) The observable transition $t = (a, e, a)$, with $e \in E$ and $a \in A$, can occur in $\Sigma$ iff: i) $\psi = a$; ii) $D_t = D(\theta, t, \tau_{age}) \cap \{ x_\theta > x_1 \} \neq \emptyset$; iii) $\mu = P_{e}(\psi) \cdot P_a(e, a) \cdot \int_{D_t} f(\theta, t, \tau_{age})(x_\theta, x_1, x_{age})dx_\theta dx_1 dx_{age} > 0$. The occurrence of $t$ in $\Sigma$ yields a location $\psi'$ = $a$ and a random vector $(\theta', t', \tau'_{age})$ distributed over $D'(\theta', t', \tau'_{age}) = [F_{\min}(\theta, t, \tau_{age})] \times [F_{\max}(\theta, t, \tau_{age})] \times D(\tau_{age})$ according to $f'(\theta', t', \tau'_{age})(x_\theta, x_1, x_{age}) = F_a(a)(x_\theta) \cdot F_{\theta}(\psi)(x_1) \cdot f(\theta, t, \tau_{age})(x_{age})$, where $D(\tau_{age})$ and $f(\tau_{age})$ are the domain and PDF of $\tau_{age}$ in $\Sigma$.

3) The unobservable transition $t = (a, b)$, can occur in $\Sigma$ iff: i) $\psi = a$; ii) $D_t = D(\theta, t, \tau_{age}) \cap \{ x_\theta < x_1 \} = \emptyset$; iii) $\mu = \int_{D_t} f(\theta, t, \tau_{age})(x_\theta, x_1, x_{age})dx_\theta dx_1 dx_{age} > 0$. If $t$ occurs in $\Sigma$, $\psi = a_*$ and $(\theta', t', \tau'_{age})$ is distributed over $D'(\theta', t', \tau'_{age}) = [F_{\min}(\theta, t, \tau_{age})] \times [F_{\max}(\theta, t, \tau_{age})] \times D(\tau_{age})$ according to $f'(\theta', t', \tau'_{age})(x_\theta, x_1, x_{age}) = F_a(a)(x_\theta) \cdot F_\psi(x_1) \cdot f(\theta, t, \tau_{age})(x_{age})$, where $D(\tau_{age})$ and $f(\tau_{age})$ are the domain and PDF of $\tau_{age}$ in $\Sigma$.

4) The observable transition $t = (a, e, a)$, with $a \in A_*$, can occur in $\Sigma$ iff: i) $\psi = a$; ii) $D_t = D(\theta, t, \tau_{age}) \cap \{ x_\theta > x_1 \} \neq \emptyset$; iii) $\mu > 0$, with $\mu = P_{e}(\psi) \cdot (1 - \sum_{a \in A} P_a(e, a) \cdot p)$ if $a = a_*$ and $\mu = P_a(e, a) \cdot p$ if $a \neq a_*$, with $p = \int_{D_t} f(\theta, t, \tau_{age})(x_\theta, x_1, x_{age})dx_\theta dx_1 dx_{age}$.

The successor of $\Sigma$ through $t$ is derived as follows:

- **Conditioning.** The assumption that $\theta$ is larger than $t$ yields a new random vector $(\theta', t', \tau'_{age})$ distributed over $D_a = D_t$ according to $f_a(x_\theta, x_1, x_{age}) = f(\theta, t, \tau_{age})(x_\theta, x_1, x_{age}) / \mu$.

- **Time advancement and projection.** The observation of event $e$ reduces $\theta$ and $\tau_{age}$ by $t$ and eliminates $\psi$ itself, yielding a random vector $(\theta', \tau'_{age})$ distributed over the projection of $D_0$ that eliminates $x_1$, i.e., $D_0 = \{(x_\theta, x_{age}) \mid \exists x_1 \in \mathbb{R}_{\geq 0}$ such that $\{ x_\theta, x_1, x_{age} \} \subseteq D_0 \}$, according to $f_0(x_\theta, x_1, x_{age}) = \int_{D_0} f_a(x_\theta + x_1, x_{age} + x_1)dx_1$, where $l$ and $u$ are the minimum and the maximum value of $x_1$ such that $\{ x_\theta + x_1, x_{age} \} \subseteq D_0$, respectively.

