Hybrid Behaviour of Markov Population Models

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

Abstract

We investigate the behaviour of population models written in Stochastic Concurrent Constraint Programming (**sCCP**), a stochastic extension of Concurrent Constraint Programming. In particular, we focus on models from which we can define a semantics of **sCCP** both in terms of Continuous Time Markov Chains (CTMC) and in terms of Stochastic Hybrid Systems, in which some populations are approximated continuously, while others are kept discrete. We will prove the correctness of the hybrid semantics from the point of view of the limiting behaviour of a sequence of models for increasing population size. More specifically, we prove that, under suitable regularity conditions, the sequence of CTMC constructed from **sCCP** programs for increasing population size converges to the hybrid system constructed by means of the hybrid semantics. We investigate in particular what happens for **sCCP** models in which some transitions are guarded by boolean predicates or in the presence of instantaneous transitions.

Keywords: Stochastic process algebras; stochastic concurrent constraint programming; stochastic hybrid systems; mean field; fluid approximation; weak convergence

1 Introduction

Stochastic Process Algebras (SPA) are a powerful framework for quantitative modelling and analysis of population processes [38]. They have been applied in a wide varieties of contexts, including computer systems [38], biological systems [24] [15] [23], ecological [58] and crowd [46] modelling.

However, their standard semantics, given in terms of Continuous Time Markov Chain (CTMC, [49]), suffers from the problem of state space explosion, which impedes the use of SPA to analyse models with a large state space. A recent technique introduced to tackle this problem is fluid-flow approximation [39], which describes the number of system components by means of continuous variables and interprets rates as flows, thus providing a semantics in terms of Ordinary Differential Equations (ODE).

The relationship between these two semantics is grounded on the law of large numbers for population processes [44], first proved by Kurtz back in the seventies [43]. Applying this theoretical framework to SPA models, one obtains that the fluid-flow ODE is the limit of the sequence of CTMC models [60, 21, 37], obtained by the standard SPA semantics for increasing system size, usually the total number of agents in the system. This also provides a link with a large body of mathematical literature on fluid and mean field approximation (see e.g. [14] for a recent review).

These results provide the asymptotic correctness of the fluid semantics and justify the use of ODEs to analyse large collective SPA models. Fluid approximation is also entering into the analysis phase in a more refined way than just by numerical simulation. For instance, in [36], the authors use fluid approximation for the computation of passage-times, while in [13] the fluid approximation scheme is used to model check properties of single agents in a large population against CSL properties.

Despite the remarkable success of fluid approximation of SPA models, its applicability is restricted to situations in which all components are present in large quantities, and all events depend continuously on the number of the different agent types. This excludes many interesting situations, essentially all those in

which some sub-populations have a fixed and small size. This is the case in biological systems, when one considers gene networks, but also in computer systems when one models some form of centralized controller. Furthermore, the description of control policies is often simplified by using forced (or instantaneous) events, happening as soon as certain condition are met, and more generally guard predicates, modulating the set of enabled actions as a function of the global state of the system.

These features of modelled systems are not easily captured in a fluid flow scheme, as they lead naturally to hybrid systems, in which discrete and continuous dynamics coexist. To deal with these situations, in [17, 18] the authors proposed a hybrid semantics for a specific SPA, namely stochastic Concurrent Constraint Programming (**sCCP**, [15]), associating with a **sCCP** model a hybrid system where continuous dynamics is interleaved with discrete Markovian jumps. In [19], also instantaneous transitions are incorporated in the framework. In this way, one can circumvent the limits of fluid-flow approximation, whilst keeping discrete only the portions of the system that cannot be safely described as continuous. Roughly speaking, this hybrid semantics works by first identifying a subset of system variables to be approximated continuously, keeping discrete the remaining ones. The latter set of variables identifies the discrete skeleton of the hybrid system, while the former defines the continuous state space. Then, each activity of agents, corresponding to a transition that modifies the state of the system, is classified as continuous, discrete/stochastic, or discrete/instantaneous. The first class of transitions is used to construct a vector field giving the continuous dynamics of the hybrid system (in each mode), while the other two transition classes define the discrete dynamics.

The advantages of working with a hybrid semantics for SPA are mainly rooted in the speed-up that can be achieved in the simulation, as discussed e.g. in [18] and [50]. Moreover, the hybrid semantics put at disposal of the modeller a broader set of analysis tools, like transient computation [61] or moment closure techniques [56, 48].

While the theory of deterministic approximation of CTMC is well developed, hybrid approximation has attracted much less attention. To the author's knowledge, the preliminary work [11] on which this paper is based was the first attempt to prove hybrid convergence results in a formal method setting. There has been some previous work on hybrid limits in [4], restricted however to a specific biological example, and in the context of large deviation theory [55], where deterministic approximation of models with level variables has been considered (but in this case transitions between modes are fast, so that the discrete dynamics is always at equilibrium in the limit). More recent work is [28], which discusses hybrid limits for genetic networks (essentially the class of models considered in [11] with some extensions).

The focus of this paper is to provide a general framework to infer consistence of hybrid semantics of SPA models in the light of asymptotic correctness. In doing this, we aimed for generality, proving hybrid limit theorems for a framework including instantaneous events, with guards possibly involving model time, random resets, and guards in continuous and stochastic transitions. The goal was to identify a broad set of conditions under which convergence holds, potentially usable in static analysis algorithmic procedures that check if a given model satisfies the conditions for convergence. We will comment on this issue in several points in the paper. To author's knowledge, this is the first attempt to discuss hybrid approximation in such generality.

We will start our presentation recalling **sCCP** (Section 2.1) and the hybrid semantics (Section 2.3). We will formally define it in terms of Piecewise Deterministic Markov Processes (Section 2.4, [31]), a class of Stochastic Hybrid Processes in which the continuous dynamics is given in terms of Ordinary Differential Equations, while the discrete dynamics is given by forced transitions (firing as soon as their guard becomes true) and by Markovian jumps, firing with state dependent rate. The hybrid semantics is defined by introducing an intermediate layer in terms of an automata based description, by the so-called Transition-Driven Stochastic Hybrid Automata (TDSHA, Sections 2.2 and 2.5, [17, 18]).

After presenting the classic fluid approximation result, recast in our framework (Section 4), we turn our attention to **sCCP** models that are converted to TDSHA containing only discrete/stochastic and continuous transitions, with no guards and no instantaneous transitions, but allowing random resets (general for discrete/stochastic transitions and restricted for continuous ones). In Section 5, we prove a limit theorem under mild consistency conditions on rates and resets, showing that the sequence of CTMC associated with a **sCCP** program, for increasing system size, converges to the limit PDMP in the sense of weak convergence. Technically speaking, the appearance of weak convergence instead of convergence in probability, in which classic fluid limit theorems are usually stated, depends on the fact that the limit process is stochastic and

can have discontinuous trajectories.

We then turn our attention to the limit behaviour in the presence of sources of discontinuity, namely instantaneous transitions (Section 6) or guards in continuous (Sections 7.2 and 7.1) or discrete/stochastic transitions (Section 7.4).

In all these cases, the situation is more delicate and the conditions for convergence are more complex. Guards in continuous transitions introduce discontinuities in the limit vector fields, requiring us to define the continuous dynamics in terms of the so-called piecewise-smooth dynamical systems [26] or, more generally, in terms of differential inclusions [3]. Here, however, we can exploit recent work in this direction [20, 35], and the hybrid convergence theorem extends easily, provided we can guarantee global existence and uniqueness of the solutions of the discontinuous differential equations.

The situation with guards for discrete/stochastic transitions and with instantaneous events is even more delicate: subtle interactions between the continuous dynamics and the surfaces in which guards can change truth status (called discontinuity or activation surfaces in the paper) can break convergence. We discuss this in detail first for instantaneous transitions (Section 6) and then for guards in discrete/stochastic transitions (Section 7.4). In these sections, we identify regularity conditions to control these subtle interactions, extending the convergence also to this setting. However, checking these conditions is more complicated, because they essentially impose restrictions on the global interactions between the vector fields and the discontinuity surfaces. A way out of this problem, hinted in the conclusions (Section 8) is to increase the randomness in the system by adding noise on resets and initial conditions or on the continuous trajectories, i.e. considering hybrid limits with continuous dynamics given by Stochastic Differential Equations or Gaussian Processes [42]. In the conclusions we will also comment on the applicability of our results to the stationary behaviour of the CTMC. Throughout the paper, starred remarks contain material that can be skipped at a first reading.

2 Preliminaries

In this section, we introduce preliminary concepts needed in the following. We will start in Section 2.1 by presenting **sCCP**, the modelling language that will be used in the paper. We will then introduce Transition-Driven Stochastic Hybrid Automata (TDSHA, Section 2.2), an high level formalism to model the limit hybrid processes of interest, namely Piecewise Deterministic Markov Processes (PDMP, Section 2.4). Finally, we will consider also how to define a hybrid semantics for **sCCP** by syntactically transforming a **sCCP** model into a TDSHA (Section 2.3) and a TDSHA into a PDMP (Section 2.5).

2.1 Stochastic Concurrent Constraint Programming

We briefly present stochastic Concurrent Constraint Programming (**sCCP**, a stochastic extension of CCP [54]). In the following we just sketch the basic notions and the concepts needed in the rest of the paper. More details on the language can be found in [10, 15].

sCCP programs are defined by a set of agents interacting asynchronously and exchanging information through a shared memory called the *constraint store*. The constraint store consists of a set of variables plus a set of constraints, which are first order predicates restricting the admissible domain of variables. By adding constraints to the store, agents refine the available information. In this paper, we consider a restricted notion of constraint store, containing only *stream variables*, i.e. variables "*a là Von Neumann*" which have a single value at any given time, and can be updated during the computation[1]. We further restrict the language by forbidding local variables. This restricted version of **sCCP**has proved to be sufficiently expressive, compact, and especially easy to manipulate for our purposes, in particular for what concerns the definition of the fluid [21] and the hybrid semantics [17, 18]. In this paper, however, we enlarge the primitives at our disposal with respect to [21, 18], as done in [19], by including also instantaneous transitions, random resets, and environment variables (which can take values in an uncountable set).

Definition 2.1. A **sCCP** program is a tuple $\mathcal{A} = (A, \mathsf{Def}, \mathbf{X}, \mathcal{D}, \mathbf{x_0})$, where

[1]Formally, one can view these variables as list, so that new values are appended at the end of the list, see [15] for further details.

1. The *initial network of agents* A and the *set of definitions* Def are given by the following grammar:

$$\begin{array}{ll} \mathsf{Def} = \emptyset \mid \mathsf{Def} \cup \mathsf{Def} \mid \{C \stackrel{\text{def}}{=} M\} & \\ A = \mathbf{0} \mid C \mid A \parallel A & M = \pi.A \mid M + M \\ \pi = [g(\mathbf{X}) \to u(\mathbf{X}, \mathbf{X}', \mathbf{W})]_{\lambda(\mathbf{X})} \mid [g(\mathbf{X}) \to u(\mathbf{X}, \mathbf{X}', \mathbf{W})]_{\infty : p(\mathbf{X})} & \end{array}$$

2. $\mathbf{X}$ is the set of stream variables of the store (with global scope). A variable $X \in \mathbf{X}$ takes values in $\mathcal{D}_X$. Variables are divided into two classes: *model* or *system variables* whose domain $\mathcal{D}_X$ has to be a countable subset of $\mathbb{R}$ (usually the integers), and *environment variables*, whose domain can be the whole $\mathbb{R}$. The state space of the model is therefore $\mathcal{D} = \prod_{X \in \mathbf{X}} \mathcal{D}_X$;

3. $\mathbf{x_0} \in \mathcal{D}$ is the *initial value* of store variables.

