Fluid approximation of CTMC with deterministic delays

Luca Bortolussi
Department of Mathematics and Geosciences,
University of Trieste, Italy.
CNR/ISTI, Pisa, Italy.
Email: luca@dmi.units.it

Jane Hillston
Laboratory for the Foundations of Computer Science,
School of Informatics, University of Edinburgh, UK.
Email: jane.hillston@ed.ac.uk

Abstract—We compare population models in terms of Continuous Time Markov Chains with embedded deterministic delays (delayed CTMC), in which an exponential timed transition can only update the state of the system after a deterministic delay, and delay differential equations (DDE). We prove a fluid approximation theorem, showing that, when the size of the population goes to infinity, the delayed CTMC converges to a solution of the DDE.

I. INTRODUCTION

Modelling approaches that incorporate deterministic delays have a long standing tradition in many disciplines, including biology [1], ecology [2], epidemics [3], and network protocol analysis [4]. Most of these approaches are based on Delay Differential Equations [2], which are (functional) differential equations in which the vector field may depend on the values of the system variables in the past.

However, there is a growing interest within computational system biology for stochastic models in which deterministic delays follow exponential transitions [5], [6]. Doing this in a stochastic setting can be relevant to understand the dynamics of biochemical networks where few molecules are involved, as is the case for genetic networks in a single cell [5].

The inclusion of deterministic delays in Markovian models make them non-Markovian, and such models have previously been used in the Performance Analysis community in various forms, e.g. [7], [8]. However, here we will consider models in which a deterministic delay always follows an exponential transition. Such paired delays can be taken to represent an event in which the stochastic occurrence of the event is separated in time from the consequence. Thus, the state under goes two modifications: one when the event occurs, after an exponentially distributed delay, and the other when the consequences become apparent, here assumed to be a further deterministic delay. This use of delays can be seen as an abstraction from sequences of biological steps. For this class of models, specialised simulation algorithms have been defined in the context of biochemical systems [5], [6], taking into account different ways of interfacing deterministic delays with biochemical reactions.

Furthermore, a process algebra specifically designed to model with these kind of delays, Bio-PEPAd, has recently been established [9]. This language supports a semantics in terms of Delay Stochastic Simulation (DSSA) [5], Generalized Semi-Markov Processes (GSMP) [7], [10], and Delay Differential Equations (DDE) [2].

The relationship between DDE and delayed CTMC has attracted some attention: in [5] it has been shown that the equation of the average of the stochastic process with deterministic delays can be approximated by a DDE. In particular, this is the DDE associated with a Bio-PEPAd model. However, as far as we are aware, there has been no attempt to relate the delayed stochastic and fluid semantics from the point of view of fluid or mean field approximation [11], i.e. formally proving the convergence of the two approaches for large populations.

In this paper, we tackle this problem, providing a mean field convergence theorem of delayed CTMC to DDE.

Related work in terms of mean field convergence are fluid limits of processes (usually queue models) with generally distributed firing times [12], [13].

This paper is organised as follows. In Section II we provide the basic definitions needed in the following, introducing a CTMC modelling language, scaling assumptions and the classical fluid approximation theorem. In Section III, we introduce deterministic delays both at the syntactic level of the modelling language, and at the semantic level, defining its semantics both in terms of GSMP and in terms of DDE. Section IV-C contains the main result of the paper, namely the fluid convergence theorem. Finally, in Section V we draw final conclusions.

II. PRELIMINARIES

In this section we introduce some background material that is needed in the following. First of all, we introduce a very simple modeling language to describe CTMC, similar to the one used by PRISM [14]. Then, we will define sequences of models, depending on a parameter describing the size of the system, usually the total size of the population of agents, and state clearly the scaling assumptions that we require on this sequence. Under these assumptions, we will present the standard deterministic limit theorem.

A. Modelling Language

In the following, we will describe a basic language for CTMC. Essentially, we have in mind models in which entities
Definition II.1. A population CTMC model is a tuple \( \mathcal{X} = (X, D, T, x_0) \), where:
1. \( X = (X_1, \ldots, X_n) \) is a vector of variables.
2. Each \( X_i \) takes values in a finite or countable domain \( D_i \subset \mathbb{R} \). Usually, but not necessarily, \( D_i \) is a subset of positive integers. Hence, \( D = \prod_i D_i \) is the state space of the model.
3. \( x_0 \in D \) is the initial state of the model.
4. \( T = \{ t_1, \ldots, t_m \} \) is the set of transitions, defined by a rate, an update vector (the net change in variable values after the transition), and a guard (specifying when the transition is active).

More precisely, the infinitesimal generator matrix \( \mathcal{Q} \) of the CTMC is the set of tuples of values that
interact through a finite set of possible actions. We will describe such systems using a finite set of variables to count how many entities of each kind are in the system, so that the state space of the CTMC is the set of tuples of values that these variables can take. The actions that can be performed are described by a set of transitions, defined by a rate, an update vector (the net change in variable values after the transition), and a guard (specifying when the transition is active). Formally:

Example. Consider a model \( \mathcal{X} \) for a simple genetic network in our modelling language, we need two properties explored in detail, both for models without delays and with delays. We introduce now the main running example of the paper: we will consider a model of a simple genetic network such a network in our modelling language, we need two properties explored in detail, both for models without delays and with delays. We introduce now the main running example of the paper: we will consider a model of a simple genetic network.

We will consider a model of a simple genetic network.

We introduce now the main running example of the paper: we will consider a model of a simple genetic network.

We introduce now the main running example of the paper: we will consider a model of a simple genetic network. The condition on the rate function \( r_{i}^{(N)} \) is known in literature as the density dependence condition. In accordance with the previous subsection, we will denote the state of the CTMC of the \( N \)-th non-normalized (resp. normalized) model at time \( t \) as \( X^{(N)}(t) \) (resp. \( X^{(N)}(t) \)).
we need to introduce a few more concepts in order to construct C. Deterministic limit theorem

However, the two scalings are equivalent for the purpose of the scaled variables represent molar concentration of species. Given the scaling introduced in the previous section fact, it is easy to check that the conditions of the theorem are satisfied, given the scaling introduced in the previous section. In Figure 1, we compare a trajectory of the fluid ODE with sustained oscillations. The model does not oscillate for any value of the parameter space presented in Section II.1 to describe delayed transitions in the sense of [5], [9]. We will allow purely exponential transitions and exponential transitions followed by a deterministic delay. After extending the modelling language, we will define Generalised Semi-Markov Processes (GSMP) and show how we can associate a GSMP with a model with delays. At the end of the section, we introduce Delay-Differential Equations (DDE) and present a semantics of the delayed model in terms of DDE. Both conversion to GSMP and to DDE have been defined also for Bio-PEPAd [9], although our modelling language is more general since we allow a mixture of delayed and purely exponential transitions.

A. Stochastic Model with Delays

The modelling language presented in the previous section is easily extended to include delayed transitions.

Definition III.1. A delayed population CTMC model is a tuple $X_d = (X, D, T_e, T_d, x_0)$, where:

1) $X$, $D$, and $x_0$ are as in Def. II.1.
2) $T_e$ is a set of exponential transitions, defined as in Def. II.1.
3) $T_d$ is a set of delayed transitions, which are tuples $\rho = (v_\rho, r_\rho(X), w_\rho, \sigma_\rho)$, where $v_\rho$ and $r_\rho(X)$ are defined as for exponential transitions, and in addition:
   a) $w_\rho \in \mathbb{R}^n$, is the post-delay update vector.
   b) $\sigma_\rho \in \mathbb{R}_{>0}$ is the constant delay.