- **Newly enabling.** The random vector $(\theta', \tau'_{age})$ is augmented with the remaining time to the next observable event, yielding $(\theta', t', \tau'_{age})$ distributed over $D' = D_0 \times [F_{\min}(\theta, t, \tau_{age})]$, according to $f'(\theta', t', \tau'_{age})(x_\theta, x_1, x_{age}) = f_0(x_\theta, x_{age}) \cdot f_\psi(x_1)$. 


II. FITTING TWO MOMENTS WHEN CV < 1

A. Preliminaries

Lemma 1. Let Z be the sum of m exponential random variables, i.e., \( Z := \sum_{i=1}^{m} X_i \) with \( X_i \sim \text{Exp}(\lambda_i) \). Then, the coefficient of variation (CV) of Z is minimum when the rates are all equal, i.e., \( \lambda_1 = \lambda_2 = \ldots = \lambda_m \).

Proof. Expected value and variance are linear operators, so the CV of Z is:

\[
CV = \frac{\sqrt{\operatorname{Var}(Z)}}{E[Z]} = \frac{\sqrt{\sum_{i=1}^{m} \operatorname{Var}(X_i)}}{\sum_{i=1}^{m} E[X_i]} = \frac{\sqrt{\sum_{i=1}^{m} \beta_i^2}}{\sum_{i=1}^{m} \beta_i}
\]

where \( \beta_i := E[X_i] = 1/\lambda_i \) and \( \beta_i^2 = 1/\lambda_i^2 = \operatorname{Var}(X_i) \).

Since \( \beta_i \geq 0 \) for all \( i \), letting \( \beta = (\beta_1, \ldots, \beta_m) \), we obtain that CV = \( \frac{\beta_1}{\beta_1^2} \). By the Cauchy-Schwarz inequality, it holds that \( \|\beta\|_1 = \beta \cdot 1_m \leq \|\beta\|_2 = \sqrt{\sum \beta_i^2} \), which, in turn, implies that CV = \( \frac{\|\beta\|_2}{\|\beta\|_1} \) is minimum when \( \beta \) and \( 1_m \) are parallel (so that their scalar product is maximum). This means that \( \beta = 1_m/\lambda \) for some \( \lambda \), i.e., all the rates of the exponentials \( X_i \) are equal. \( \square \)

B. Using the sum of an Erlang and an Exponential distribution

Let Z be the sum of a random variable having Erlang distribution with shape \( n \) and rate \( \lambda_1 \) and an exponential random variable with rate \( \lambda_2 \). Z is an hypo-exponential, namely the sum of \( m := n + 1 \) exponentials. Hence, by Lemma 1, Z has minimum CV equal to \( 1/\sqrt{m} \) when \( \lambda_2 = \lambda_1 \).

If \( \mu \) is the mean and \( S \) is the standard deviation of our samples such that CV = \( S/\mu < 1 \), our strategy is to have \( m \) large enough for the input CVing and then, given \( m, \mu \), and \( S \) using \( \lambda_1 \) and \( \lambda_2 \). To this end, first we pick \( m = \lceil \mu/S \rceil \) that is, the minimum \( m \) guaranteeing that \( 1/\sqrt{m} \leq S/\mu \), so that the hypoexponential has enough phases to have an equal or lower CV than the input data (by hypothesis, \( S/\mu < 1 \)). According to this, \( n \) is set equal to:

\[
n = \left\lceil \frac{1}{\sqrt{CV^2}} \right\rceil - 1.
\]

Now, we impose \( \mu = n/\lambda_1 + 1/\lambda_2 \) and \( S^2 = n/\lambda_1^2 + 1/\lambda_2^2 \), where \( n = m - 1 \). From the first equation, we obtain that \( 1/\lambda_2 = \mu - n/\lambda_1 \). Plugging this equality into the second equation, we get:

\[
CV^2 = \frac{n/\lambda_1^2 + (\mu - n/\lambda_1)^2}{CV^2 + \mu^2/\mu^2 - 2\mu/\lambda_1^2 + \mu^2/\mu^2 - 2\mu/\lambda_1^2 + \mu^2},
\]

and rearranging:

\[
(\mu^2 - CV^2 \mu^2)\lambda_1^2 - (2\mu \lambda_1) + (n^2 + n) = 0.
\]

The discriminant of the latter equation is always non-negative. In fact: \( 4n^2 \mu^2 - 4(\mu^2 - CV^2 \mu^2)(n^2 + n) \geq 0 \iff \frac{n^2 + n}{\mu^2} \geq 1/n \), which is true because \( S/\mu \geq 1/\sqrt{m} \) for our choice of \( n = m - 1 \). And, at least one solution must be non-negative (because \(-2\mu\lambda_1 \) is negative and \( \mu^2 - CV^2 \mu^2 \) is positive), so we evaluate \( \lambda_1 \) as a non-negative solution of the considered second-degree equation.