System variables usually describe the number of individuals of a given population, like the number of molecules in a biochemical mixture or the number of jobs waiting in a queue. Environment variables, on the other hand, are useful to describe properties of the "environment", like the temperature of a biochemical system, or the value of a controllable parameter that may change at run-time. Examples of the use of environment variables will be given in Section 5.

In the previous definition, a basic action π (called throughout the paper also *event* or *transition*) is a *guarded update* of (some of the) store variables. In particular:

- the *guard* $g(\mathbf{X})$ is a quantifier-free first order formula whose atoms are inequality predicates on variables $\mathbf{X}$;
- the *update* $u(\mathbf{X}, \mathbf{X}')$ is a predicate on $\mathbf{X}, \mathbf{X}'$, a conjunction of *atomic updates* of the form $\bigwedge X' = r(\mathbf{X}, \mathbf{W})$ (where X' denotes variable X after the update), where each variable X' appears only once. Here r is a function with values in $\mathcal{D}_X$, and can depend on the store variables $\mathbf{X}$ and on a *random* vector $\mathbf{W}$ in $\mathbb{R}^h$ (for some $h > 0$), which can also depend on the current state of variables $\mathbf{X}$. Updates will be referred to also as *resets*.
- The *rate function* $\lambda : \mathcal{D} \to \mathbb{R}_{\geq 0}$ is the (state dependent) rate of the exponential distribution associated with π, which specifies the *stochastic duration* of π;
- if, instead of λ, an action π is labelled by $\infty : p(\mathbf{X})$, it is an instantaneous action. In this case, $p : \mathcal{D} \to \mathbb{R}_{\geq 0}$ is the *weight* function associated with the action.

Updates can be seen as (random) functions from $\mathcal{D}$ to itself, and they can be very general. However, in the following we will need to restrict them in order to define the fluid semantics. An atomic reset is a *constant increment update* if it is of the form $X' = X + k$, with $k \in \mathbb{R}$ such that $X' \in \mathcal{D}_X$ whenever $X \in \mathcal{D}_X$ (usually $X, k \in \mathbb{Z}$) and it is a *random increment update* if it is of the form $X' = X + \mu$, with μ a random number, such that $|\mu|$ has *finite* expectation. An update is a constant/random increment if all its atomic updates are so.

Example 2.1. We introduce now a simple example that will be used for illustrative purposes throughout the paper. We will consider a simple client-server system, consisting of a population of clients which request a service and, after having obtained an answer, process it for some time before asking for another service, in a loop. The servers, instead, reply to client's request at a fixed rate. We ignore any internal behaviour of servers, for simplicity. However, servers can break down and need some time to be repaired. We can model such system in **sCCP** by using 4 variables, two counting the number of clients requesting a service (X_r) and processing data (X_t), and two modelling the number of idle servers ready to reply to a request (X_i) and the number of broken servers (X_b). The initial network is then `client` $\parallel$ `server`, with initial conditions $X_r = X_b = 0$, $X_t = N_1$, and $X_i = N_2$. The `client` and `server` agents are defined as follows ($*$ stands for true):

$$\begin{array}{lcl} \texttt{client} & \stackrel{\text{def}}{=} & [* \to X_r' = X_r - 1 \wedge X_t' = X_t + 1]_{\min\{k_r X_r, k_s X_i\}}\texttt{.client} + \\ & & [* \to X_r' = X_r + 1 \wedge X_t' = X_t - 1]_{k_t X_t}\texttt{.client} \\ \texttt{server} & \stackrel{\text{def}}{=} & [* \to X_i' = X_i - 1 \wedge X_b' = X_b + 1]_{k_b X_i}\texttt{.server} \\ & + & [* \to X_i' = X_i + 1 \wedge X_b' = X_b - 1]_{k_f X_b}\texttt{.server} \end{array}$$

Note in the previous code how the rate at which information is processed by clients corresponds to the global rate of observing an agent finishing its processing activity. Observe also that we defined the service rate as the minimum between the total request rate of clients and the total service rate of servers. This use of minimum is consistent with the bounded capacity interpretation of queueing theory and of the stochastic process algebra PEPA [38]. This global interaction-based modelling style is typical of **sCCP**, see [15] for a discussion in the context of systems biology. Furthermore, although we want to keep all variables ≥ 0, we are not using any guard in the transitions. However, non-negativity is automatically ensured by rates, which, by being equal to zero, disallow transitions that would make one variable negative.

In order to simplify the definition of the fluid and hybrid semantics, we will work with a restricted subclass of **sCCP** programs, that we will call *flat*. A flat **sCCP** program satisfies the following two additional restrictions: (a) each component C is of the form $C = \pi_1.C + \ldots + \pi_h.C$, i.e. it always calls itself recursively, and (b) the initial network A is the parallel composition of all components, i.e. $A = \|_{C \in \mathsf{Def}} C$. Note that the client-server model of Example 2.1 is flat.

The requirement of being flat is not a real restriction, as each **sCCP** program respecting Definition 2.1 can be turned into an equivalent flat one, by adding fresh variables counting how many copies of each component C are in parallel in the system. Guards, resets, rates and priorities have to be modified to update consistently these new variables. (see Appendix B for an example)

In the following definitions, we will always suppose to be working with flat **sCCP** models, possibly obtained by applying the flattening recipe. Given a (flat) **sCCP** model $\mathcal{A} = (A, \mathsf{Def}, \mathbf{X}, \mathcal{D}, \mathbf{x_0})$, we will denote by $\mathsf{action}_s(C)$ the set of stochastic actions of a component C and by $\mathsf{action}_i(C)$ the set of its instantaneous actions. We will use the following notation:

- For an action $\pi \in \mathsf{action}_s(C) \cup \mathsf{action}_i(C)$, we denote by $\mathbf{guard}[\pi](\mathbf{X})$ or $\mathbf{g}_\pi(\mathbf{X})$ its guard.
- For an action $\pi \in \mathsf{action}_s(C) \cup \mathsf{action}_i(C)$, we denote by $\mathbf{reset}[\pi](\mathbf{X}, \mathbf{W})$ or $\mathbf{r}_\pi(\mathbf{X}, \mathbf{W})$ its update function (so that $\mathbf{X}' = \mathbf{r}_\pi(\mathbf{X}, \mathbf{W})$).
- For an action $\pi \in \mathsf{action}_s(C) \cup \mathsf{action}_i(C)$, if π has a constant increment update, we will denote the increment vector by $\mathbf{k}_\pi$ (so that $\mathbf{X}' = \mathbf{X} + \mathbf{k}_\pi$), while if π has a random increment update, we will denote it by μ_π. We also let ν_π be either $\mathbf{k}_\pi$ or μ_π.
- For an action $\pi \in \mathsf{action}_s(C)$, we denote by $\mathbf{rate}[\pi]$ or λ_π its rate function.
- For an action $\pi \in \mathsf{action}_i(C)$, we denote by $\mathbf{weight}[\pi]$ or $\mathbf{p}_\pi$ its weight.

A **sCCP** program with all transitions stochastic can be given a standard semantics in terms of Continuous Time Markov Chains, in the classical Structural Operational Semantics style, along the lines of [15]. For a flat **sCCP** model, the derivation of the labelled transition system (LTS) is particularly simple. First, the state space of CTMC corresponds to the domain $\mathcal{D}$ of the **sCCP** variables. Secondly, each stochastic action $\pi \in \mathsf{action}_s(C)$ of a component C defines a set of edges in the LTS. In particular, if in a point $\mathbf{x}$ it holds that $\mathbf{g}_\pi(\mathbf{x})$ is true and $\mathbb{P}\{\mathbf{r}_\pi(\mathbf{x}, \mathbf{W}) = \mathbf{y}\} = p_{\mathbf{y}} > 0$, then we have a transition from $\mathbf{x}$ to $\mathbf{y}$ with rate $p_{\mathbf{y}} \lambda_\pi(\mathbf{x})$. As customary, the rates of the edges of the LTS connecting the same pair of nodes are summed up to get the corresponding rate in the CTMC. Instantaneous transitions, on the other hand, can be dealt with in the standard way as in [45], by partitioning states of $\mathcal{D}$ into vanishing (in which there is an active instantaneous transition) and non-vanishing (in which there is no active instantaneous transition), and removing vanishing states in the LTS, solving probabilistically any non-deterministic choice between instantaneous transitions with probability proportional to their weight.

We will indicate by $\mathbf{X}(t)$ the state at time t of the CTMC associated with a **sCCP** program $\mathcal{A}$ with variables $\mathbf{X}$.

If all transitions of a **sCCP** program are stochastic and have constant increment updates, they can be interpreted as flows, and a fluid semantics can be defined [21]. However, to properly deal with random resets and instantaneous transitions, it is more convenient to consider a more general semantics for **sCCP**, in terms of stochastic hybrid automata [17, 18, 19]. This approach will also allow us to partition variables and transitions into discrete and continuous, so that only a portion of the state space will be approximated as fluid.

2.2 Transition-driven Stochastic Hybrid Automata

Transition-Driven Stochastic Hybrid Automata (TDSHA, [17, 18]) has proved to be a convenient intermediate formalism to associate a Piecewise Deterministic Markov Process with a **sCCP** program. The emphasis of TDSHA is on *transitions* which, as always in hybrid automata, can be either discrete (corresponding to jumps) or continuous (representing flows acting on system variables). Discrete transitions can be of two kinds: either stochastic, happening at random jump times (exponentially distributed), or instantaneous, happening as soon as their guard becomes true.

In this paper, we consider a slight variant of TDSHA, in which discrete modes of the automaton are described implicitly by a set of discrete variables (variables taking values in a discrete set), rather than explicitly. This syntactic variant is similar to the one used in [12], and is introduced in order to simplify the mapping from flat **sCCP** models.

Definition 2.2. A Transition-Driven Stochastic Hybrid Automaton (TDSHA) is a tuple $\mathcal{T} = (\mathbf{Z}, Q, \mathbf{Y}, \mathfrak{TC}, \mathfrak{TD}, \mathfrak{TS}, init)$, where:

- $\mathbf{Z} = \{Z_1, \ldots, Z_k\}$ is the set of discrete variables, taking values in the countable set $Q \subset \mathbb{R}^k$. Each value $q \in Q$, $q = (z_1, \ldots, z_k)$ is a *control mode* of the automaton.
- $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ is a set of real valued *system variables*, taking values in $\mathbb{R}^n$. We let $\mathbf{X} = \mathbf{Z} \cup \mathbf{Y}$ be the vector of TDSHA variables, of size $m = k + n$.[2]
- $\mathfrak{TC}$ is the multi-set[3] of *continuous transitions or flows*, containing tuples $\eta = (\mathbf{k}, f)$, where $\mathbf{k}$ is a real vector of size m (identically zero on components corresponding to $\mathbf{Z}$), and $f : Q \times \mathbb{R}^n \to \mathbb{R}$ is a piecewise continuous function for each fixed $q \in Q$ (usually, but not necessarily, Lipschitz continuous[4]). We will denote them by ν_η, and $\mathbf{f}_\eta$, respectively.
- $\mathfrak{TD}$ is the multi-set of *discrete or instantaneous transitions*, whose elements are tuples $\eta = (G, R, p)$, where: $p : Q \times \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ is a *weight* function used to resolve non-determinism between two or more active transitions, G is the *guard*, a quantifier-free first-order formula with free variables among $\mathbf{X}$, and R is the *reset*, a conjunction of atoms of the form $X' = r(\mathbf{X}, \mathbf{W})$, where $r : Q \times \mathbb{R}^n \times \mathbb{R}^h \to \mathbb{R}$, is the reset function of X, depending on variables $\mathbf{X}$ as well as on a random vector $\mathbf{W}$ in $\mathbb{R}^h$. Note that the guard can depend on discrete variables, and the reset can change the value of discrete variables $\mathbf{Z}$. In the following, we will interpret the reset as a vector of $k+n$ functions, $R : Q \times \mathbb{R}^n \times \mathbb{R}^h \to Q \times \mathbb{R}^n$, equal to $X' = r(\mathbf{X}, \mathbf{W})$ in the component corresponding to X if $X' = r(\mathbf{X}, \mathbf{W})$, and equal to the identity function for all those variables X unchanged by the reset. The elements of a tuple η are indicated by $\mathbf{g}_\eta$, $\mathbf{r}_\eta$, and $\mathbf{p}_\eta$, respectively.
- $\mathfrak{TS}$ is the multi-set of *stochastic transitions*, whose elements are tuples $\eta = (G, R, \lambda)$, where G and R are as for transitions in $\mathfrak{TD}$, while $\lambda : Q \times \mathbb{R}^n \to \mathbb{R}^+$ is a function giving the state-dependent rate of the transition. Such a function is indicated by λ_η.
- $init = (\mathbf{z_0}, \mathbf{y_0}) \in Q \times \mathbb{R}^n$ is the *initial state* of the system.