The semantics of delayed transitions is the following: the transition is initially governed by an exponential distribution with rate $r_\rho(X)$. When the transition fires in state $x$ at time $t$, it updates the variables according to the update vector, i.e. from $x$ to $x + v_\rho$, and schedules a delayed update after $\sigma_\rho$ units of time. At time $t + \sigma_\rho$, the variables are again updated from their current value $y$ to $y + w_\rho$. 

III. INTRODUCING DELAYS

In this section, we will extend the modelling language presented in Section II.1 to describe delayed transitions in the sense of [5], [9]. We will allow purely exponential transitions and exponential transitions followed by a deterministic delay. After extending the modelling language, we will define Generalised Semi-Markov Processes (GSMP) and show how we can associate a GSMP with a model with delays. At the end of the
We stress how the fact that deterministic delays are always guarded by exponential transitions and cannot be interrupted is crucial for the following analysis, as it introduces a controlled memory in the system, see also [7]. We further notice that guards could have been added both for the exponential transition and for the delayed one, but this would have introduced discontinuities in the system, see [23], [24] for a treatment of the purely exponential case.

Example. Consider again the running example introduced in Section II-A. As noted in Section II-C, the gene network without delays is not able to show oscillatory behaviour. However, the introduction of delays changes the picture, and oscillations can be produced both by stochastic and by deterministic models [5], [16]. From a biological point of view, the introduction of deterministic delays is a reasonable extension of the model, to take into account the fact that both transcription and translation are not atomic processes, but are composed of many complex steps. Abstracting such a sequence of steps with a deterministic delay, following an exponentially distributed trigger, is a more reasonable policy than simply abstracting them as by an exponential step [5].

In our language, we can easily introduce delays by replacing transcription and translation by delayed transitions:

1. \((0, 0), \sigma M, (1, 0), \sigma M\): delayed transcription;
2. \((0, 0), \alpha P M, (0, 1), \sigma P\): delayed translation.

Delayed CTMC models can be simulated by an extension of the Stochastic Simulation Algorithm (SSA) of Gillespie [25], called Delayed SSA (DSSA) [5]. Essentially, the simulation works by storing the scheduled delayed events in a priority queue and comparing at each simulation step the firing time of the next scheduled delayed event, if any, with the next firing time of an exponential event, sampled in the SSA-style. Then, the event first occurring is executed.

Going back to the running example, in Figure 2(a) we show a simulated trace of the evolution of protein and mRNA. As we can see, contrary to the non-delayed model, we observe sustained oscillations.

Given a sequence of delayed models \( \mathcal{X}^{(N)} \), depending on the size parameter \( N \), their associated normalized models \( \mathcal{X}^{(N)} \) are defined along the lines of Section II-B, with the delayed update scaled in the same way as the normal update. The delay duration, instead, remains unchanged. In particular, we require density dependence of rates.

B. Generalized Semi-Markov Processes

A Generalized Semi-Markov Processes (GSMP) [7], [10] is a general model of a non-Markovian process, allowing transitions to fire at generally distributed firing times. Here we consider a variant of the standard definition, see Remark III.1 for a discussion.

Definition III.2. A Generalized Semi-Markov Processes is a tuple \( \mathcal{G} = (S, s_0, E, r, p, F) \), where:

1) \( S \) is a countable set of states and \( s_0 \in S \) is the initial state.
2) \( E \) is a countable set of events, and \( E : S \rightarrow \varphi(E) \) associates with each state \( s \) a subset of active events.
3) With each event \( e \in E \), we associate a clock taking values in the non-negative reals. The space of clock valuations is thus \( C = \mathbb{R}_+^\infty = \mathbb{R}_+ \times \mathbb{R}_+ \times \ldots \), a clock valuation is denoted by \( c \in C \), and the value for the clock associated with \( e \) is \( c_e \).
4) \( r = S \times E \rightarrow \mathbb{R}_+ \) associates with each state and each event a rate, which is the rate at which the clock evolves. It satisfies \( r(s, e) = 0 \) if \( e \notin E(s) \), and \( r(s, e) > 0 \) for at least one \( e \in E(s) \).
5) \( p(s', s, e) \) is the transition probability of jumping to state \( s' \) from state \( s \), when event \( e \in E(s) \) fired.
6) \( F(\cdot, e) \) is the cumulative distribution function (cdf) of the clock associated with event \( e \in E \).

The dynamics of a GSMP process is informally defined in the following way. The initial state \( s_0 \) and the initial valuation of clocks is defined by sampling each associated cdf. The transitions of the process are then executed according to

\note{2Note that this translation transition assumes that mRNA molecules remain available when acting as a template to produce a protein. A more precise formulation of translation is therefore the following: \(((-1, 0), \alpha P M, (1, 1), \sigma P)\), in which mRNA is unavailable for the duration of the delay. However, we will stick to the former version for simplicity.}

\note{3A general distribution over the state space can also be considered.}
the following policy. For each state \( s \in S \) and clock valuation \( c \), define the following quantities:

\[ t^e = t^e(s, c) = \inf \{ t \geq 0 \mid \min_{e \in E(s)} \{ c_e - t \cdot r(s, e) \} \geq 0 \} \]
\[ c^*_e = c^*_e(s, c) = c_e - t^e(s, c) r(s, e) \]
\[ e^* = e^*(s, c) = \min \{ e \in E(s) \mid c^*_e(s, c) = 0 \} \] (assuming an ordering between events).

The event \( e^* \) is the event that fires in state \( s \), where \( c \) represents the current clock valuation. \( t^e \), instead, is the time elapsed in state \( s \). After the firing of event \( e^* \), the state is changed from \( s \) to \( s' \) with probability \( p(s', s, e^*) \). Clock valuation \( c \), instead, is changed as follows:

- The clocks of all events \( e \in R(s', s, e^*) \) are updated by sampling from their cdf \( F(\cdot, e) \), where \( R(s', s, e^*) = \{ e^* \} \cup (E(s) \setminus E(s')) \) is the set of disabled events plus the event that fired.
- The clocks of all events \( e \in W(s', s, e^*) \) are modified to take into account the elapsed time, i.e. they are set to the value \( c^*_e \). \( W(s', s, e^*) = E(s') \cap (E(s) \setminus \{ e^* \}) \) is the set of all events remaining active after the transition.

Remark III.1. The definition of GSMP presented above is slightly different from the standard one. In particular, we assumed that the cdf of each clock is independent both of the transition that fired and of the current state. This allows us to assume a different update policy of clocks: Instead of resetting a clock when the associated event becomes active, as customary, we reset the clock only when the associated event fires or when it is deactivated. As the cdf is independent of the enabling transition and of the current state, this is equivalent to resetting them once they become active. This modification is introduced here in order to simplify the following proofs.

We are now ready to define the GSMP associated with a model \( \mathcal{X}_d \) with delays.