Then, we use the first equation again to evaluate \( \lambda_2 \):

\[
\lambda_2 = \frac{1}{\mu - n/\lambda_1}.
\]

Finally, once the parameters of the Erlang-distributed random variable and the exponentially distributed random variable have been determined, the PDF of their sum can be derived by convolution, for instance by applying the method of: M. Akkouchi, “On the convolution of exponential distributions”, Journal of the Chungcheong Mathematical Society, vol. 21, no. 4, pp. 501–510, 2008.

III. THEOREM PROOFS

Theorem 1. \( \forall \) class \( \Sigma = (\psi, D_{\theta_0, \tau_{\text{age}}}, f_{\theta_0, \tau_{\text{age}}}) \in \Gamma_z \), with \( n > 0 \), the random variables \( \theta, \iota, \) and \( \tau_{\text{age}} \) are independently distributed, and the random vector \( (\theta_{\text{age}}, \tau_{\text{age}}) \) has domain \( D_{\theta_0, \tau_{\text{age}}} = D_\theta \times [F_{\min}^{\max}, F_{\max}^{\max}] \times [0, T] \) and PDF \( f_{\theta_0, \tau_{\text{age}}}(x_\theta, x_{\tau}) = f_{0}(x_{\theta}) f_{1}(x_{\tau}) \delta(x_{\tau}, \theta) \), where \( D_\theta \) and \( f_0 \) are the marginal domain and the marginal PDF of \( \theta \), respectively. If entering class \( \Sigma \) accounts for starting activity \( \psi \), then \( D_\theta = [F_{\min}^{\max}, F_{\max}^{\max}] \) and PDF \( f_0(x_{\theta}) = F_{\theta}(x_{\theta}) \). Otherwise (i.e., if entering \( \Sigma \) represents a continuation of \( \psi \)), \( D_\theta = \{ x_{\theta} | x_{\theta} + \Delta_n \in D_\theta \cap \{ x_{\theta} \geq \Delta_n \} \} \) and PDF \( f_0(x_{\theta}) = f_0(x_{\theta} + \Delta_n) / f_{\Delta}(x_{\theta}) dy \), where \( \Delta_n = T_n - T_{n-1} \), and \( D_\theta \) and \( f_0 \) are the marginal domain and the marginal PDF, respectively, of the remaining sojourn time \( \theta_p \) in the class \( \Sigma_p \in \Gamma_p \) (at the previous event) that yields \( \Sigma \).

Proof. As illustrated in Sect. IV-B1, the set of classes \( \Gamma_n \) is derived in two steps: first, each class \( \Sigma_{n-1} \in \Gamma_{n-1} \) is conditioned to the observation of event \( e_n \), yielding the intermediate set of classes \( \Gamma_{n-1} \); then, each class \( \Sigma_{n-1} \) is conditioned to the observation time \( t_n \) imposing \( \tau_{\text{age}} = t_n - t_{n-1} \), yielding class \( \Sigma_n \) of the current transition \( \psi \) (see the previous event) that yields \( \Sigma \).

The conditioning \( \tau_{\text{age}} = t_n - t_{n-1} \) does not affect the current activity, hence \( \psi = \psi \).

Given that \( \Sigma_{n-1} \) is reached through an observable transition \( (\cdot, e_n, \iota, \psi) \) representing the occurrence of \( e_n \), the remaining time \( \iota \) to the next observable event is independent of \( \tau_{\text{age}} \) and the remaining sojourn time \( \theta \), and it is distributed according to \( F_{\psi} \) over \( [F_{\min}^{\max}, F_{\max}^{\max}] \), hence, \( \iota \) is not affected by the fact that \( \tau_{\text{age}} = \Delta_n - T_n - T_{n-1} \), and thus \( \tau_{\text{age}} \).