A TDSHA has three types of transitions. Continuous transitions represent flows and, for each $\eta \in \mathfrak{TC}$, ν_η and $\mathbf{f}_\eta$ give the *magnitude* and the *form* of the flow of η on each variable $Y \in \mathbf{Y}$, respectively (see also Section 2.5). Instantaneous transitions $\eta \in \mathfrak{TD}$, instead, are executed as soon as their guard $\mathbf{g}_\eta$ becomes true. When they fire, they can reset both discrete and continuous variables, according to the reset policy $\mathbf{r}_\eta$, which can be either deterministic or random. Weight $\mathbf{p}_\eta$ is used to resolve probabilistically the simultaneous activation of two or more instantaneous transitions, by choosing one of them with probability proportional to $\mathbf{p}_\eta$. Finally, stochastic transitions $\eta \in \mathfrak{TS}$ happen at a specific rate λ_η, given that their guard $\mathbf{g}_\eta$ is true and they can change system variables according to reset $\mathbf{r}_\eta$. Rates define a random race in continuous time, giving the delay for the next spontaneous jump.
The dynamics of TDSHA will be defined in terms of PDMP, see Section 2.5 or [17, 18] for a more detailed discussion.

[2] Notation: the time derivative of Y_j is denoted by $\dot{Y}_j$, while the value of Y_j after a change of mode is indicated by Y'_j.

[3] Multi-sets are needed to take into account the proper multiplicity of transitions.

[4] A function $f : \mathbb{R}^m \to \mathbb{R}$ is Lipschitz continuous if and only if there is a constant $L > 0$, such that $\|f(\mathbf{x_1}) - f(\mathbf{x_2})\| \leq L\|\mathbf{x_1} - \mathbf{x_2}\|$

Composition of TDSHA. We consider now an operation to combine two TDSHA with the same vectors of discrete and continuous variables, by taking the union of their transition multi-sets. Given two TDSHA $\mathcal{T}_1 = (\mathbf{Z}, Q, \mathbf{Y}, \mathfrak{TC}_1, \mathfrak{TD}_1, \mathfrak{TS}_1, init)$ and $\mathcal{T}_2 = (\mathbf{Z}, Q, \mathbf{Y}, \mathfrak{TC}_2, \mathfrak{TD}_2, \mathfrak{TS}_2, init)$, agreeing on discrete and continuous variables and on the initial state, their composition $\mathcal{T} = \mathcal{T}_1 \uplus \mathcal{T}_2$ is simply $\mathcal{T} = (\mathbf{Z}, Q, \mathbf{Y}, \mathfrak{TC}_1 \cup \mathfrak{TC}_2, \mathfrak{TD}_1 \cup \mathfrak{TD}_2, \mathfrak{TS}_1 \cup \mathfrak{TS}_2, init)$, where the union symbol $\cup$ refers to union of multi-sets.

2.3 From sCCP to TDSHA

In this section we recall the definition of the semantics for **sCCP** in terms of TDSHA [18]. We will assume to work with flat **sCCP** models, so that we can ignore the structure of agents and focus our attention on system variables. In this respect, this approach differs from the one of [18], but it provides a more homogeneous treatment.

The mapping proceeds in three steps. First we will partition variables into discrete and continuous. Then, we will convert each component into a TDSHA, and finally we will combine these TDSHA by the composition construction defined in the previous section.

The first step is to consider a flat **sCCP** model $\mathcal{A} = (A, \mathsf{Def}, \mathbf{X}, \mathcal{D}, \mathbf{x_0})$, and partition its set of variables $\mathbf{X}$. Recall that variables $\mathbf{X}$ are divided into model variables $\mathbf{X}_m$ and environment variables $\mathbf{X}_e$. Model variables $\mathbf{X}_m$ are partitioned into two subsets: $\mathbf{X}_d$, to be kept discrete, and $\mathbf{X}_c$, to be approximated continuously. Hence $\mathbf{X} = \mathbf{X}_d \cup \mathbf{X}_c \cup \mathbf{X}_e$. How to perform this choice depends on the specificity of a given model: some guidelines will be discussed in Remarks 2.1 and 5.1. We stress here the double nature of environment variables: they will be treated like discrete variables in terms of the way they can be updated, but as continuous variables for what concerns their domain, i.e. they are part of the continuous state space of the TDSHA.

Once variables have been partitioned, we will process each component $C \in \mathsf{Def}$ separately, subdividing its stochastic actions $\mathsf{action}_s(C)$ into two subsets: $\mathsf{disc}(C)$, those to be maintained discrete, and $\mathsf{cont}(C)$, those to be treated continuously. This choice confers an additional degree of freedom to the mapping, but has to satisfy the following constraint:

Assumption 1. Continuous transitions must have a constant increment update or a random increment update, i.e. $\mathbf{r}_\pi = \mathbf{X} + \nu_\pi$. Furthermore, their reset cannot modify any discrete or environment variable, i.e. $\nu_\pi[X] = 0$, for each $X \in \mathbf{X}_d \cup \mathbf{X}_e$.

We will now sketch the main ideas behind the definition TDSHA associated with a component C.

Continuous transition. With each $\pi \in \mathsf{cont}(C)$, we associate $\eta \in \mathfrak{TC}$ with rate function $\mathbf{f}_\eta(\mathbf{X}) = \mathbf{I}\{\mathbf{g}_\pi(\mathbf{X})\} \cdot \lambda_\pi(\mathbf{X})$, where $\mathbf{I}\{\cdot\}$ is the indicator function of the predicate $\mathbf{g}_\pi(\mathbf{X})$, equal to 1 if $\mathbf{g}_\pi(\mathbf{X})$ is true, and to zero if it is false. The update vector is $\mathbf{k}_\pi$, if π has a constant increment update. If π has random increment μ_π, we define the update vector as $\mathbb{E}[\mu_\pi]$, the expected value of the random vector μ_π.[5]

Stochastic transitions. Stochastic transitions are defined in a very simple way: guards, resets, and rates are copied from the **sCCP** action $\pi \in \mathsf{disc}(C)$.

Instantaneous transitions. Instantaneous transitions are generated from **sCCP** instantaneous actions $\pi \in \mathsf{action}_i(C)$, by copying guards, resets and priorities.

We can define formally the TDSHA of a **sCCP** component as follows:

Definition 2.3. Let $\mathcal{A} = (A, \mathsf{Def}, \mathbf{X}, \mathcal{D}, \mathbf{x_0})$ be a flat **sCCP** program and $(\mathbf{X}_d, \mathbf{X}_c, \mathbf{X}_e)$ be a partition of the variables $\mathbf{X}$. Let C be a component, with stochastic actions $\mathsf{action}_s(C)$ partitioned into $\mathsf{disc}(C) \cup \mathsf{cont}(C)$, in agreement with Assumption 1. The TDSHA associated with C is $\mathcal{T}(C, \mathsf{disc}(C), \mathsf{cont}(C)) = (\mathbf{Z}, Q, \mathbf{Y}, \mathfrak{TC}, \mathfrak{TD}, \mathfrak{TS}, init)$, where

[5] Alternatively, we could have considered the support $\{\mu_\pi^1, \ldots, \mu_\pi^k, \ldots\}$ of μ_π, with probability density $P(\mu_\pi^1), \ldots, P(\mu_\pi^k), \ldots$, and generated a family of continuous transitions with rate $P(\mu_\pi^k)\mathbf{f}_\eta(\mathbf{X})$ and update vector μ_π^k. However, if we add up these transitions as required to construct the vector field (see Section 2.5), we obtain $\mathbb{E}[\mu_\pi]\mathbf{f}_\eta(\mathbf{X})$, i.e. the two approaches are equivalent.

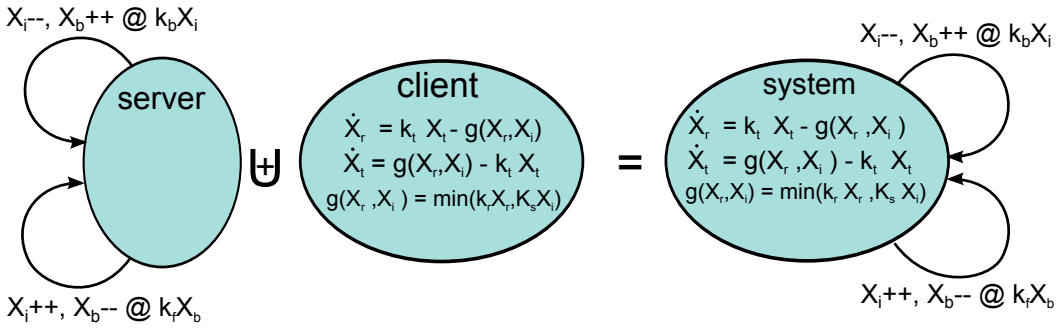


Figure 1: TDSHA associated with `client` and `server` components of Example 2.1, together with their composition. Continuous transitions are rendered into a set of ODEs, as explained in Section 2.5.

- $\mathbf{Z}$ is equal to $\mathbf{X}_d$, while $\mathbf{Y} = \mathbf{X}_c \cup \mathbf{X}_e$. Q is the domain of $\mathbf{X}_d$ in $\mathcal{A}$, and $init = \mathbf{x_0}$.
- With each $\pi \in \mathsf{cont}(C)$ with constant increment reset $\mathbf{r}_\pi = \mathbf{X} + \mathbf{k}_\pi$, we associate $\eta = (\mathbf{k}_\pi, \mathbf{f}_\eta) \in \mathfrak{TC}$, where $\mathbf{f}_\eta(\mathbf{X}) = \mathbf{1}\{\mathbf{g}_\pi(\mathbf{X})\} \cdot \lambda_\pi(\mathbf{X})$.
- With each $\pi \in \mathsf{cont}(C)$ with random increment reset $\mathbf{r}_\pi = \mathbf{X} + \mu_\pi$, we associate $\eta = (\mathbb{E}[\mu_\pi], \mathbf{f}_\eta) \in \mathfrak{TC}$, where $\mathbf{f}_\eta$ is defined as above.
- With each $\pi \in \mathsf{disc}(C)$ we associate $\eta = (\mathbf{g}_\pi(\mathbf{X}), \mathbf{r}_\pi(\mathbf{X}), \lambda_\pi(\mathbf{X})) \in \mathfrak{TG}$.
- With each $\pi \in \mathsf{action}_i(C)$ we associate $\eta = (\mathbf{g}_\pi(\mathbf{X}), \mathbf{r}_\pi(\mathbf{X}), \mathbf{p}_\pi(\mathbf{X})) \in \mathfrak{TD}$.