Definition III.3. Let \( \mathcal{X}_d = (X, D, \mathcal{T}_d, \mathcal{T}_d, x_0) \) be a delayed CTMC model. The GSMP \( \mathcal{G} = \mathcal{G}(\mathcal{X}_d) = (S, s_0, E, E', r, p, F) \) associated with \( \mathcal{X}_d \) satisfies:

- The set of events \( E \) contains an event \( e_\tau \) for each exponential transition \( \tau \in \mathcal{T}_e \), with a cdf exponentially distributed with rate 1, \( F(\cdot, e) \sim \text{Exp}(1) \). Moreover, for each delayed transition \( \rho \in \mathcal{T}_d \) there is one event \( e_\rho \) with a cdf exponentially distributed with rate 1, and an infinite set of events \( e_{\rho,j} \), \( j \geq 1 \), deterministically distributed with cdf \( F(\cdot, e_{\rho,j}) = \delta_{\tau_\rho} \). We need an infinite set of events because there is no upper bound on the number of delayed events that can be scheduled for transition \( \rho \).
- In the state space \( S \), we need to keep track not only of the value of variables, but also of the delayed events scheduled to fire, as they cannot be inferred from the current state in \( D \). Hence, \( S = D \times \mathcal{Y}_f \), so that each state \( s = (x, A) \) contains a point in the state space of \( \mathcal{X}_d \) and the subset \( A \) of active events in it, i.e. \( E(s) = A \). Notice that we assume that only a finite set of events can be active in each state. In particular, we will consider only states \( s = (x, A) \) for which \( A \) contains the set \( E_0 = \{ e_\rho \mid \rho \in \mathcal{T}_d \} \cup \{ e_\tau \mid \tau \in \mathcal{T}_e \} \).
- The initial state is \( s_0 = (x_0, E_0) \).
- The rate function \( r \) in state \( s = (x, A) \) is such that \( r(s, e_\rho) = r_\rho(x) \), \( r(s, e_\tau) = r_\tau(x) \), and \( r(s, e_{\rho,j}) = 1 \) if \( e_{\rho,j} \in A \), and 0 otherwise. Since \( E_0 \subseteq A \), this definition is consistent.
- The transition probability \( p \) in state \( s = (x, A) \) is defined as follows: \( p(s', s, e_\tau) = 1 \) if \( s' = (x + v_\tau, A) \) and 0 otherwise; \( p(s', s, e_\rho) = 1 \) if \( s' = (x + w_\rho, A \setminus \{ e_{\rho,j} \}) \) and 0 otherwise.

In the previous definition, we reset the exponential clocks only after they have fired. Then, the clocks are updated using different clock speeds in each state, equal to the rate of the exponential distribution in that state. This mechanism is essentially the one used to define the Gibson-Bruck stochastic simulation algorithm [26], also known as next reaction method, and it is equivalent to resetting each exponential clock after each jump.

C. Delay Differential Equations

Delay Differential Equations (DDE) [2] are differential equations in which the derivative can depend also on past values of the function. More precisely, a DDE is a functional differential equation of the form

\[ \frac{dx(t)}{dt} = F(t, x_t), \]

where \( x_t \) is the function from \([-d, 0]\) to \( \mathbb{R}^n \), defined by \( x_t(s) = x(t + s) \), for \( s \in [-d, 0] \). In a DDE, the right hand side is in fact a functional from the Banach space \( \Omega = C([-d, 0], \mathbb{R}^n) \times \mathbb{R} \) to \( \mathbb{R}^n \), associating with each function from \([-d, 0] \) to \( \mathbb{R}^n \), and each time \( t \), a vector in \( \mathbb{R}^n \). Notice that the initial condition of a DDE is no longer a single point, but rather a function \( \varphi : [-d, 0] \to \mathbb{R}^n \), specifying the behaviour up to \( d \) units of time in the past. Existence and uniqueness theorems for DDE reflect those of ODE: if \( F \) is Lipschitz continuous, then the DDE has a unique solution, for any initial function \( \varphi : [-d, 0] \to \mathbb{R}^n \).

In this paper, we will be interested in DDE with constant delays, i.e. equations of the form

\[ \frac{dx(t)}{dt} = F(t, x(t), x(t - \tau_1), \ldots, x(t - \tau_n)). \]

We consider now a mapping from a delayed CTMC model to a set of DDE:

Definition III.4. Let \( \mathcal{X}_d = (X, D, \mathcal{T}_d, \mathcal{T}_d, x_0) \) be a delayed CTMC model. The set of DDE with constant delays associated with \( \mathcal{X}_d \) is defined as:

\[ \frac{dx(t)}{dt} = \sum_{\tau \in \mathcal{T}_e} v_\tau r_\tau(X(t)) + \sum_{\rho \in \mathcal{T}_d} v_\rho r_\rho(X(t)) + \sum_{\rho \in \mathcal{T}_d} w_\rho r_\rho(X(t - \sigma_\rho)). \]
As can be seen from the previous formula, the delayed transitions give rise to two different kinds of flows, one non-delayed (corresponding to the exponential step) and one delayed (corresponding to the delay step).

If we write down the DDEs associated with the running example, we obtain the following set of equations.

\[
\begin{align*}
\frac{dm(t)}{dt} &= \frac{\alpha_m(t - \sigma_m)}{1 + (t - \sigma_m)/\rho_m} - \beta_m m(t) \\
\frac{dp(t)}{dt} &= \alpha_p(t - \sigma_p) - \beta_p p(t)
\end{align*}
\]

A solution of these equations for the initial function \( \varphi \equiv 0 \) is shown in Figure 2(b). Comparing it with Figure 2(a), we observe a similar behaviour. In Figure 3, we show different stochastic trajectories for increasing system sizes. As can be observed, the stochastic trajectory seems to behave deterministically already for \( N = 100 \). Furthermore, it looks almost identical to the DDE solution, if we ignore the initial lag for the CTMC model. This behaviour suggests a possible fluid limit, connecting delayed CTMC models and DDE. This will be the topic of the next section.

IV. MEAN FIELD CONVERGENCE

In this section we will prove the main result of the paper, namely the convergence of the sequence of GSMP, associated with a sequence of delayed CTMC models depending on system size \( N \), to the DDE syntactically defined in Section III-C.

In order to do this, we will rely on two similar approximations of deterministic delays, both for GSMP and for DDE. In the former case, we will replace a deterministic delay \( \sigma \) with an Erlang distribution with \( k \) stages and rate \( k/\sigma \), for increasing \( k \). As the Erlang distribution is a special case of a phase type distribution, we can expand the GSMP into an appropriate Markov Chain. In the latter case, we will approximate the deterministic delay by introducing a set of \( k \) variables [27], governed by a vector field that turns out to be (equivalent to) the drift of the Markov Chain obtained in the previous step. Then, we will obtain the convergence result by applying the standard fluid limit of Kurtz [21] to these two approximations. However, some care is needed, as the Erlang approximation increases the dimension of the state space in both the CTMC and the ODE approximations, affecting the speed of convergence, so that we need to increase the system size and the stages of the Erlang distribution in a coordinated and controlled way. Furthermore, we need also to deal appropriately with the initial conditions of the DDE, that are given in terms of a function. In contrast, the delayed CTMC has initial conditions defined in terms of a single point. In the following, we will also show how we can incrementally construct the initial solution of the DDE, by taking into account only a subset of delays.

In the rest of this section, we will first tackle the approximation of a GSMP by replacing deterministic delays with Erlang distributions, tailoring the convergence result to the case in which we increase both the stages of the Erlang distribution, \( k \), and the system size \( N \). Then, we will turn our attention to the Erlang approximation of DDE, recalling the convergence result of [27]. Finally, we will compare these two approximations and prove the fluid limit theorem for CTMC with delays, showing also how to construct the initial condition of the DDE.

A. Erlang approximation of delays for GSMP

We start this section by recalling the basic properties of the Erlang distribution. A random variable \( Z \) with values in the positive real numbers has an Erlang distribution with \( k \in \mathbb{N} \) stages and rate \( \lambda \in \mathbb{R}_+ \), \( Z \sim \text{Erlang}(k, \lambda) \), if \( Z = \sum_{i=1}^{k} Y_i \), where each \( Y_i \) is exponentially distributed with rate \( \lambda \). The density function of the Erlang distribution is \( f_Z(x, k, \lambda) = \lambda^k \frac{x^{k-1} e^{-\lambda x}}{(k-1)!} \), while its cumulative distribution function is \( \Phi_Z(x, k, \lambda) = 1 - \sum_{n=0}^{k-1} \frac{e^{-\lambda x} (\lambda x)^n}{n!} \). In particular, the sequence \( Z_k \sim \text{Erlang}(k, \lambda k) \) converges to the deterministic value \( \frac{\lambda}{\lambda} \) (this follows from the Law of Large Numbers applied to the sum of exponentials). Therefore, we can approximate a deterministic delay \( \sigma \) with an Erlang random variable \( Z_k \sim \text{Erlang}(k, \lambda k/\sigma) \), for large \( k \).