If \( \theta \) and \( \tau_{\text{age}} \) are independently distributed, which occurs if entering \( \Sigma_{n-1} \) corresponds to starting the activity \( \psi \) in \( A_n \) (e.g., in Fig. 5, entering \( \Sigma_1 \) corresponds to starting \( a_1 \)), then \( \theta \) is not conditioned by the value of \( \tau_{\text{age}} \), and \( \theta = \hat{\theta} \) is distributed over \( D_\theta = [F_{\min}^{\max}, F_{\max}^{\max}] \) according to PDF \( f_0(x_{\theta}) = F_{\psi} \).

If \( \theta \) and \( \tau_{\text{age}} \) are dependent random variables, which occurs if entering \( \Sigma_{n-1} \) corresponds to continuing the activity \( \psi \) in \( A_n \) (e.g., in Fig. 5, entering \( \Sigma_1 \) corresponds to continuing \( a_1 \)), then the fact that \( \tau_{\text{age}} = t_n - t_{n-1} \) (that \( \theta \) is not conditioned by the value of \( \tau_{\text{age}} \), and \( \theta \) is distributed according to PDF \( f_0(x_{\theta}) = F_{\psi} \)).
Therefore, the random variables θ, τ, and τage are independently distributed, and the random vector (θ, τ, τage) has domain \( D_{θ,τ,τage} = D_θ \times [F_{min}^1, F_{max}^1] \times [0, 0] \) and PDF \( f_{θ,τ,τage}(x, y, z) = f_θ(x) F_1(\delta_z) \delta_0(\tauage) \).

**Theorem 2.** ∀ \( n > 0 \), the number of classes in \( \Gamma_n \) is bounded by \( Q_{max} = (|A| + 1) \cdot 2^C \), where \(|A|\) is the number of annotated activities, and \( C \) is the maximum number of consecutive events that may represent the start or the continuation of an activity.

**Proof.** According to the semantics of the model (defined in Section III-A) and the calculus of the plausible classes after a sequence of observations (illustrated in the Proof of Theorem 1), starting from a class where the logical location is an annotated activity \( a \in A \) (e.g., \( a_2 \) in class \( \Sigma_1^b \) of Fig. 5), the system behavior may evolve in three different manners:

1. \( a \) is continued, leading to a class where \( θ \) is reduced by the time elapsed between the last two observed events (e.g., \( \Sigma_1^c \), \( Σ_1^d \), and \( Σ_1^e \) in Fig. 5);
2. \( a \) is completed and an activity \( ψ \in A \) is started, yielding a class where \( θ, τ \), and \( τage \) are independently distributed according to \( F_2(ψ) \), \( F_1(ψ) \), and \( δ(xage) \), over \([F_{min}^2, F_{max}^2]\), \([F_{min}^1, F_{max}^1]\), and \([0, 0]\), respectively (e.g., \( Σ_1^b \) and \( Σ_1^c \) in Fig. 5);
3. \( a \) is completed and the Idling activity is started and then continued, also yielding a class where \( θ \) is reduced by the time elapsed between the last two observed events (e.g., \( Σ_1^b \), \( Σ_1^c \), \( Σ_1^d \), and \( Σ_1^e \) in Fig. 5).

Similarly, if the logical location of the considered starting class were Idling, either Idling is continued (which is similar to case 1) or it is completed and an annotated activity is started (which is similar to case 2).

On the one hand, all the classes yielded by case 2 with the same logical location are equal to each other and known a priori, so that case 2 yields at most a class for each annotated activity, i.e., at most \(|A|\) classes. On the other hand, cases 1 and 3 yield one class each, which is derived based on the observed time-stamps. But, the class representing a continuation of Idling (case 3) is the same for all plausible classes that have been reached by the same observation sequence and have a tagged activity as logical location.