Finally, the TDSHA of the whole **sCCP** program is obtained by taking the composition of the TDSHA of each component, as follows:

Definition 2.4. Let $\mathcal{A} = (A, \mathsf{Def}, \mathbf{X}, \mathcal{D}, \mathbf{x_0})$ be a flat **sCCP** program and $(\mathbf{X}_d, \mathbf{X}_c, \mathbf{X}_e)$ be a partition of variables $\mathbf{X}$. The TDSHA $\mathcal{T}(\mathcal{A})$ associated with $\mathcal{A}$ is

$$\mathcal{T}(\mathcal{A}) = \biguplus_{C \in \mathsf{Def}} \mathcal{T}(C, \mathsf{disc}(C), \mathsf{cont}(C)).$$

Example. Consider the **sCCP** program of Example 2.1. The TDSHA associated with its two components (`client` and `server`) and their composition are shown in Figure 1. In this case, we partitioned variables by making all client variables continuous, i.e. X_r and X_t, and all server variables discrete, i.e. X_i and X_b. This describes a situation in which there are few servers that have to satisfy the requests of many clients. Consequently, we considered all `client` transitions as continuous and all `server` transitions as discrete.

Remark 2.1. Choosing how to partition variables into discrete and continuous is a complicated matter, and depends on specific features of the model under study. We postpone a more detailed discussion on this issue to Remark 5.1 in Section 5, as this choice can depend on the notion of system size, which has still to be introduced. Here we just note that a non-flat **sCCP** model may naturally suggest a candidate subset of variables to be kept discrete, namely state variables of a sequential **sCCP** component (i.e. an agent changing state but never forking or killing itself) present in a single copy in the network. This is the approach followed e.g. in [17, 19] to define the control modes of the TDSHA. However, the approach presented here is more general: different partitions of model variables and stochastic transitions lead to different TDSHA, which can be arranged in a lattice, as done in [18].

2.4 Piecewise Deterministic Markov Processes

The dynamical evolution of Transition Driven Stochastic Hybrid Automata is defined by mapping them to a class of stochastic processes known as Piecewise Deterministic Markov Processes (PDMP, [31]). They have a continuous dynamics given by the solution of a set of ODE and a discrete and stochastic dynamics given by a Markov jump process. The following definition deviates slightly from the classical one for PDMP in the way the discrete state space is described.

Definition 2.5. A PDMP is a tuple $(\mathbf{Z}, Q, \mathbf{Y}, E, \phi, \lambda, R)$, such that:

- $\mathbf{Z}$ is a set of discrete variables, taking values in the countable set $Q \subset \mathbb{R}^k$, the set of *modes* or *discrete states*. (Hence $q \in Q$ is of the form $(z_1, \ldots, z_k)$.) $\mathbf{Y}$ is a vector of variables of dimension $|\mathbf{Y}| = n$. For each $q \in Q$, let $E_q \subset \mathbb{R}^n$ be an open set, the continuous domain of mode q. E, the *hybrid state space*, is defined as the disjoint union of E_q sets, namely $E = \bigcup_{q \in Q} \{q\} \times E_q$. A point $\mathbf{x} \in E$ is a pair $\mathbf{x} = (q, \mathbf{y})$, $\mathbf{y} \in E$.[6] In the following, we will denote $\mathbf{Z} \cup \mathbf{Y}$ by $\mathbf{X}$, so that variables $\mathbf{X}$ range over E.
- With each mode $q \in Q$ we associate a vector field $F_q : E_q \to \mathbb{R}^n$. The ODE $\dot{\mathbf{y}} = F_q(\mathbf{y})$ is assumed to have a unique solution starting from each $\mathbf{y}_0 \in E_q$, globally existing in E_q (i.e., defined until the time at which the trajectory leaves E_q). The (semi)flow $\phi_q : E_q \times \mathbb{R}^+ \to \mathbb{R}^n$ of such vector field is assumed to be continuous in both arguments. $\phi_q(t, \mathbf{y}_0)$ denotes the point reached at time t starting from $\mathbf{y}_0 \in E_q$.[7]
- $\lambda : E \to \mathbb{R}^+$ is the *jump rate* and it gives the hazard of executing a discrete transition. It is assumed to be (locally) integrable.
- $R : E \cup \partial E \times \mathcal{B} \to [0, 1]$ is the *transition measure* or *reset kernel*. It maps each $y \in E \cup \partial E$ on a probability measure on $(E, \mathcal{B})$, where $\mathcal{B}$ is the Borel σ-algebra of E. $R(\mathbf{x}, A)$ is required to be measurable in the first argument and a probability measure for each $\mathbf{x}$.

The idea underlying the dynamics of PDMP is that, within each mode q, the process evolves along the flow ϕ_q. While in a mode, the process can jump spontaneously with hazard given by the rate function λ_q. Moreover, a jump is immediately performed whenever the boundary of the state space of the current mode is hit.

In order to formally capture the evolution, we need to define the sequence of jump times and target states of the PDMP, given by random variables $T_1, \chi_1, T_2, \chi_2, \ldots$. Let $t_*(\mathbf{x}) = \inf\{t > 0 \mid \phi_q(t, \mathbf{x}) \in \partial E_q\}$ (with $\inf \emptyset = \infty$) be the hitting time of the boundary ∂E_q starting from $\mathbf{x} = (q, \mathbf{y}) \in E$. We can define the survivor function of the first jump time T_1, given that the process started at $\mathbf{x} = (q, \mathbf{y})$, by $F(t, \mathbf{x}) = \mathbb{P}(T_1 \geq t) = I_{t < t_*(\mathbf{x})} \exp\left(-\int_0^t \lambda(q, \phi_q(s, \mathbf{x}))\mathrm{d}s\right)$.
This defines the probability distribution of the first jump time T_1, which can be simulated, as customary, by solving for t the equation $F(t, \mathbf{x}) = U_1^1$, with U_1^1 uniform random variable in $[0, 1]$. Once the time of the first jump has been drawn, we can sample the target point χ_1 of the reset map from the distribution $R(\mathbf{x}_{T_1}^-, \cdot)$, with $\mathbf{x}_{T_1}^- = \phi_q(T_1, \mathbf{x})$, using another independent uniform random variable U_1^2. From $\chi_1 = (q_1, \mathbf{x}_1)$, the process follows the flow $\phi_{q_1}(t - T_1, \mathbf{x}_1)$, until the next jump, determined by the same mechanism presented above.

Using two independent sequences of uniform random variables U_N^1 and U_N^2, we are effectively construct a realization of the PDMP in the Hilbert cube $[0, 1]^\omega$. A further requirement is that, letting $N_t = \sum_k \mathbf{I}\{t > T_k\}$ be the random variable counting the number of jumps up to time t, it holds that N_t is finite with probability 1, i.e. $\forall t, \mathbb{E}N_t < \infty$, see [31] for further details. If this holds, then the PDMP is called *non-Zeno*.

Remark 2.2. In [11], we proved some limit results restricting the attention to the case in which no instantaneous jump can occur. This amounts to requiring that each E_q has no boundaries, i.e. $E_q = \mathbb{R}^n$, or, more precisely, that $t^*(\mathbf{x}) = \infty$ for each $\mathbf{x} \in E$. If, in addition to this description, we also require the vector field to be Lipschitz continuous and the stochastic jumps to be described by a finite set of transitions η with rate λ_η and reset given by a constant increment ν_η, we obtain the so called *simple PDMP* [11].

[6] See Appendix C for a brief discussion on metric and topological properties of hybrid state spaces.

[7] Usually, F_q is *locally Lipschitz continuous*, hence the solution exists and is unique, provided trajectories do not explode in finite time. However, as in the paper we will consider also situations in which the vector field con be discontinuous due to the presence of guards, we have chosen this more general condition.

2.5 From TDSHA to PDMP

The mapping of TDSHA into PDMP is quite straightforward, with the exception of the definition of the reset kernel. Essentially, the problem lies in the fact that each discrete transition of a PDMP has to jump in the interior of the state space E, which will be defined as the set of points in which no guard of any instantaneous transition is active. However, in a TDSHA we do not have any control over this fact, and we may define guards of transitions $\eta \in \mathfrak{TD}$ in such a way that an infinite sequence of them can fire in the same time instant. For instance, the transitions $(X >= 1, X' = 0, 1)$ and $(X <= 0, X' = 1, 1)$ will loop forever if one of them is triggered. In order to avoid such bad behaviours, we will forbid by definition the possibility that two discrete transitions fire in the same time instant. We will call *chain-free* a TDSHA with this property. This condition is unnecessarily restrictive, and can be relaxed allowing the firing of a finite number of finite sequences (*loop-free* TDSHA), as done in [18], but it allows a simpler definition of the reset kernel of the PDMP. The interested reader is referred to [18] for the construction of the reset kernel for loop-free TDSHA. The good news here is that all the results in this paper extend immediately to loop-free TDSHA. The bad news is that checking if a TDSHA is loop-free is in general undecidable, as one can easily encode an Unlimited Register Machine in a TDSHA [18]. However, some sufficient conditions in terms of acyclicity of a graph constructed from transitions in $\mathfrak{TD}$ have been discussed in [34]. Practically, most models will satisfy the chain-free condition, as the discrete controller described by instantaneous transitions is usually simple. More advanced controllers will perform some form of local computation, which can then result in a loop-free model. Violation of the loop-free property, instead, usually indicates an error in the model.

We now briefly introduce some notation, and then define chain-free TDSHA and the PDMP associated with a chain-free TDSHA.

Let $\mathcal{T} = (\mathbf{Z}, Q, \mathbf{Y}, \mathfrak{TC}, \mathfrak{TD}, \mathfrak{TS}, init)$ be a TDSHA. Given a transition $\eta \in \mathfrak{TD}$, we let $G_\eta = \{\mathbf{x} \in Q \times \mathbb{R}^n \mid \mathbf{g}_\eta(\mathbf{x})\ true\}$, and $R_\eta = \{\mathbf{x} \in Q \times \mathbb{R}^n \mid \mathbf{x} \in \mathbf{r}_\eta(\bar{G}_\eta, \mathbf{W})\}$. R_η is the set of points that can be reached after the firing of η, defined as the image under $\mathbf{r}_\eta$ of the closure $\bar{G}_\eta$ of the activation set G_η of the guard. Similarly, for $\eta \in \mathfrak{TS}$, we let $R_\eta = \{\mathbf{x} \in Q \times \mathbb{R}^n \mid \mathbf{x} \in \mathbf{r}_\eta(\{\mathbf{x}_1 \mid \lambda_\eta(\mathbf{x}_1) > 0\}, \mathbf{W})\}$, the set of points that can be reached by a stochastic jump.

Definition 2.6. A TDSHA is chain-free if and only if, for each $\eta_1 \in \mathfrak{TD} \cup \mathfrak{TS}$ and each $\eta_2 \in \mathfrak{TS}$, $R_{\eta_1} \cap \bar{G}_{\eta_2} = \emptyset$.