Let \( \mathcal{G} \) be a delayed CTMC model, and let \( \mathcal{G} = \mathcal{G}(\mathcal{X}_d) \) be the GSMP associated with it according to Definition III.2. Fix an integer \( k > 0 \), and let \( \mathcal{G}_k \) be the GSMP which is defined like \( \mathcal{G} \), but with the cdf for each event \( e_{\rho,j}, \rho \in T_d \) equal to \( F(x, e_{\rho,j}) = F(x, k, k/\sigma) \). The sequence of GSMPs \( \mathcal{G}_k \) approximates \( \mathcal{G} \) in the following sense:

**Lemma IV.1.** Let \( \mathcal{G} = \mathcal{G}(\mathcal{X}_d) \) for a delayed CTMC model \( \mathcal{X}_d \) and \( \mathcal{G}_k \) be defined as above. Then \( \mathcal{G}_k \) converges weakly (i.e., in distribution) to \( \mathcal{G} \): \( \mathcal{G}_k \Rightarrow \mathcal{G} \).

**Proof:** The proof is a simple adaptation of the proof of Theorem I of [28], in which a general weak convergence
result is proved for GSMP with continuous cdf. The key point is that, as deterministic delayed events always follow an exponentially distributed event, the probability that two events fire at the same time is zero. Details are provided in [29].

Consider now a sequence of delayed CTMC models \( X_d^{(N)} \), depending on the size parameter \( N \). We want to understand what happens when we approximate the GSMP \( G^{(N)} \) by GSMP \( G_{kN}^{(N)} \) with \( kN \)-stage Erlang delays, and then take the limit for \( N \to \infty \). In particular, we will (only) assume that \( kN \to \infty \) as \( N \) increases.

**Lemma IV.2.** Let \( G^{(N)} = G(X_d^{(N)}) \) be the sequence of GSMP associated with a sequence of delayed CTMC models, for increasing system size \( N \). Let \( kN \) be a sequence of integers, such that \( kN \to \infty \) as \( N \to \infty \). Finally, let \( G_k^{(N)} \) be the GSMP with Erlang delays of \( kN \) stages replacing deterministic ones. Then, \( G_k^{(N)} \Rightarrow G^{(N)} \) as \( N \to \infty \).

**Proof:** The proof is an adaptation of that of Theorem I in [28], to the case in which the limit process is not fixed, but varies with \( N \). The key steps are to define the transition kernels of the embedded Discrete Time Markov Processes, and to show that these kernels satisfy the conditions of Theorem I of [30], adapted to the case of a variable limit. Details are provided in [29].

Each Erlang distribution with \( k \) stages and rate \( \lambda \) can be clearly represented as a CTMC with \( k + 1 \) states \( S = \{0, 1, \ldots, k\} \) and \( k \) transitions, jumping from state \( i \) to \( i + 1 \) (\( i = 0, \ldots, k - 1 \)) with the same rate \( \lambda \). Owing to this fact, we can convert the GSMP \( G = G(X_d) \) into a CTMC with countable state space. We find it convenient to define this transformation at the level of the modelling language, associating with each delayed model \( X_d \), a non-delayed model \( X_k \), approximating deterministic delays with stage \( k \) Erlangs.

**Definition IV.1.** Let \( X_d = (X, D, \mathcal{T}_e, \mathcal{T}_d, x_0) \) be a delayed model and \( k > 0 \). The corresponding approximating CTMC model \( X_k = (Y, D_k, \mathcal{T}_k, y_0) \) is defined in the following way:

1. \( Y = X \cup \{Z_{\rho, j} \mid \rho \in \mathcal{T}_d, j = 1, \ldots, k\} \) is the set of variables. Each delay variable \( Z_{\rho, j} \) represents one state in the Erlang expansion for the delay of transition \( \rho \) (excluding the absorbing state). We let \( \pi_X(y) \) be the projection of \( y \) on \( X \) coordinates and, for \( x \in D, \pi_Y(x) \) be the point \( y \in D_k \) equal to zero for all coordinates not in \( X \).
2. The domain is \( D_k = D \times \mathbb{N}^{k|\mathcal{T}_d|} \), i.e. each variable \( Z_{\rho, j} \) takes values in the natural numbers.
3. The initial state \( y_0 \) is \( y_0 = \pi_Y(x_0) \), i.e. it coincides with \( x_0 \) on variables \( X \), and equals 0 for the other variables.
4. The set of transitions, \( \mathcal{T}_k \), contains a transition \( (\pi_Y(v), r(\pi_X(Y))) \) for each exponential transition \( \tau \in \mathcal{T}_e, \tau = (v, r(X)) \). In addition, for each \( \rho \in \mathcal{T}_d, \rho = (v, r(X), w, \sigma) \), \( \mathcal{T}_k \) contains the transitions \( (\pi_Y(v) + 1_{Z_{\rho, 1}}, r(\pi_X(Y))) \), where \( 1_{Z_{\rho, 1}} \) is the vector equal to 1 for coordinate \( Z_{\rho, 1} \) and zero elsewhere, \( (1_{Z_{\rho, 1}} - 1_{Z_{\rho, 1}}, k/\sigma) \), for \( i = 1, \ldots, k - 1 \), and \( (1_{Y}(w) - 1_{Z_{\rho, k}}, k/\sigma) \).

In Definition IV.1 we replaced a delayed transition \( \rho \) by an exponential transition followed by a sequence of \( k \) transitions with rate \( k/\sigma \), encoding the \( k \) stages of the CTMC representation of an Erlang distribution. Thus it follows that

**Proposition IV.1.** The CTMC defined by \( X_k \) is equivalent to the GSMP \( G_k \).

**Example.** Consider again the running example. If we want to approximate the delayed CTMC model with 2-step Erlang delays, we need to introduce two new variables for transcription, \( Z_{1,1} \) and \( Z_{1,2} \), and two new variables for transition, \( Z_{2,1} \) and \( Z_{2,2} \). Then, we obtain the following set of transitions replacing transcription: \( (1_{Z_{1,1}}, \alpha_M + 1/(\sigma_M)), (1_{Z_{1,2}} - 1_{Z_{1,1}}, 2/\sigma_M) \), and \( (1_{M} - 1_{Z_{2,1}}, 2/\sigma_M) \). Delayed translation is replaced by a similar set of transitions.

**B. Erlang approximation of delays for DDE**

In this section, we discuss a method to approximate a set of DDE with constant delays by a set of ODE. The approximation works by introducing an additional set of variables for each delayed term, which essentially corresponds to the Erlang approximation discussed in the previous section.