According to this, after the observation of the first event, the initial class may yield at most \(|A| + 1\) classes, which occurs if the first event represents the continuation of the initial Idling activity or the start of any annotated activity. After the observation of the second event, each of these at most \(|A| + 1\) classes may yield at most 1 new different class, which occurs if the second event represents the continuation of any activity (including Idling). Moreover, overall these classes may also yield \( A \) classes representing the start of a tagged activity, and a class representing a continuation of Idling. Therefore, after the second event, the number of plausible classes is bounded by \((|A| + 1) \cdot 2^C\). After the observation of \( n \) consecutive events that may start or continue any activity, the number of classes is bounded by \((|A| + 1) \cdot 2^n\). Whenever an event is observed that may not continue the activity of a plausible class or may not start an activity from each plausible class, the number of possible new classes decreases. Therefore, if \( C \) is the maximum number of consecutive events that may represent the start or the continuation of an activity, the number of plausible classes after any observation is bounded by \((|A| + 1) \cdot 2^C\). □

**Theorem 3.** ∀ \( n > 0 \) and ∀ \( \Sigma_n^m \in Γ_n \), it holds that:

\[
\chi(\Sigma_0, \Sigma_1^{\rightarrow}, \Sigma_m^n) = \sum_{\Sigma_{n-1}^{\rightarrow} \in Γ_{n-1}^{parent}} \chi(\Sigma_0, \Sigma_1^{\rightarrow}, \Sigma_m^n_{-1}) \cdot P(\Sigma_{m}^{n} \mid \Sigma_{m-1}^{n})
\]

where: \( Γ_{n, m} \subseteq Γ_{n-1} \) is the set of classes that, conditioned on \( ω_{n-1} = (\epsilon_{n-1}, t_{n-1}) \), yield \( \Sigma_{m-1}^{n} \); \( P(\Sigma_{m}^{n} \mid \Sigma_{m-1}^{n}) \) is the probability that \( ω_{n} \) is observed and class \( \Sigma_{m}^{n} \) is reached at time \( t_{n} \) given that class \( \Sigma_{m-1}^{n} \) was entered at time \( t_{n-1} \); and, \( \chi(\Sigma_0, \Sigma_1^{\rightarrow}, \Sigma_0) = 1 \) is the probability of the initial class.

**Proof.** By induction, in the base case \( n = 1 \), \( \chi(\Sigma_0, \Sigma_1^{\rightarrow}, \Sigma_1^{m}) = \chi(\Sigma_0, \Sigma_1^{\rightarrow}, \Sigma_0)P(\Sigma_1^{m} \mid \Sigma_0) \). Hence, by the conditional probability, \( \chi(\Sigma_0, \Sigma_1^{\rightarrow}, \Sigma_1^{m}) = P(\Sigma_1^{m} \mid \Sigma_0) \). In the inductive step, we assume that the thesis holds for some natural number \( n - 1 > 0 \) and we prove that it holds for \( n \). By the definition of forward variable and the definition of stochastic state class, \( \chi(\Sigma_0, \Sigma_1^{\rightarrow}, \Sigma_1^{m}) = \sum_{\Sigma_{n-1}^{\rightarrow} \in Γ_{n-1}^{parent}} P(\Sigma_{m}^{n} \mid \Sigma_0, \Sigma_1^{\rightarrow}, \Sigma_{n-1}^{m-1}) \cdot P(\Sigma_0, \Sigma_1^{\rightarrow}, \Sigma_{n-1}^{m-1}) \). By the definition of conditional probability, we obtain that \( \chi(\Sigma_0, \Sigma_1^{\rightarrow}, \Sigma_m^n) = \sum_{\Sigma_{n-1}^{\rightarrow} \in Γ_{n-1}^{parent}} P(\Sigma_0, \Sigma_1^{\rightarrow}, \Sigma_{n-1}^{m-1}) \). By the law of total probability, we have \( \chi(\Sigma_0, \Sigma_1^{\rightarrow}, \Sigma_m^n) = P(\Sigma_0, \Sigma_1^{\rightarrow}, \Sigma_m^n) \), which proves the thesis. □

**Theorem 4.** ∀ \( n > 0 \) and ∀ \( t \in (t_n, t_{n+1}) \), it holds that:

\[
L^\rightarrow_0(a, t, \Sigma_0, \Sigma_1^{\rightarrow}) = \sum_{\Sigma_m^n \in Γ_n} P(a, t, \Sigma_m^n) r^\rightarrow_0(\Sigma_m^n).
\]

**Proof.** By the law of total probability, it holds that:

\[
L^\rightarrow_0(a, t, \Sigma_0, \Sigma_1^{\rightarrow}) = \sum_{\Sigma_m^n \in Γ_n} P(a, t) = a \sum_{\Sigma_m^n \in Γ_n} P(\Sigma_0, \Sigma_1^{\rightarrow})
\]