Consider now a chain-free TDSHA $\mathcal{T} = (\mathbf{Z}, Q, \mathbf{Y}, \mathfrak{TC}, \mathfrak{TD}, \mathfrak{TS}, init)$. Then, its associated PDMP $\mathcal{P} = (\mathbf{Z}, Q, \mathbf{Y}, E, \phi, \lambda, R)$ is defined by:

- Discrete and continuous variables, and discrete modes Q, are the same both in $\mathcal{T}$ and in $\mathcal{P}$.
- The state space of the PDMP, encoding the *invariant region* of continuous variables in each discrete mode, is defined as the set of points in which no instantaneous transition is active:

$$E = \bigcap_{\eta \in \mathfrak{TD}} \overline{G}_\eta^c.$$

  Note that E_q is open, because we are intersecting the complement of the closure of each set G_η.

- The vector field is constructed from continuous transitions, by adding their effects on system variables:

$$F(\mathbf{x}) = \sum_{\eta \in \mathfrak{TC}} \nu_\eta \cdot \mathbf{f}_\eta(\mathbf{x}). \tag{1}$$

- The rate function λ is defined by adding point-wise the rates of all active stochastic transitions:

$$\lambda(\mathbf{x}) = \sum_{\eta \in \mathfrak{TS}} \mathbf{I}\{\mathbf{g}_\eta(\mathbf{x})\}\lambda_\eta(\mathbf{x}). \tag{2}$$

- The reset kernel R for $\mathbf{x} \in E$ is obtained by choosing the reset of one active stochastic transition in $\mathbf{x}$ with a probability proportional to its rate. As all such resets jump to points in the interior of E by the chain-free property of the TDSHA, we have

$$R(\mathbf{x}, A) = \sum_{\eta \in \mathfrak{T}\mathfrak{S}} \left(\frac{\mathbf{I}\{\mathbf{g}_\eta(\mathbf{x})\}\lambda_\eta(\mathbf{x})}{\lambda(\mathbf{x})} \mathbb{P}\{\mathbf{r}_\eta(\mathbf{x}, \mathbf{W}) \in A\} \right), \tag{3}$$

  where $A \in \mathcal{B}$, the Borel σ-algebra of E. If the reset of η is deterministic, then $\mathbb{P}\{\mathbf{r}_\eta(\mathbf{x}, \mathbf{W}) \in A\} = \delta_{\mathbf{r}_\eta(\mathbf{x}, \mathbf{W})}$, where $\delta_{\mathbf{x}_1}(A)$ is the Dirac measure on the point $\mathbf{x}_1 \in E$, assigning probability 1 to $\mathbf{x}_1$ and 0 to the rest of the space.

- The reset kernel R on the boundary ∂E is defined from resets of instantaneous transitions. If more than one transition is active in a point $\mathbf{x} \in \partial E$, we choose one of them with probability proportional to their weight. Let $\mathbf{p}(\mathbf{x}) = \sum_{\eta \in \mathfrak{T}\mathfrak{D} \mid \mathbf{g}_\eta(\mathbf{x}) \ true} \mathbf{p}_\eta(\mathbf{x})$, then

$$R(\mathbf{x}, A) = \sum_{\eta \in \mathfrak{T}\mathfrak{D} \mid \mathbf{g}_\eta(\mathbf{x}) \ true} \left(\frac{\mathbf{p}_\eta(\mathbf{x})}{\mathbf{p}(\mathbf{x})} \mathbb{P}\{\mathbf{r}_\eta(\mathbf{x}, \mathbf{W}) \in A\} \right). \tag{4}$$

- The initial point is $\mathbf{x}_0 = init$.

From now on, we implicitly assume that all the TDSHA obtained by the **sCCP** models we consider are chain-free. In general this may not be true and has to be checked. However, the property will hold straightforwardly in all the examples of this paper, and it will also be true in many practical examples. Indeed, as it is enough to consider loop-free TDSHA [18], this check may be automatically performed by the method of [34].

3 System Size and Normalisation

In this paper we are concerned with the correctness of the hybrid semantics of **sCCP** in terms of approximation or limit results. Essentially, we want to show that "taking the system to the limit", the standard CTMC semantics of **sCCP** converges (in a stochastic sense) to the PDMP defined by the hybrid semantics.

Clearly, this idea of convergence requires us to have a sequence of models. This sequence will depend on the size γ of the system. Hence, we will be concerned with the behaviour of a **sCCP** program, when the system size goes to infinity.

The concrete notion of system size depends on the model under examination and the type of system being modelled. In general, it is related to the *size of the population*, intended as the number of agents or entities in the system (which in flat **sCCP** models are counted by the system variables). For instance, in the client/server example (Example 2.1), this can be the total number of clients or the total number of clients and servers. In an epidemic model, this can be the size of the total population, or of the initial population, if we allow also birth and death events. However, the notion of system size can also be connected to the size of the population in an area or a volume. In this case, when the size increases, both the number of agents and the area or volume increase, usually keeping constant the density (number to area or volume ratio). The classical examples here are biochemical systems, in which we consider molecules in a given volume. Furthermore, in a model of bacteria's growth (like that of Example B.1), we may be interested in increasing the number of bacteria together with the area of the Petri dish in which the culture is grown.

In order to make the notion of size explicit, we will decorate a **sCCP** model with the corresponding population size.

Definition 3.1. A population-**sCCP** program $(\mathcal{A}, \gamma)$ consists of a **sCCP** program $\mathcal{A}$ together with the population size $\gamma \in \mathbb{R}^+$.

It is intended that rates of transitions, and even updates, of a population-**sCCP** program can depend on the population size γ. We further stress that, in a population-**sCCP** program, model variables usually take integer values.

Example 3.1. We go back to the client-server model of Example 2.1, and consider the population-**sCCP** model in which the size γ corresponds to the total population of clients and servers, namely $\gamma = N_1 + N_2 = \mathbf{X}_r(0)+\mathbf{X}_t(0)+\mathbf{X}_i(0)+\mathbf{X}_b(0)$. In this scenario, we are interested in what happens when the total population increases, maintaining constant the client-to-server ratio.

A different notion of size can be envisaged, corresponding to a different scaling law. More specifically, we can consider $\gamma = N_1 = \mathbf{X}_r(0) + \mathbf{X}_t(0)$, the total number of clients in the system. Increasing this notion of size, we are effectively increasing the number of clients requesting information to a fixed number of servers. Intuitively, these two different scalings for the client-server system should correspond to two different limit behaviours (taking γ to infinity).

In order to compare models for increasing values of the size γ, we need to normalize models to the same scale. This is done by the *normalization* operation. Essentially, we will divide system variables by the system size (in fact, only those that will be approximated continuously), and express guards, rates, and resets in terms of such normalized variables.

We formalize now the operation of normalization. Consider a population-**sCCP** program $(\mathcal{A}, \gamma)$, with $\mathcal{A} = (A, \mathsf{Def}, \mathbf{X}, \mathcal{D}, \mathbf{x_0})$, let $\mathbf{X}(t)$ be the associated CTMC, and assume that variables $\mathbf{X}$ are partitioned into discrete $\mathbf{X}_d$, continuous $\mathbf{X}_c$, and environment variables $\mathbf{X}_e$. Then the normalized CTMC $\hat{\mathbf{X}}(t)$ is constructed as follows:

- Normalized variables are $\hat{\mathbf{X}} = (\mathbf{X}_d, \hat{\mathbf{X}}_c, \mathbf{X}_e)$, with $\hat{\mathbf{X}}_c = \gamma^{-1}\mathbf{X}_c$;
- Given a stochastic action $\pi = (\mathbf{g}_\pi(\mathbf{X}), \mathbf{r}_\pi(\mathbf{X}), \lambda_\pi(\mathbf{X})$, we define:
  - $\hat{\mathbf{g}}_\pi(\hat{\mathbf{X}}) = \mathbf{g}_\pi(\mathbf{X})$, the guard predicate with respect to normalized variables;
  - Let $X' = \mathbf{r}_\pi[X](\mathbf{X}, \mathbf{W})$. If $X \in \mathbf{X}_d \cup \mathbf{X}_e$, then $\hat{\mathbf{r}}_\pi[X](\hat{\mathbf{X}}, \mathbf{W}) = \mathbf{r}_\pi[X](\mathbf{X}, \mathbf{W})$. Otherwise, if $X \in \mathbf{X}_c$, then $\hat{\mathbf{r}}_\pi[X](\hat{\mathbf{X}}, \mathbf{W}) = \gamma^{-1}\mathbf{r}_\pi[X](\mathbf{X}, \mathbf{W})$ (hence, we replaced $\mathbf{X}_c$ variables with their normalized counterpart in the reset function, but also rescaled the reset of $\hat{\mathbf{X}}_c$ variables by dividing the reset function for γ);
  - $\hat{\lambda}_\pi(\hat{\mathbf{X}}) = \lambda_\pi(\mathbf{X})$.
- Instantaneous transitions are rescaled in the same way (expressing the weight function $\hat{\mathbf{p}}$ in terms of normalized variables like the rate of stochastic transitions);
- Normalized initial conditions are $\hat{\mathbf{x}}_0 = (\mathbf{x}_{d,0}, \gamma^{-1}\mathbf{x}_{c,0}, \mathbf{x}_{e,0})$.

Applying this transformation to a **sCCP** program, we can construct the normalized CTMC $\hat{\mathbf{X}}(t)$ along the lines of the construction of Section 2.1. Furthermore, we can construct the TDSHA associated with a **sCCP** program by considering normalized transitions and variables, instead of non-normalized ones. As we will always compare normalized processes, we will always assume that this construction has been carried out.

Given a population-**sCCP** program $(\mathcal{A}, \gamma)$, our goal is to understand what will be the limit behaviour of a sequence of normalized CTMC $\hat{\mathbf{X}}^{(N)}(t)$, constructed from $(\mathcal{A}, \gamma)$ and a sequence of system sizes $\gamma_N \to \infty$ as $N \to \infty$. In order to properly do this, we need to get a better grasp on some related questions, namely:

1. how to split variables into discrete and continuous;
2. how rates and updates scale with the system size.

These two questions are somehow dependent; the last one, in particular, is crucial, as the correct form of the limit depends on the scaling of rates and updates. Investigating these issues, moreover, will give us some hints on how to choose discrete and continuous variables and transitions to define the hybrid semantics of **sCCP**.

We will start by considering the fluid case, in which all variables are approximated as continuous. Rates and updates will be required to scale in a consistent way, and we will refer to these conditions as the *continuous scaling*. Then, we will turn our attention to hybrid scaling and hybrid limits.

Before doing this, we stress that the normalization operation extends naturally to the TDSHA associated with a population-**sCCP** program and, consequently, to the PDMP associated with the so-obtained TDSHA. In particular, if we have a sequence $(\mathcal{A}, \gamma_N)$ of population-**sCCP** models, we can construct its normalization for each N, and associate a TDSHA with each element of the sequence. We call $\hat{\mathcal{T}}(\mathcal{A}, \gamma_N)$ such a TDSHA. However, in the rest of the paper we are interested in the limit behaviour, i.e. in models independent of γ_N. The scaling conditions for each transition that we will introduce will naturally lead to the construction of a limit TDSHA, independent of any notion of size, referred to as $\hat{\mathcal{T}}(\mathcal{A})$ in the rest of the paper.

4 Continuous Scaling and Fluid Limit

We discuss now the standard fluid limit [43] [44] [42] [29] [30] in our context. We will consider a sequence of population-**sCCP** programs $(\mathcal{A}, \gamma_N)$ with divergent population size $\gamma_N \to \infty$ as $N \to \infty$.