To have an idea of how the method works, consider the DDE \( \frac{dx(t)}{dt} = f(x(t - \sigma)) \). If we let \( z(t) = f(x(t) - \sigma) \), then obviously \( \frac{dz(t)}{dt} = z(t) \) and \( z(t + \sigma) = f(x(t)) \). Fix an integer \( k > 0 \), and introduce \( k \) new variables \( z_{1}, \ldots, z_{k} \), and let \( z_{j}(t + \sigma/k) = f(x(t)) \) and \( z_{j+1}(t + \sigma/k) = z_{j}(t) \), for \( j = 1, \ldots, k - 1 \). Clearly, \( z_{k}(t) = z(t) \). Now, by a first order Taylor expansion of \( z_{j+1}(t + \sigma/k) \), we find from the previous relation an approximate ODE for \( z_{j+1} \), namely

\[
\frac{dz_{j+1}(t)}{dt} = \frac{k}{\sigma} (z_{j}(t) - z_{j+1}(t)).
\]

In summary, we obtain the following system of ODE:

\[
\begin{align*}
\frac{dz_{1}(t)}{dt} &= \frac{k}{\sigma} (f(x(t)) - z_{1}(t)) \\
\vdots \\
\frac{dz_{j+1}(t)}{dt} &= \frac{k}{\sigma} (z_{j}(t) - z_{j+1}(t)) \\
\vdots \\
\frac{dz_{k}(t)}{dt} &= z_{k}(t)
\end{align*}
\]

Consider now a delayed CTMC model \( X_d \), whose associated DDE equals \( \frac{dx(t)}{dt} = f(x(t - \sigma)) \). It has one variable \( X \), and only one delayed transition: \( \rho = (0, f(X), 1, \sigma) \) (for simplicity we assume \( f \) is positive). If we apply the Erlang approximation with \( k \) steps, according to Def. IV.1, then we introduce \( k \) new variables \( Z_{\rho, j} = Z_{j} \), and \( k \) new transitions. If we construct the fluid approximation of this CTMC model, we get the following
Although these two sets of ODE are not the same, they are easily shown to be equivalent; just apply the substitution $Z_j = \frac{2}{k} z_j$ and $X = x$.

Consider now a generic delayed CTMC model $\mathcal{X}_d$, and construct the associated DDE according to Def. III.4, with initial conditions of the DDE.

Let us deduce that $\mathcal{X}_d$ is independent of $N$, by virtue of the scaling assumptions. The latter, instead, can be obtained by applying the construction of Def. IV.1 to any model $\mathcal{X}_d^{(N)}$, and then constructing the fluid ODE from this model, according to Sec. II-C. Also in this case, by virtue of the scaling assumption and on the linearity of the rates of the added transitions, the system of ODE obtained is independent of $N$ (although dependent on the steps of the Erlang approximation). We will denote by $\mathcal{Y}_k(t)$ the solution of the set of ODE approximating the DDE with $k$ steps, with $\mathcal{X}_k(t)$ their projection on the variables of $\mathcal{X}_d^{(N)}$, and with $x(t)$ the solution of the DDE. As in the fluid case, we consider a subset $E \subseteq \mathbb{R}^n$ such that the solution of the DDE (and all approximating ODE, when projected in $E$) are defined in $E$ for all $t \geq 0$.

1) The problem of initial conditions: Differently from ODE, the initial conditions of $\mathcal{X}_d$ are given by a function in the time interval $[t_0 - \sigma_M, t_0]$, where $\sigma_M$ is the largest delay. On the contrary, the delayed CTMC model starts from a point $x_0$, and delays start to have an effect not at a time $t_0$, as for DDE, but at time $t_0 + \sigma_1$, where $\sigma_1$ is the smallest delay. The full effect of delays is sensed only after time $t_0 + \sigma_M$. This immediately raises two problems. Firstly the delayed CTMC and the DDE are out of phase, and should be synchronised by comparing time $t$ of the DDE solution with time $t + \sigma_M$ of the delayed CTMC. The second, more delicate problem is how to define the initial conditions of the DDE in such a way that they capture the limit dynamics of the delayed CTMC in the initial transient phase, i.e. in the time interval $[0, \sigma_M]$. Note that this problem has repercussions for the definition of the initial conditions of the Erlang approximation of the DDE. Indeed, we need to fix the initial values of the step variables propagating the delay according to the function defining the initial conditions of the DDE.

In contrast, the CTMC approximation of the delayed CTMC does not suffer from this problem, since the newly added variables are set to zero, corresponding to the fact that the effects of delays take place in the future and there is no effect.
from the past. Therefore, we should consider the solution of the DDE starting not from time \( t_0 \), but from time \( t_0 + \sigma_M \), and construct the initial condition for the DDE from the behaviour of the delayed CTMC in \([t_0, t_0 + \sigma_M]\).

In order to fix the notation, let \( \sigma_1, \ldots, \sigma_M \) be the different delays appearing in the delayed CTMC, sorted in ascending order. Also let \( \sigma_0 = 0 \) and \( \sigma_{M+1} = +\infty \). The intuition is the following. In the time interval \([0, \sigma_1]\), there is no delay in force in \( X^{(N)} \), hence we can ignore all delays, obtaining a standard CTMC, from which we can construct the usual fluid limit in terms of ODE. In the interval \([\sigma_1, \sigma_2]\), only the delays of duration \( \sigma_1 \) are in effect, hence we can restrict \( X^{(N)} \) only to transitions without delays and to transitions with delays equal to \( \sigma_1 \). Then we prove convergence to the corresponding DDE, with initial conditions defined by the solution of the fluid ODE up to time \( \sigma_1 \). In this way, we can extend convergence up to time \( \sigma_2 \). Iterating the same procedure, we can extend convergence up to time \( \sigma_M \), from which all delays are in force.

Then, we have constructed the initial solution for the limit DDE, and we can now prove convergence up to any finite time horizon \( T > \sigma_M \).

In order to properly formalize this procedure, we need the following definition

**Definition IV.2.** Let \( X_d = (X, D, T_e, T_d, x_0) \) be a delayed CTMC and \( X_k = (Y, D_k, T_k, y_0) \) be its approximating CTMC with \( k \)-step Erlang delays. Let \( 0 = \sigma_0, \sigma_1, \ldots, \sigma_M \) be the different delays of \( X_d \) in ascending order. For \( m \leq M \), we define the \( m \)-reduced delayed CTMC \( X^m_d = (X, D, T_e, T_d, x_0) \) replacing each transition \( \rho = (\text{true}, v, r, (X)) \), true, \( w, \rho, (X) \) with \( \sigma_0 > \sigma_m \), by the exponential transition \( \tau_\rho = (\text{true}, v, r, (X)) \).

The \( m \)-reduced \( k \)-step Erlang approximating CTMC \( X^m_k \), instead, is defined by replacing each final transition in the \( k \)-step Erlang approximation of a delay, i.e. \((w_\rho - 1_{Z_{\rho,k}}, -\frac{1}{\sigma_\rho} Z_{\rho,k})\), for each \( \sigma_0 > \sigma_m \), by \((-1_{Z_{\rho,k}}, -\frac{1}{\sigma_\rho} Z_{\rho,k})\) (i.e. by removing its effect on \( X \) variables).

Note that, in the above definition, \( X^0_d = X_d \) and \( X^0_k = X_k \), while \( X^0_d \) is a model without any delayed transition, hence a CTMC.

Given the \( m \)-reduced models \( X^m_d \) and \( X^m_k \), we denote by \( F^m \) and \( F^m_k \) the DDE and ODE vector fields associated with them. Furthermore, we define \( x^m(t) \) inductively as the solution of the DDE \( \frac{dx^m}{dt} = F^m(x^m(t), x^m(t-\sigma_0), \ldots, x^m(t-\sigma_m)) \), for \( t \geq t_0 = \sigma_m \), with initial function \( x^m(t) \) in \([0, \sigma_m]\), for \( m \geq 1 \); whereas \( x^0(t) \) is the solution of the ODE \( \frac{dx^0}{dt} = F^0(x^0(t)) \) and initial conditions \( x^0(0) = x_0 \) in \([0, \sigma_1]\).

**D. Proof of convergence**

In order to prove convergence, we first show the following

**Lemma IV.4.** Suppose \( x^{(N)}_0 \rightarrow x_0 \). There exists a divergent and non-decreasing sequence \( k_N \) such that, for each \( t \geq 0 \),

\[
\|X^{(N)}_{k_N}(t) - x(t)\| \rightarrow 0 \text{ in probability},
\]

where \( x(t) = x^m(t), \) for \( t \in [\sigma_m, \sigma_{m+1}] \) and \( x(t) = x^M(t), \) for \( t \geq \sigma_M \).