Following from conditional probability, \( L^\rightarrow_0(a, t, \Sigma_0, \Sigma_1^{\rightarrow}) = \sum_{\Sigma_m^n \in Γ_n} r_0^\rightarrow P(a, t) = a \sum_{\Sigma_m^n \in Γ_n} P(\Sigma_0, \Sigma_1^{\rightarrow}) \), which finally proves the thesis. □

**Theorem 5.** ∀ \( n > 0 \) and ∀ \( \Sigma_m^n \in Γ_n \), it holds that:

\[
p^\rightarrow(\Sigma_m^n) = \frac{\chi(\Sigma_0, \Sigma_1^{\rightarrow}, \Sigma_m^n)\beta(\Sigma_m^n, \Sigma_{m+1}^n)}{\sum_{\Sigma_i^m \in Γ_m} \chi(\Sigma_0, \Sigma_1^{\rightarrow}, \Sigma_m^n)\beta(\Sigma_i^m, \Sigma_{m+1}^n)},
\]

where \( \beta(\Sigma_m^n, \Sigma_{m+1}^n) := P(\Sigma_{m+1}^n \mid \Sigma_m^n) \) is a backward variable defined as the probability of the observation sequence \( \omega_{n+1, N} \) conditioned to the fact that \( \Sigma_m^n \) is entered at time \( t_n \).

**Proof.** By the definition of conditional probability and by the law of total probability, \( p^\rightarrow(\Sigma_m^n) = P(\Sigma_0, \Sigma_1^{\rightarrow}, \Sigma_m^n) / \sum_{\Sigma_i^m \in Γ_m} P(\Sigma_0, \Sigma_1^{\rightarrow}, \Sigma_i^m) \). By the definition of conditional probability and the definition of stochastic state class:

\[
p^\rightarrow(\Sigma_m^n) = \frac{P(\Sigma_0, \Sigma_1^{\rightarrow}, \Sigma_m^n) P(\Sigma_1^{\rightarrow}, \Sigma_{m+1}^n) P(\Sigma_m^n)}{\sum_{\Sigma_i^m \in Γ_m} P(\Sigma_0, \Sigma_1^{\rightarrow}, \Sigma_i^m) P(\Sigma_1^{\rightarrow}, \Sigma_{m+1}^n) P(\Sigma_i^m)}.
\]
Again by the definition of conditional probability, we have:

\[ p_{\text{off}}(\Sigma_{n}^m) = \frac{p(\Sigma_0, \omega_{1,n}^+, \Sigma_{n}^m) p(\omega_{n+1,N}^+ | \Sigma_{n}^m)}{\sum_{\Sigma_i \in \Gamma_n} p(\Sigma_0, \omega_{1,n}^+, \Sigma_i) p(\omega_{n+1,N}^+ | \Sigma_i)}, \]

which finally proves the thesis. \( \square \)

**Theorem 6.** \( \forall 0 < n < N \) and \( \forall \Sigma_{n}^m \in \Gamma_n > 0 \), it holds that:

\[ \beta(\Sigma_{n+1}^m, \omega_{n+1,N}^+) = \sum_{\Sigma_i \in \Gamma_{n+1}^\text{child}} \beta(\Sigma_{n+1}, \omega_{n+2,N}^+) \cdot \frac{p(\omega_{n+1}, \Sigma_{n+1}^m | \Sigma_n^m)}{p(\omega_{n+1}, \Sigma_{n+1}^m | \Sigma_n^m)} \]

where: \( \Gamma_{n,m} \subseteq \Gamma_{n+1} \) is the set of classes yielded by \( \Sigma_{n}^m \) conditioned to observation \( \omega_{n+1} \); \( p(\omega_{n+1}, \Sigma_{n+1}^m | \Sigma_n^m) \) is derived through Eq. (7); and, \( \beta(\Sigma_{n}^m, \omega_{n+1,N}^+) = 1 \forall \Sigma_{n}^m \in \Gamma_N. \)