In the rest of this section, we will require the following assumptions:

- All variables $\mathbf{X}$ *are continuous* and thus normalized according to the recipe of the previous section (hence there are no discrete or environment variables).
- There is *no instantaneous transition* in $\mathcal{A}$.
- All stochastic transitions are *unguarded* and have *constant or random increment updates*.

In order to define the continuous scaling, we consider the domain $E \subseteq \mathbb{R}^m$ of normalized variables (note that here E is not a hybrid state space), which depends on possible values that non-normalized variables can take in $\mathbb{R}^m$ (usually in $\mathbb{Z}^m$, see also Remark 4.2 below). In particular, we assume that E contains the domain of the normalized variables of a population-**sCCP** program $(\mathcal{A}, \gamma_N)$ for any $N \geq 0$, so that also the limit process will be defined in E.

Now we state the continuous scaling assumptions:

Scaling 1 (Continuous Scaling). A normalized **sCCP** transition $\hat{\pi} = (true, \hat{\mathbf{X}}' = \hat{\mathbf{X}} + \hat{\nu}_\pi^{(N)}, \hat{\lambda}_\pi^{(N)}(\hat{\mathbf{X}}))$ of a population-**sCCP** program $(\mathcal{A}, \gamma_N)$, with $E \subseteq \mathbb{R}^m$ the domain of normalised variables $\hat{\mathbf{X}}$, has *continuous scaling* if and only if:

1. there is a function $g_\pi^{(N)} : E \to \mathbb{R}_{\geq 0}$ such that $\hat{\lambda}_\pi^{(N)}(\hat{\mathbf{X}}) = \gamma_N g_\pi^{(N)}(\hat{\mathbf{X}})$. Furthermore, $g_\pi^{(N)}$ converges uniformly to a locally Lipschitz continuous and locally bounded function $g_\pi : E \to \mathbb{R}_{\geq 0}$ (rates are $O(\gamma_N)$);
2. There is a constant or random vector $\nu_\pi \in \mathbb{R}^m$ such that the non-normalized increments $\nu_\pi^{(N)}$ converge weakly to ν_π, $\nu_\pi^{(N)} \Rightarrow \nu_\pi$.[8] Furthermore, $\nu_\pi^{(N)}$ and ν_π have *bounded* and *convergent first order moments*, i.e. $\mathbb{E}[\|\nu_\pi^{(N)}\|] < \infty$, $\mathbb{E}[\|\nu_\pi\|] < \infty$, $\mathbb{E}[\nu_\pi^{(N)}] \to \mathbb{E}[\nu_\pi]$, and $\mathbb{E}[\|\nu_\pi^{(N)}\|] \to \mathbb{E}[\|\nu_\pi\|]$. In particular, it follows that normalized increments are $\Theta(\gamma_N^{-1})$.

The intuition behind the previous conditions is that, as the system size increases, rates increase, leading to an increase of the density of events on the temporal axis. Furthermore, the increments become smaller and smaller, suggesting that the behaviour of the CTMC will become deterministic, with instantaneous variation equal to its mean increment. This will produce a limit behaviour described by the solution of a differential equation.

Remark 4.1. Scaling 1 can be generalized in some way, see for instance [30, 29]. However, the version stated here is sufficiently general to deal with **sCCP** programs. If we further restrict the previous scaling condition, requiring that $g_\pi^{(N)}(\hat{\mathbf{X}}) = g_\pi(\hat{\mathbf{X}})$, where $g_\pi(\hat{\mathbf{X}})$ is a locally Lipschitz function independent of γ_N, and $\nu_\pi^{(N)} = \nu_\pi$, then we obtain the so-called *density dependent* scaling. For instance, all transitions in the client/server model of Example 2.1 are density dependent, as easily checked.

[8] The concept of weak convergence is introduced in Appendix C.

*Remark** 4.2. The structure of the domain E of normalized variables depends mainly on conservation properties of the system modelled. For instance, a closed population model (i.e. without birth and death events) will preserve the total population (this is the case for the client/server model of Example 2.1), hence the domain of the normalized variables will be contained in the unit simplex in $\mathbb{R}^m$, which is a compact set. For open systems, for instance a model of growth of a population of bacteria (see Example B.1), in which the population can (in principle) become unbounded, the domain can be the whole $\mathbb{R}^m$. However, it is unlikely that populations actually diverge (one may question the reliability of the model itself, if this happens), hence one can usually find a compact set that contains the interesting part of the state space (at least up to a finite time horizon). In particular, some of the hypotheses that we will state afterwards, like locally Lipschitzness or local boundedness, rely on this implicit assumption (i.e., that we can restrict our attention to a compact set in any finite time horizon). We will further discuss these issues while proving main theorems, once they emerge.

In order to state the fluid limit theorem, we need to construct the limit ODE. This is done according to the recipe of equation 1. More specifically, for each N we construct the drift or mean increment in $\hat{\mathbf{x}}$ as

$$F^{(N)}(\hat{\mathbf{x}}) = \sum_{\pi} \mathbb{E}[\hat{\nu}_{\pi}^{(N)}]\hat{\lambda}_{\pi}^{(N)}(\hat{\mathbf{x}}) = \sum_{\pi} \mathbb{E}[\nu_{\pi}^{(N)}]g_{\pi}^{(N)}(\hat{\mathbf{x}}), \tag{5}$$

where the sum ranges over all stochastic actions of the **sCCP** program $\mathcal{A}$. If all **sCCP** transitions satisfy the continuous scaling assumption, $F^{(N)}$ converges uniformly to

$$F(\hat{\mathbf{x}}) = \sum_{\pi} \mathbb{E}[\nu_{\pi}]g_{\pi}(\hat{\mathbf{x}}). \tag{6}$$

The limit ODE is therefore $\frac{\mathrm{d}\hat{\mathbf{x}}(t)}{\mathrm{d}t} = F(\hat{\mathbf{x}}(t))$, whose solution starting from $\mathbf{x}_0$ is denoted by $\hat{\mathbf{x}}(t)$. Note that this limit ODE can be obtained in terms of TDSHA with continuous transitions only, by the construction of Section 2.5. In particular, the limit TDSHA corresponding to the fluid ODE has a continuous transition of the form $(\mathbb{E}[\hat{\nu}_{\pi}], \gamma_N g_{\pi}) = (\gamma_N^{-1}\mathbb{E}[\nu_{\pi}], \gamma_N g_{\pi})$ for each normalized **sCCP** transition $\hat{\pi}$.

Theorem 4.1 (Kurtz [43, 29, 30])**.** *Let $(\mathcal{A}, \gamma_N)$ be a sequence of population-**sCCP** models for increasing system size $\gamma_N \to \infty$, satisfying the conditions of this section, and with all **sCCP**-actions π satisfying the continuous scaling condition. Let $\hat{\mathbf{X}}^{(N)}(t)$ be the sequence of normalized CTMC associated with the **sCCP**-program and $\hat{\mathbf{x}}(t)$ be the solution of the fluid ODE.*
If $\hat{\mathbf{x}}_0^{(N)} \to \hat{\mathbf{x}}_0$ almost surely, then for any $T < \infty$, $\sup_{t \le T} \|\hat{\mathbf{X}}^{(N)}(t) - \hat{\mathbf{x}}(t)\| \to 0$ as $N \to \infty$, almost surely. □

Example. Consider again the client/server model of Example 2.1, in which both the number of clients and of servers is increased. Therefore, consider a sequence of models with size γ_N equal to the total number of clients and servers. It is easy to see that its normalized models all live in the unit simplex E in $\mathbb{R}^4$, and that all its transitions are density dependent, hence satisfy the continuous scaling. Assume that $\hat{\mathbf{x}}_0 = (c, 0, s, 0)$, with $c + s = 1$, so that $\mathbf{X}_0^{(N)} = (cN, 0, sN, 0)$, meaning that we keep constant the client-to-server ratio. The fluid ODE associated with this model is

$$\begin{cases} \frac{\mathrm{d}x_r}{\mathrm{d}t} &= k_t x_t - \min\{k_r x_r, k_s x_i\} \\ \frac{\mathrm{d}x_t}{\mathrm{d}t} &= \min\{k_r x_r, k_s x_i\} - k_t x_t \\ \frac{\mathrm{d}x_i}{\mathrm{d}t} &= k_f x_b - k_b x_i \\ \frac{\mathrm{d}x_b}{\mathrm{d}t} &= k_b x_i - k_f x_b \end{cases}$$

Hence, we can apply Theorem 4.1 to infer convergence of the CTMC sequence $\hat{\mathbf{X}}^{(N)}$ to its solution.

Remark 4.3. The version of Kurtz theorem we presented here is similar to the one of [29], but with scaling taken from [30]. The point of the scaling is to prove that noise goes to zero, which is usually shown either by some martingale inequality or by using the law of large number of Poisson random variables, using a Poisson representation of CTMC. In Appendix D, we present a proof based on the Poisson representation.

*Remark** 4.4. In continuous transitions with random increments, we assumed for simplicity that the distribution of the increment is independent from the current state of the system. However, this restriction

can be safely dropped, provided that we require uniform boundedness (in any compact $K \subset E$) of the limit first order moments of the increments, i.e. $\sup_{\mathbf{x}\in K} \mathbb{E}[\nu(\mathbf{x})] < \infty$ and $\sup_{\mathbf{x}\in K} \mathbb{E}[\|\nu(\mathbf{x})\|] < \infty$, and uniform convergence of the expectation of $\nu^{(N)}(\mathbf{x})$ to $\nu(\mathbf{x})$, i.e. $\sup_{x\in K} \|\mathbb{E}[\nu^{(N)}(\mathbf{x})] - \mathbb{E}[\nu(\mathbf{x})]\| \to 0$ and $\sup_{x\in K} |\mathbb{E}[\|\nu^{(N)}(\mathbf{x})\|] - \mathbb{E}[\|\nu(\mathbf{x})\|]| \to 0$. Given these conditions, it is easy to check that the resulting sequence of CTMC still satisfy the conditions of [43] (restricted to a suitable compact K), hence Theorem 4.1 continues to hold.

5 Hybrid Scaling and Hybrid Fluid Limits

In this section we will introduce a scaling for transitions that cannot be approximated continuously, roughly speaking because their frequency remains constant as the population size grows. We will then prove that the sequence of normalized CTMC converges to the PDMP associated with the normalized **sCCP** model. This proof will be first given under a suitable set of restrictions (essentially, restricting to unguarded stochastic actions with generic random resets for transitions kept discrete), in order to clarify the main ingredients that guarantee convergence. In the next sections, we will remove some of these restrictions, considering more complex hybrid limits.

The first step in the construction of the hybrid limit, which coincides with the first step in constructing the **sCCP** hybrid semantics, is the separation of model variables into discrete and continuous. This step is delicate and is model-dependent, as the same model can be interpreted in different ways. For example, the client/server model of Example 2.1 can be interpreted continuously, assuming that the number of both clients and servers is increased with γ_N, or in a hybrid way, assuming that only the number of clients increases, while the number of servers remains constant. In this case, the service rate has also to be increased in order to match the larger demand. This can be justified by thinking of an increased number of cores on the same machine, in such a way that the breakdown of a server will affect all its cores. We will discuss the partitioning of variables in Remark 5.1 below, after introducing the hybrid scaling conditions.

To this end, we need to modify the conditions of Scaling 1 for continuous transitions. In particular, we need to allow the possibility of activating a transition only in a subset of discrete modes. This is enforced by guards depending only on discrete (and environment) variables.