**Proof:** We prove the result by induction with \( t \in [\sigma_0, \sigma_1] \) as the base case (0 delays), and with \( t \in [\sigma_m, \sigma_{m+1}] \) as the inductive steps, assuming the result holds up to time \( \sigma_m \).

The base case is essentially an application of Theorem II.1, as there are no delays. However, care must be taken to properly choose the sequence \( k_N \), as the increased dimension of the state space of \( y_{k_N}(t) \) has an impact on the speed of convergence of \( Y^{(N)}_{k_N}(t) \) to \( y_{k_N}(t) \). In the inductive step, the key point is proving that the induction hypothesis guarantees that the initial conditions \( z_{p,j}(\sigma_m) \) of the Erlang variables of \( y_{k_N}(t) \) converge to the values needed to apply Lemma IV.3. The full proof can be found in [29].

We are now ready to prove the main result of the paper.

**Theorem IV.1.** Let \( X^{(N)} \) be a sequence of delayed CTMC models, satisfying the scaling assumptions. Suppose \( x^{(N)}_0 \rightarrow x_0 \), and let \( x(t) = x^m(t), \) for \( t \in [\sigma_m, \sigma_{m+1}] \) and \( x(t) = x^M(t), \) for \( t \geq \sigma_M \) be the solution of the DDE associated with the sequence \( X^{(N)} \). Then, for each \( T \geq 0 \)

\[
\lim_{N \to \infty} \sup_{0 \leq t \leq T} \|X^{(N)}(t) - x(t)\| = 0 \quad \text{in probability.}
\]

**Proof:** From Lemma IV.4, we know that, for each \( t \geq 0 \), \( \|X^{(N)}_{k_N}(t) - x(t)\| \rightarrow 0 \) in probability. Applying Theorem 3, pg. 92 of [31], we can deduce that \( X^{(N)}_{k_N} \rightarrow x \) weakly. From Lemma IV.2, we know that \( X^{(N)}_{k_N} \rightarrow X^{(N)} \). Combining these two facts, we can therefore show that \( X^{(N)} \rightarrow x \) [32]. But weak convergence to a deterministic limit is equivalent to uniform convergence in probability for any finite time horizon (cf. [32]).

**Example.** Returning to the example, we can visually observe convergence in Figure 4. As for the construction of the initial condition, we can easily see that both protein and mRNA quantity will remain equal to zero until the production of a first molecule of mRNA, after time \( t = \sigma_M \). Hence, the initial solution of the DDE is, in this case, the constant function zero.

**V. Conclusions**

In this paper we considered the problem of relating the stochastic behaviour of CTMC population models with deterministic delays and models in terms of DDE. In particular, we proved a mean field result, showing the convergence of the delayed CTMC for increasing population size to a DDE that is defined syntactically at the level of the modelling language. We believe this result can be helpful for analysing such stochastic models exploiting fluid approximation as done for Markovian Stochastic Process Algebras. Indeed, the DDE limit has a simulation cost independent of the population size, while the computational cost of simulating a delayed population CTMC becomes prohibitive for large populations.

Error bounds will be an interesting extension of the theorem presented, although the standard error bounds for fluid
approximation, despite being interesting from a theoretical point of view, are of little help in practice, due to their doubly exponential dependence on the time horizon.

Other extensions of this work, from a theoretical perspective, include understanding the relationship between higher order approximation schemes for DDE, and the corresponding CTMC. In fact, this could suggest more effective phase-type approximations of deterministic delays.

REFERENCES


Fluid approximation of CTMC with deterministic delays
Supplementary online material

Luca Bortolussi
Department of Mathematics and Geosciences,
University of Trieste, Italy.
CNR/ISTI, Pisa, Italy.
Email: luca@dmi.units.it

Jane Hillston
Laboratory for the Foundations of Computer Science,
School of Informatics, University of Edinburgh, UK.
Email: jane.hillston@ed.ac.uk

Disclaimer: This document is a self-managed online supplementary material of the paper “Fluid approximation of CTMC with deterministic delays”, by L. Bortolussi and J. Hillston, published at QEST 2012. This is an updated version of the original document, fixing a problem with the proof of Lemma IV.2, which has been discovered after a discussion with Richard Hayden and Miklos Telek, which we thank for having drawn our attention on the Lemma’s proof.

I. PROOFS OF LEMMAS

Lemma (Lemma IV.1). Let \( G = G(\mathcal{X}_d) \) for a delayed CTMC model \( \mathcal{X}_d \) and \( G_k \) be defined as above. Then \( G_k \) converges weakly (i.e. in distribution) to \( G \): \( G_k \Rightarrow G \).

Proof: The proof is a simple adaptation of that of Theorem I of [1], in which a general weak convergence result is proved for GSMP with continuous cdf. In particular, that proof converts the GSMP into a Discrete Time Markov Process (DTMP), with state space enriched by clock valuations and the model time, and then uses Theorem I in [2] to conclude. The result in [2] itself relies on a continuity property of the transition kernel, specifically the expected value of a uniform continuous function after one jump, has to be continuous as a function of the state from which the chain jumps. In order to enforce this condition for GSMP, in [1] the author restricts the clock valuation set to the open subset of \( \mathbb{R}^n_+ \) in which only one event can fire (i.e., no two events can fire at the same time). In state \( s \in S \), therefore, the allowed set of clock valuations is

\[
C_s = \{ c \in \mathbb{R}^n_+ | c_{e_1} r(s, e_1)^{-1} \neq c_{e_2} r(s, e_2)^{-1}, \\
\text{for } e_1 \neq e_2, \text{ and } c_{e_1} r(s, e_1) c_{e_2} r(s, e_2) > 0 \}.
\]

The key point is that continuity of the cdf implies that the probability that two events fire at the same time is zero, hence the probability of the complementary set of \( C_s \), at any time \( t \), is always zero. In our setting, the cdf of deterministic delays is not continuous. However, we can always restrict to the set \( C_s \), because each deterministic delay fires after an exponentially distributed delay, so that the probability that any two events fire at the same time remains zero.

Lemma (Lemma IV.2). Let \( G^{(N)} = G(\mathcal{X}^{(N)}_d) \) be the sequence of GSMP associated with a sequence of delayed CTMC models, for increasing system size \( N \). Let \( k_N \) be a sequence of integers, such that \( k_N \to \infty \) as \( N \to \infty \). Finally, let \( G_k^{(N)} \) be the GSMP with Erlang delays of shape \( k_N \) replacing deterministic ones. Then, \( G_k^{(N)} \Rightarrow G^{(N)} \) as \( N \to \infty \).

Instead of proving this lemma, we show a weaker version, which is sufficient for our purposes:

Lemma 1.1. Let \( G^{(N)} = G(\mathcal{X}^{(N)}_d) \) be the sequence of GSMP associated with a sequence of delayed CTMC models, for increasing system size \( N \) and let \( \hat{X}^{(N)}(t) \) the process on normalised variables. Let \( k_N \) be a sequence of integers, such that \( k_N \to \infty \) as \( N \to \infty \), and \( k_N = o(N^{1+}) \). Finally, let \( G_k^{(N)} \) be the GSMP with Erlang delays of shape \( k_N \) replacing deterministic ones, and let \( X_k^{(N)}(t) \) be the corresponding process on normalised variables. Then, for any \( t \geq 0 \), \( X_k^{(N)}(t) \Rightarrow \hat{X}^{(N)}(t) \) as \( N \to \infty \).