**Proof.** The proof is by induction. In the base case \( n = N - 1 \), \( \beta(\Sigma_N, \omega_{N,N}^+) = \sum_{\Sigma_i \in \Gamma_N^\text{child}} p(\omega_N, \Sigma_N | \Sigma_N^m) \), which yields \( \beta(\Sigma_N, \omega_{N,N}^+) = P(\omega_{N,N}^+ | \Sigma_N^m) \) by the law of total probability. In the inductive step, we assume that the expression of \( \beta(\Sigma_{n}^m, \omega_{n+1,N}^+) \) holds for some \( n + 1 \in (0, N - 1) \) and we prove that it holds for \( n \). According to this, by the definition of backward variable and stochastic state class, \( \beta(\Sigma_{n}^m, \omega_{n+1,N}^+) \) can be expressed as:

\[ \beta(\Sigma_{n}^m, \omega_{n+1,N}^+) = \sum_{\Sigma_i \in \Gamma_{n+1}^\text{child}} p(\omega_{n+2,N}^+ | \Sigma_i) \cdot \frac{p(\omega_{n+1}, \Sigma_{n+1}^m | \Sigma_n^m)}{p(\omega_{n+1}, \Sigma_{n+1}^m | \Sigma_n^m)}. \]

By the chain rule, we obtain that \( \beta(\Sigma_{n}^m, \omega_{n+1,N}^+) = \sum_{\Sigma_i \in \Gamma_{n+1}^\text{child}} p(\omega_{n+1,N}^+, \Sigma_{n+1}^m | \Sigma_{n+1}^m) / p(\Sigma_{n+1}^m) \). By the total probability law and the definition of conditional probability, \( \beta(\Sigma_n^m, \omega_{n+1,N}^+) = P(\omega_{n+1,N}^+ | \Sigma_n^m) \), proving the thesis. \( \square \)

**Theorem 6.** \( \forall n > 0 \) and \( \forall t \in (t_n, t_{n+1}) \), it holds that:

\[ L_{\text{off}}(a, t, \Sigma_0, \omega_{1,N}^+) = \sum_{\Sigma_n^m \in \Gamma_n} p(a, t, \Sigma_n^m) p_{\text{off}}(\Sigma_n^m). \]

**Proof.** By the definition of conditional probability and by the law of total probability, \( L_{\text{off}}(a, t, \Sigma_0, \omega_{1,N}^+) \) can be written as:

\[ L_{\text{off}}(a, t, \Sigma_0, \omega_{1,N}^+) = \sum_{\Sigma_n^m \in \Gamma_n} \frac{P(\alpha(t) = a, \Sigma_0, \omega_{1,N}^+, \Sigma_n^m)}{\sum_{\Sigma_i \in \Gamma_n} P(\Sigma_0, \omega_{1,N}^+, \Sigma_i)}. \]

By the definition of conditional probability and stochastic state class, we can express \( L_{\text{off}}(a, t, \Sigma_0, \omega_{1,N}^+) \) first as:

\[ L_{\text{off}}(a, t, \Sigma_0, \omega_{1,N}^+) = \sum_{\Sigma_n^m \in \Gamma_n} \frac{P(\alpha(t) = a, \Sigma_0, \omega_{1,n}^+, \Sigma_n^m) p(\omega_{n+1,N}^+ | \Sigma_n^m) P(\Sigma_n^m)}{\sum_{\Sigma_i \in \Gamma_n} P(\Sigma_0, \omega_{1,N}^+, \Sigma_i) P(\omega_{n+1,N}^+ | \Sigma_i) P(\Sigma_i)}. \]

and then as:

\[ L_{\text{off}}(a, t, \Sigma_0, \omega_{1,N}^+) = \sum_{\Sigma_n^m \in \Gamma_n} \frac{P(\alpha(t) = a | \Sigma_n^m) P(\Sigma_0, \omega_{1,n}^+, \Sigma_n^m) P(\omega_{n+1,N}^+ | \Sigma_n^m)}{\sum_{\Sigma_i \in \Gamma_n} P(\Sigma_0, \omega_{1,n}^+, \Sigma_i) P(\omega_{n+1,N}^+ | \Sigma_i)}. \]

Therefore, according to Theorem 5, we finally obtain that \( L_{\text{off}}(a, t, \Sigma_0, \omega_{1,N}^+) = \sum_{\Sigma_n^m \in \Gamma_n} p(a, t, \Sigma_n^m) L_{\text{off}}(\Sigma_n^m). \) \( \square \)