Scaling 2 (Hybrid Continuous Scaling). A normalized **sCCP** transition $\hat{\pi} = (\hat{\mathbf{g}}_\pi(\hat{\mathbf{X}}), \hat{\mathbf{X}}' = \hat{\mathbf{X}} + \hat{\nu}_\pi^{(N)}, \hat{\lambda}_\pi^{(N)}(\hat{\mathbf{X}}))$ of a population-**sCCP** program $(\mathcal{A}, \gamma_N)$, with discrete variables $\mathbf{X}_d$, continuous variables $\mathbf{X}_c$, and environment variables $\mathbf{X}_e$, and with $\hat{\mathbf{X}} \in E$, has *hybrid continuous scaling* if and only if:

1. the rate $\hat{\lambda}_\pi^{(N)}(\hat{\mathbf{X}})$ and the update $\hat{\mathbf{X}}' = \hat{\mathbf{X}} + \hat{\nu}_\pi^{(N)}$ satisfy the same conditions of Scaling 1
2. The guard predicate $\hat{\mathbf{g}}_\pi(\hat{\mathbf{X}})$ depends only on discrete ($\mathbf{X}_d$) and environment ($\mathbf{X}_e$) variables.

Additionally, we need to define the scaling for discrete stochastic transitions. Also in this case, we will assume that their guard depends only on discrete or environment variables.

Scaling 3 (Discrete Scaling for Stochastic Transitions). A *normalized* **sCCP** transition with *random reset* $\hat{\pi} = (\hat{\mathbf{g}}_\pi(\hat{\mathbf{X}}), \hat{\mathbf{X}}' = \hat{\mathbf{r}}_\pi^{(N)}(\hat{\mathbf{X}}, \mathbf{W}^{(N)}(\hat{\mathbf{X}})), \hat{\lambda}_\pi^{(N)}(\hat{\mathbf{X}}))$ of a population-**sCCP** program $(\mathcal{A}, \gamma_N)$ with discrete variables $\mathbf{X}_d$, continuous variables $\mathbf{X}_c$, and environment variables $\mathbf{X}_e$, with $\hat{\mathbf{X}} \in E$, has *discrete scaling* if and only if:

1. the guard predicate $\hat{\mathbf{g}}_\pi(\hat{\mathbf{X}})$ depends only on discrete ($\mathbf{X}_d$) and environment ($\mathbf{X}_e$) variables;
2. $\hat{\lambda}_\pi^{(N)}(\hat{\mathbf{X}}) = O(1)$, $\hat{\lambda}_\pi^{(N)}(\hat{\mathbf{X}})$ converges uniformly in each compact $K \subset E$ to the continuous function $\hat{\lambda}_\pi(\hat{\mathbf{X}})$;
3. Resets converge weakly (uniformly on compact (sub)sets), i.e. for each $\hat{\mathbf{x}}^{(N)} \to \hat{\mathbf{x}}$ in E, $\hat{\mathbf{r}}_\pi^{(N)}(\hat{\mathbf{x}}^{(N)}, \mathbf{W}^{(N)}(\hat{\mathbf{x}}^{(N)})) \Rightarrow \hat{\mathbf{r}}_\pi(\hat{\mathbf{x}}, \mathbf{W}(\hat{\mathbf{x}}))$, as random elements on E.

Remark 5.1. The choice on how to partition variables into discrete and continuous is a crucial step. This choice is usually model dependent, and relies heavily on the knowledge and intuition of the modeller. However, as a general guideline, we can look at two aspects of the model:

Conservation Laws: Very often, the identification of discrete variables can be made by looking at conservation laws, i.e. at subsets of variables whose total mass is conserved during the evolution of the system, as pursued in [16]. In fact, conserved variables usually are related to internal states of an agent which is present in one or very few copies. The identification of these sets can be carried out using algorithms like the Fourier-Motzkin elimination procedure [25], or using a constraint based approach [57]. In **sCCP**, when describing non-flat models, these sets of variables, corresponding to state variables, are usually evident (cf. Remark 2.1).

Scaling of Rates: in describing a population-**sCCP** model, a modeller is forced to make explicit the dependence of rates on the system size γ_N. Given this knowledge, it is possible to identify some variables that cannot be continuous, otherwise both scaling 1 and 3 would be violated. For instance, if we have a rate like kX_1X_2, then at least one of X_1 and X_2 has to be discrete, otherwise the normalized rate would depend quadratically on γ_N. On the contrary, $k\gamma_N^{-1}X_1X_2$ is not compatible with both X_1 and X_2 discrete, otherwise the rate would vanish. Clearly, not all rate functions are informative; for instance, linear rates are compatible both with discrete and continuous scaling.

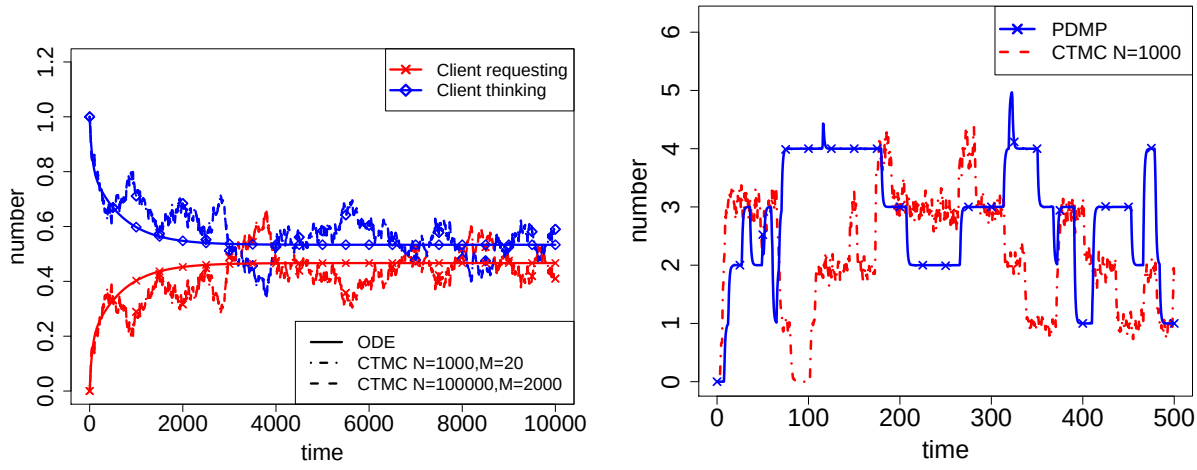
The two previous arguments can be used to set up an algorithmic procedure to suggest a possible partition of variables into discrete and continuous, given a population-**sCCP** model. However, we leave this for future work.

We stress that, in general, if the modeller does not know how rates depend on the system size, she may choose a partition of variables and a scaling for each transition and impose a dependence of rates on system size that is correct with respect to the partition. This dependence has then to be validated a-posteriori, checking if it is meaningful in the context of the model. For instance, in a practical modelling scenario for the client/server example of Section 2.1, one usually has a fixed number of clients and servers and fixed parameters. To apply the convergence results of this paper, a specific scaling has to be assumed, and the parameters of the limit model have to be computed consequently. If, for instance, the number of servers is kept fixed, we obtain a meaningful limit if the *service rate per client* is constant. If this cannot be assumed, namely if it is the global service rate of servers that remains constant, then the service rate per client depends on their number N, and goes to zero as N increases. Hence, in the limit model the service rate is zero. However, for a fixed population size, we can still obtain a hybrid process that approximates closely the CTMC, using the size-dependent rates. This phenomenology (uninformative limit, but good size-dependent approximation) happens also in the fluid limit setting, see for instance [47].

Consider now a population-**sCCP** model $(\mathcal{A}, \gamma_N)$ with only stochastic actions, in which transitions satisfy either the continuous scaling 1 or the discrete scaling 3. The *limit TDSHA* $\hat{\mathcal{T}}(\mathcal{A})$ constructed from this model has continuous transitions of the form $(\mathbb{E}[\nu_\pi], g_\pi)$, for each **sCCP** action π satisfying continuous scaling and stochastic transitions of the form $(true, \hat{\mathbf{r}}_\pi, \hat{\lambda}_\pi)$, for each **sCCP** action π satisfying the discrete scaling. The limit PDMP is obtained from this TDSHA by the construction of Section 2.5.

Example 5.1. We consider a new example with a biological flavour, namely a simple genetic network. Genes are the storage units of biological information: they encode in a string of DNA the information to produce a protein. Each cell has a biochemical machine that is capable of reading the information in a gene, first copying it into a mRNA molecule and then translating this molecule into a protein. Genes are in fact more than simple storage units: they are also part of the software that controls their own expression. In fact, expression is regulated by specific proteins, called transcription factors, which physically bind to the DNA close to a gene and activate or repress transcription. There are genes encoding for transcription factors that act as self-repressors. We model such a scenario here.

To construct a population-**sCCP** model, we need two integer-valued variables: M, counting the amount of mRNA, and P, counting the amount of protein. Here the size of the system γ is the volume times the Avogadro number, so that normalized variables represent molar concentrations (see for instance [62, 15]). We will consider a model with one agent for the gene (which can be on or off), and agents for translation of mRNA into protein and degradation of both protein and mRNA.



(a) Client Server, Example 2.1

(b) Gene Network, Example 5.1

Figure 2: Left: comparison of stochastic trajectories and fluid ODE for the client-server model of Example 2.1, with scaling discussed in Section 4. Parameters are $k_r = 2$, $k_s = 0.8$, $k_t = 1/50$, $k_b = 1/2000$, $k_f = 1/1000$, and initial conditions are $X_r^{(N)}(0) = N_1 = 100N$, $X_i(0) = N_2 = 2N$. In the plot, the CTMC trajectory for $N = 100000$ and $M = 2000$ fully overlap with the solution of the fluid ODE. Right: comparison of a trajectory of the limit PDMP and of the CTMC for the gene network model of Example 5.1. Parameters are $k_p = 0.1$, $k_t = 1$, $k_{d_p} = 1$, $k_{d_m} = 0.01$, $k_b = 0.1$, $k_u = 0.1$, and initial conditions are $P(0) = M(0) = G_{off}(0) = 0$, $G_{on}(0) = 1$. Note that both the stochastic and the hybrid system show a multi-modal behaviour.

$$
\begin{array}{lcl}
\texttt{gene\_on} & \stackrel{\text{def}}{=} & [* \to M' = M + 1]_{k_p}.\texttt{gene\_on} \\
 & + & [* \to P' = P - 1]_{k_b \gamma^{-1} P}.\texttt{gene\_off} \\
\texttt{gene\_off} & \stackrel{\text{def}}{=} & [* \to P' = P + 1]_{k_u}.\texttt{gene\_on} \\
\texttt{translate} & \stackrel{\text{def}}{=} & [* \to P' = P + 1]_{\gamma k_t M}.\texttt{translate} \\
\texttt{degrade} & \stackrel{\text{def}}{=} & [* \to M' = M - 1]_{k_{dm} M}.\texttt{degrade} \\
 & + & [* \to P' = P - 1]_{k_{dp} P}.\texttt{degrade}
\end{array}
$$

Inspecting the previous model, we can see that it is not flat. To convert it into a flat model, we need to add two additional variables, G_{on} and G_{off}, with domain $\{0, 1\}$, encoding the state of the gene agent. The structure of the gene agent itself reveals a conservation pattern in the system, namely that $G_{on} + G_{off} = 1$, as they are indicator variables of the state of the gene. Inspecting transitions, we can notice how translation has a rate depending on γ, suggesting that M has also to be treated as a discrete variable. On the other hand, repression scales as γ^{-1}, i.e. it depends on the concentration of P, rather than on the number of molecules (repression depends only on the molecules close to the gene, the only ones that can bind to it). With this partitioning of variables, we obtain the following normalized TDSHA:

- Discrete variables are G_{on}, G_{off}, M, while $\hat{P}$ is the continuous variable. $Q = \{0, 1\} \times \{0, 1\} \times \mathbb{N}$ and $\hat{P}$ has domain $[0, \infty)$.
- Continuous transitions are $(*, \hat{P}' = \hat{P} + \gamma^{-1}, \gamma k_t M)$ and $(*, \hat{P}' = \hat{P} - \gamma^{-1}, \gamma k_{dp} \hat{P})$;
- Discrete transitions are $(*, G'_{on} = 0, G'_{off} = 1, \hat{P}' = \hat{P} - \gamma^{-1}, k_b \hat{P} G_{on})$, $(*, G'_{on} = 1, G'_{off} = 0, \hat{P}' = \hat{P} + \gamma^{-1}, k_u G_{off})$, $(*, M' = M + 1, k_p G_{on})$, $(*, M' = M - 1, k_{dm} M)$.