Proof: We will use two Poisson representations of \( \hat{X}^{(N)}(t) \) and \( X_k^{(N)}(t) \). The representation for \( \hat{X}^{(N)}(t) \) is based on the equivalent description of the process as a CTMC introduced in Definition IV.2. Call this CTMC, on the extended state space, \( \hat{Z}^{(N)}(t) \), and the projection on the coordinates of the state space of the population process by \( \pi_x \). The Poisson representation of \( \hat{Z}^{(N)}(t) \) is obtained using a standard argument [3]. The Poisson representation for \( \hat{X}^{(N)}(t) \), instead, has been proved in [4]. Hence, we have:

\[
\hat{X}^{(N)}(t) = \hat{X}^{(N)}(0) + \sum_{\tau \in T_e} \frac{w_e}{N} \mathcal{Y}_\tau \left( N \int_0^t f^{(N)}_\tau (\hat{X}^{(N)}(s)) \, ds \right) + \sum_{p \in T_c} \frac{w_p}{N} \mathcal{Y}_p \left( N \int_0^t f^{(N)}_p (\hat{X}^{(N)}(s)) \, ds \right)\
+ \sum_{\rho \in T_c} \frac{w_\rho}{N} \mathcal{Y}_\rho \left( N \int_0^t f^{(N)}_\rho (\hat{X}^{(N)}(s)) \, ds \right)
\]
and

$$X^{(N)}_{k,N}(t) = \pi_x \left( \sum_{\eta \in T_d} v_\eta f^{(N)}_\eta (\pi_x Z^{(N)}(s))ds \right) + \frac{k_N}{N} \sum_{\rho \in T_\eta} N \rho \left( N \int_0^t f^{(N)}_\rho (\pi_x Z^{(N)}(s))ds \right) + \frac{k_N}{N} \sum_{\rho \in T_\eta} \sum_{j=1}^{k_N} w_{\rho,j} N \rho \left( N \int_0^t \frac{k_N}{\sigma_\rho} Z^{(N)}_{\rho,j}(s)ds \right) + \hat{Z}^{(N)}(0)$$

In the previous equation, the vectors $w_{\rho,j}$ are the update vectors of the phase-type expansion of the Erlang distributions, having rate $\frac{k_N}{\sigma_\rho} \hat{Z}^{(N)}(s)$.

Now, we will show that $\| \hat{X}^{(N)}(t) - X^{(N)}_{k,N}(t) \|$ goes to zero in probability, which implies weak convergence. We will use a standard argument to bound centred Poisson processes, defined as $\mathbb{Y}(\lambda(t)) = \mathbb{Y}(\lambda(t)) - \lambda(t)$, assuming the rate of each function $f_\eta$ is bounded by $M_\eta$ (which follows from a standard compactness/tightness argument, and by continuity of rates, see [3]). In particular, observe that $\| \mathbb{Y}(N f_\eta^{(N)} \lambda(s)ds) \|$ can be stochastically bounded by $\| \mathbb{Y}(N M t) \|$, where $M$ is an upper bound for $\lambda(s)$. By adding and removing the arguments of Poisson processes, and by adding and subtracting an expression for $X^{(N)}_{k,N}(t)$ similar to the one for (the arguments of Poisson processes of) $X^{(N)}(t)$, we get:

$$\left\| \hat{X}^{(N)}(t) - X^{(N)}_{k,N}(t) \right\| \leq \left\| \hat{X}^{(N)}(0) - X^{(N)}_{k,N}(0) \right\| + \left\| \sum_{\eta \in T_d} v_\eta \int_0^t f^{(N)}_\eta (\hat{X}^{(N)}(s)) - f^{(N)}_\eta (\hat{X}^{(N)}_{k,N}(s))ds \right\| + \left\| \sum_{\rho \in T_d} w_{\rho} \int_0^t f^{(N)}_\rho (\hat{X}^{(N)}(s)) - f^{(N)}_\rho (\hat{X}^{(N)}_{k,N}(s))ds \right\| + \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4^{(N)}$$

where

$$\varepsilon_1 = \sum_{\eta \in T_d} \left\| v_\eta \right\| \left\| N \hat{Y}(N M t) \right\|$$

$$+ \sum_{\rho \in T_d} \left\| w_{\rho} \right\| \left\| N \hat{Y}(N M t) \right\|$$

$$\varepsilon_2 = \sum_{\eta \in T_d} \left\| v_\eta \right\| \left\| N \hat{Y}(N M t) \right\|$$

$$\varepsilon_3 = \sum_{\rho \in T_d} \left\| w_{\rho} \right\| \left\| N \hat{Y}(N M t) \right\|$$

and

$$\varepsilon_4^{(N)} = \left\| \pi_x \left( \sum_{\eta \in T_d \cup T_\eta} v_\eta \int_0^t f^{(N)}_\eta (\pi_x Z^{(N)}(s))ds \right) + \frac{k_N}{N} \sum_{\rho \in T_\eta} \sum_{j=1}^{k_N} w_{\rho,j} \frac{k_N}{\sigma_\rho} Z^{(N)}_{\rho,j}(s)ds + \hat{Z}^{(N)}(0) \right\|$$

To bound the $\varepsilon_4$ terms, observe that $\| \frac{1}{N} \hat{Y}(N M t) \|$ converges almost surely to zero, and therefore in probability [3]. Hence, it is less than $\varepsilon$ with probability going to 1 as $N$ increases. This implies that $\varepsilon_1$ and $\varepsilon_2$ can be made arbitrary small with arbitrarily large probability. To deal with $\varepsilon_3$, apply the Chebyshev bound to the Poisson random variables, to deduce that the probability that $\frac{1}{N} \hat{Y}(N M t)$ is less than $\delta$ is smaller than $\frac{\varepsilon}{N \delta^2}$. By choosing $\delta = \frac{\varepsilon}{N \delta^2}$, we then obtain that $\| \frac{1}{N} \hat{Y}(N M t) \|$ is less than $\varepsilon N \delta^2$ with probability $\frac{1}{N} \hat{Y}(N M t)$, implying $\varepsilon_3 < \max \{ w_{\rho,j} \| T_\rho \| \delta \}$. Here, in order for $k_N = o(N^{\frac{1}{2}})$, we need to go to zero for fixed $\delta'$, as $N$ increases, we need to have $k_N = o(N^{\frac{1}{2}})$. Finally, the term $\varepsilon_4^{(N)}$ goes to zero as $N$ grows, in virtue of Lemma IV.3. Note that, as $Z_{\rho,j}(t)$ propagates the value of $f_\rho(t - \sigma_\rho)$ along the delay, it is always bounded by $M_\rho$. In fact, terms in $\varepsilon_4$ are the integral representation of the solution of the DDE and of its Erlang-like ODE approximation, hence their norm converges to zero.

Now, let $L_\eta$ be the Lipschitz constant of function $f_\eta$. We can bound expression 1 by

$$\left\| \hat{X}^{(N)}(t) - X^{(N)}_{k,N}(t) \right\| \leq \left\| \hat{X}^{(N)}(0) - X^{(N)}_{k,N}(0) \right\| + \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4^{(N)}$$

$$+ \sum_{\eta \in T_d \cup T_\eta} \left\| v_\eta \right\| L_\eta \int_0^t \left\| \hat{X}^{(N)}(s) - X^{(N)}_{k,N}(s) \right\| ds$$

$$+ \sum_{\rho \in T_d} \left\| w_{\rho} \right\| L_\rho \int_0^t \left\| \hat{X}^{(N)}(s) - X^{(N)}_{k,N}(s) \right\| ds$$

By an application of the Grönwall inequality [3], calling $L = \sum_{\eta \in T_d \cup T_\eta} \left\| v_\eta \right\| L_\eta + \sum_{\rho \in T_d} \left\| w_{\rho} \right\| L_\rho$, and $\alpha_N = \| \hat{X}^{(N)}(0) - X^{(N)}_{k,N}(0) \| + \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4^{(N)}$, we get that

$$\| \hat{X}^{(N)}(t) - X^{(N)}_{k,N}(t) \| \leq \alpha_N e^{-Lt},$$

which, fixed $\varepsilon > 0$, by an appropriate choice of $\varepsilon_j$, $j = 1, 2, 3$, is less than $\varepsilon$ for $N$ large enough, with probability that goes to 1 as $N$ goes to infinity.
Remark: The difference between the current version of the proof and the previous one is quite radical. In the previous version, we used an argument that was an extension of the proof of Lemma IV.1, essentially trying to tweak the results of [1] and [2] to the case in which both processes depend on $N$. The problem with this argument is that it shows the convergence of the processes by convergence of finite dimensional distributions of an associated DTMC, i.e. convergence is proved for the distribution after any fixed number $m$ of steps. However, we are interested in convergence of the processes at any fixed time $t > 0$, but the number of steps to reach such a time increases with $N$, as the rates of the exponential transitions are proportional to $N$, and so is the density of events. Now, convergence for each $m$ does not imply convergence along an increasing sequence $m_N$ of jumps, required to reach time $t$ for system size $N$. This is because the speed of convergence will decrease as $m$ increases, for any fixed $N$, hence in increasing $m$ with $N$, the error does not necessarily go to zero. The current proof essentially introduces an additional argument, showing that the increased density of events, matched by a decrease in the jump size, cancels out the differences between the Erlang and the deterministic delays. However, as the argument is based on showing that noise goes to zero for a specific representation of the processes, this proof also implies the following lemma and the theorem of the paper (in fact, Hayden uses this argument in [5] to prove a result similar to Theorem IV.1). As for the proof of the original Lemma, first note that convergence (in probability) for each time $t$ implies weak convergence of the finite dimensional distributions, hence of the processes. This holds for any $k_N$ growing sufficiently slowly. However, increasing $k_N$ we are also increasing the accuracy of the Erlang approximation, so intuitively the error committed will be smaller than for slower $k_N$. We conjecture that a stochastic ordering argument (applied to $|X_k(t) - X(t)|$ for fixed $N$) may be used here to conclude. We are also investigating possible alternative ways of proving the Lemma, not relying on the Poisson representation of [4].

**Lemma (Lemma IV.3).** Suppose $x_0(N) \to x_0$. There exists a divergent and non-decreasing sequence $k_N$ such that, for each $t \geq 0$,

$$\|X_{k_N}(t) - x(t)\| \to 0 \quad \text{in probability,}$$

where $x(t) = x_m(t)$, for $t \in \left[\sigma_m, \sigma_{m+1}\right]$ and $x(t) = x_M(t)$, for $t \geq \sigma_M$.

**Proof:** We prove the result inductively first for $t \in [\sigma_0, \sigma_1]$, then for $t \in \left[\sigma_m, \sigma_{m+1}\right]$, assuming the result holds up to time $\sigma_m$. An alternative way of seeing the induction we use is in the number of different delays appearing in $X_d$.  

a) **Base case:** $t \in [0, \sigma_1]$ (0 delays). In this case, we consider the sequence of models $(X_{d}^{0}(N))$, which are in fact simple CTMC. Their $k_{N}$-step Erlang approximation $(\text{cal} X_{k_{N}}^{0}(N))$, when restricted to variables of $(X_{d}^{0}(N))$, in fact coincides with $(X_{d}^{0}(N))$. Furthermore, $x^0(t)$ and $y_{k_{N}}^0(t)$ are both solutions of differential equations, and $x^0(t) = x_{k_{N}}^0(t)$, where $x_{k_{N}}^0(t)$ is the projection of $y_{k_{N}}^0(t)$ to the coordinates corresponding to $x$ variables. Therefore, the proof follows from the application of Theorem ?? for a given $t$, for any sequence $k_N$.

However, we will introduce a more complex argument, that is needed in the following, i.e. when $x^0(t)$ and $x_{k_{N}}^0(t)$ will coincide only in the limit.

Essentially, the problem in this case is that the dimension of the state space of $y_{k_{N}}^0(t)$, and of the sequence of CTMC $(Y_{k_{N}}^{0}(N))(t)$ converging to it in virtue of Kurtz theorem, increases with $k_{N}$. This increment has an effect on the speed of convergence, hence if we do not choose carefully the sequence $k_{N}$, we may fail to show convergence to $x^0(t)$. In [6], an exponential error bound for Kurtz theorem is proved, which, adapted to our case, reads (we drop the index 0, as this bound is valid for any $m$):

$$\|Y_k^{N}(t) - y_k(t)\| \leq 2 \cdot n_k \cdot e^W,$$

where $n_k$ is the dimension of the system, and $W$ stands for $-\frac{1}{2} \cdot \tau_{\sigma_2} \cdot L_k$, is the Lipschitz constant of $F_k$, and the inequality holds in a set of probability at least $1 - \varepsilon$. Now, in the previous formula, we have two quantities depending on $k$: the dimension $n_k$ and the Lipschitz constant $L_k$. Both quantities depend linearly on $k$: for the dimension this is obvious, while for the Lipschitz constant, observe that the transitions for the Erlang approximation of a delay $\sigma$ contribute to the vector field (and to the Lipschitz constant) with a linear term which is proportional to $\frac{1}{\sigma}$. Therefore, if we choose a sequence $k_N$ that grows fast, the previous bound will fail to converge to zero. However, it is easy to check that $k_N = \lfloor \ln \ln N \rfloor$ will work. With this choice of $k_N$, we have that

$$\|X_{k_N}(t) - x(t)\| \leq \|X_{k_N}(t) - x_{k_N}(t)\| + \|x_{k_N}(t) - x(t)\|,$$

whose right hand side goes to zero as $N \to \infty$.

b) **Induction step:** $t \in [\sigma_m, \sigma_{m+1}]$ (m delays). Consider the DDE
d$$\frac{dx^m(t)}{dt} = F^m(x_m(t - \sigma_0), x_m(t - \sigma_1), \ldots, x_m(t - \sigma_m))$$

with initial conditions given by function $x(t)$, $t \leq \sigma_m$. By the induction hypothesis, $X_{k_N}^{(N)}(t)$ converges to $x(t)$ in probability for any $t \leq \sigma_m$. Let $t \in [\sigma_m, \sigma_{m+1}]$, as $x(t) = x^m(t)$, it holds:

$$\|X_{k_N}^{(N)}(t) - x^m(t)\| \leq \|X_{k_N}^{(N)}(t) - x_{k_N}(t)\| + \|x_{k_N}(t) - x^m(t)\|.$$
depend only on the values of $X$ variables (the rate at which Erlang delays are triggered depends only on $X$ and the rate of Erlang steps is independent of $X$), hence the infinitesimal generators of $Y_{k,N}^{(N)}$ and $(Y_{m-1}^{m-1})^{(N)}$ converge. It follows that $Y_{k,N}^{(N)} = (Y_{m-1}^{m-1})^{(N)}$ in $[0, \sigma_m]$, see [7]. Now, from the fact that and $(Y_{m-1}^{m-1})^{(N)} \Rightarrow y_{m-1}^{m-1}$, we conclude $Y_{k,N}^{(N)} \Rightarrow y_{k,N}^{m-1}$ in $[0, \sigma_m]$. But as it also holds that $Y_{k,N}^{(N)} = y_{k,N}$, and as weak convergence to a deterministic limit is equivalent to (uniform) convergence in probability for any finite time [7], it follows that $\|y_{k,N}^{(N)}(\sigma_m) - y_{k,N}^{m-1}(\sigma_m)\| \rightarrow 0$, as desired.

With the previous argument, we have in fact proved that $\|x_{k,N}(t) - x^{m}(t)\|$, for $t \in [\sigma_m, \sigma_{m+1}]$, converges to zero in probability.

 REFERENCES