*Remark** 5.2. Scaling 3 forbids discrete transitions to have a fast, $O(\gamma)$ rate. If this would be the case, the dynamics of discrete transitions in the limit would be faster and faster, and one would expect that

the discrete subsystem affected by these transitions reaches immediately the equilibrium (in a stochastic sense). This is what actually happens, under some regularity conditions on fast discrete dynamic, namely the possibility of isolating a discrete subsystem affected by fast discrete transitions, which is ergodic (considering only fast discrete transitions), and with fast rates depending continuously on continuous variables. In this case, one can compute the equilibrium distribution (as a function of other variables) of the fast discrete subsystem, remove the fast discrete variables and average the rate functions depending on fast discrete variables according to the equilibrium distribution. In case one has only fast discrete variables, the fluid limit is given in terms of ODE [5]. This scaling can be integrated quite easily in our framework, using the limit theorem of [5] instead of Theorem 4.1 and defining syntactically the averaging at the level of the TDSHA, given a method to compute the equilibrium distribution.

We now turn to discuss the limit behaviour of a model showing hybrid scaling, i.e. with both discrete and continuous transitions. We will stick to further simplifying assumptions for the moment: the **sCCP** program has no instantaneous transitions, all stochastic actions are unguarded and have continuous rates, variables and transitions have been partitioned into discrete and continuous, discrete transitions have deterministic resets and satisfy discrete scaling 3, and continuous transitions satisfy continuous scaling 1.

We are now ready to state the main result of this section, namely that, under these restrictions, a normalized CTMC constructed from a **sCCP** program converges weakly to the PDMP constructed from the normalized TDSHA associated with the **sCCP** program.

Theorem 5.1 ([11])**.** *Let $(\mathcal{A}, \gamma_N)$ be a sequence of population-**sCCP** models for increasing system size $\gamma_N \to \infty$, satisfying the conditions of this section, with variables partitioned into discrete $\mathbf{X}_d$, continuous $\mathbf{X}_c$, and environment ones $\mathbf{X}_e$. Assume that discrete actions satisfy scaling 3 and continuous actions satisfy scaling 2. Let $\hat{\mathbf{X}}^{(N)}(t)$ be the sequence of normalized CTMC associated with the **sCCP** program and $\hat{\mathbf{x}}(t)$ be the PDMP associated with the limit normalized TDSHA $\hat{\mathcal{T}}(\mathcal{A})$.*

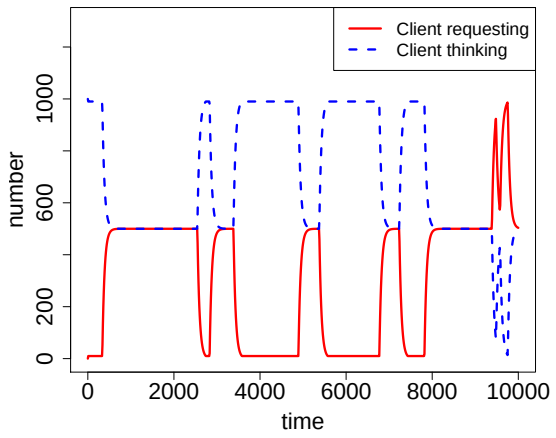
If $\hat{\mathbf{x}}_0^{(N)} \Rightarrow \hat{\mathbf{x}}_0$ (weakly) and the PDMP is non-Zeno, then $\hat{\mathbf{X}}(t)$ converges weakly to $\hat{\mathbf{x}}(t)$, $\hat{\mathbf{X}} \Rightarrow \hat{\mathbf{x}}$, as random elements in the space of cadlag function with the Skorohod metric.[9]

Proof. We just sketch the proof here. A detailed proof can be found in Appendix D. The main idea is to exploit the fact that we can restrict our attention to CTMC and PDMP that do at most m discrete jumps. This is sufficient to obtain the weak convergence of the full processes, for two reasons. The first is related to the nature of the Skorohod metrics, which discounts the future (i.e. only $1/2^T$ of the distance comes from time instants greater than T), while the second is the non-Zeno nature of the limit PDMP, which implies that we can consider no more than m jumps up to time T, with probability $1 - \varepsilon_m$, for $\varepsilon_m \to 0$ as $m \to \infty$.

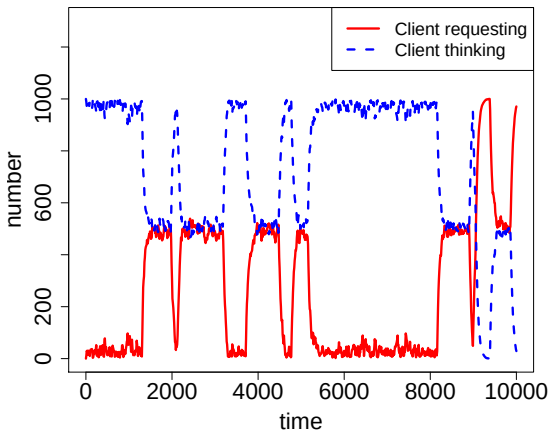
In order to prove weak convergence of $\hat{\mathbf{X}}_m^{(N)}$, the CTMC with at most m jumps of discrete transitions, to $\hat{\mathbf{x}}_m$, the PDMP with at most m jumps, we can exploit the piecewise deterministic nature of PDMP, applying Theorem 4.1 inductively. At the first step, we will prove that the time $\tau_1^{(N)}$ of the first stochastic jump for $\hat{\mathbf{X}}^{(N)}$ converges weakly to τ_1, the first jump time of $\hat{\mathbf{x}}$ (Lemma D.2 in Appendix D), and also the state $\hat{\mathbf{X}}^{(N)}(\tau_1^{(N)})$ after time $\tau_1^{(N)}$ converges weakly to $\hat{\mathbf{x}}(\tau_1)$ (Corollary D.1). This shows convergence of the processes up to the first stochastic jump. Exploiting this and the strong Markov property, we can restart $\hat{\mathbf{x}}(t)$ at time τ_1 from $\hat{\mathbf{x}}(\tau_1)$ and $\hat{\mathbf{X}}(t)$ from $\hat{\mathbf{X}}(\tau_1^{(N)})$ at time $\tau_1^{(N)}$ and apply Theorem 4.1 and its corollaries again (actually, a minor modification of Theorem 4.1, allowing to sample probabilistically the initial conditions of the ODE), to prove weak convergence of the CTMC to the PDMP up to the m-th jump, for any m. Note that this argument is based on the continuity of vector fields, rates and resets, which holds in our setting as their guards depend only on discrete and environment variables, hence their values do not change in each deterministic piece of the PDMP dynamics. □

Example. We consider again the simple client server network of Example 2.1, but with a different scaling compared to Section 4. In particular, we consider as size γ_N the number of clients, assuming that the number of servers remains constant, but with service rate depending linearly on γ_N. In this way, the rate of the request transition of `client` agents is $\gamma_N \min\{k_r \hat{X}_r^{(N)}, k_s X_i\}$, and it satisfies the continuous scaling. Breakdown and repair transitions, on the other hand, will be kept discrete as they modify only the number of available servers. As their rate is independent of γ_N and their reset is constant and also independent of N, they both clearly satisfy the discrete scaling. The limit TDSHA that we obtain in this way is shown

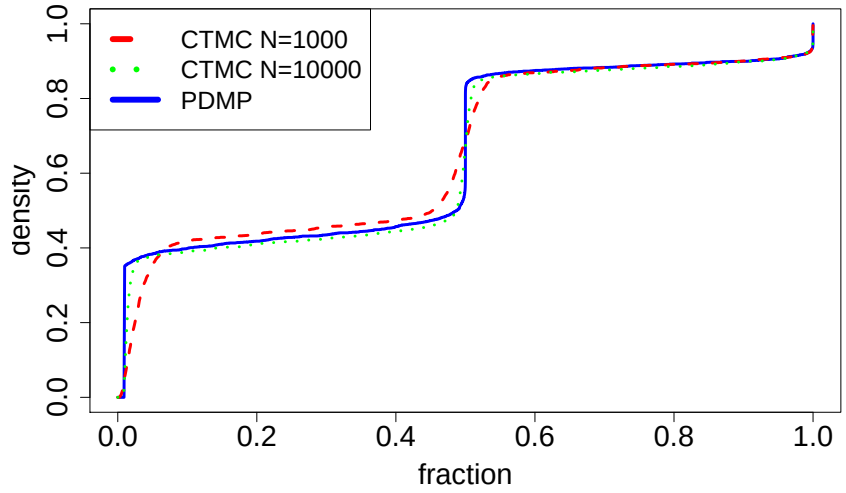
[9] See Appendix C for a breif introduction of these concepts.



(a) PDMP trajectory



(b) CTMC trajectory, $N = 1000$



(c) $t = 10000$, PDMP and CTMC cumulative distributions

Figure 3: Client Server model of Example 2.1, compared with the hybrid limit scaling keeping the number of server fixed to 2. Parameters are $k_r = 2$, $k_s = 0.01$, $k_t = 1/50$, $k_b = 1/2000$, $k_f = 1/1000$, and initial conditions are $X_r^{(N)}(0) = N$, $X_i(0) = 2$. Figures 3(a) and 3(b) show one trajectory of the PDMP and the CTMC for $N = 1000$, respectively. Figure 3(c), instead, compares the empirical cumulative distribution of the PDMP limit and the CTMC, for $N = 1000$ and $N = 10000$, at time $t = 10000$, generated from 2500 sampled trajectories.

in Figure 1. As the hypotheses of Theorem 5.1 are satisfied, the sequence of CTMC models obtained from **sCCP** with the standard stochastic semantics converges (weakly) to the limit TDSHA.

This can be seen in Figure 3, where we compare a trajectory of the CTMC with a trajectory of the PDMP, and the distribution of the number of clients requesting service at time $t = 10000$.

Example. We reconsider now the genetic network model of Example 5.1. Also in this case, we can expect a bimodal behaviour for the CTMC semantics, due to the gene working as a discrete switch. If the binding strength of the repressor is large, meaning that the protein will remain bound to the gene for a long time, then the gene will be switched off for long periods, and we may expect to see a bursty behaviour. This is indeed the case, as can be seen in Figure 2(b). Moreover, the hybrid limit constructed in Example 5.1 matches perfectly this behaviour, as can be seen in Figure 2(b). As the model satisfies the (scaling) assumptions of Theorem 5.1, we can conclude that this is indeed the consequence of the (weak) convergence of the sequence of CTMC models to the hybrid limit.

5.1 More on random resets

The scaling condition 3 requires us to check a convergence condition on resets that seems quite complicated at first glance, as it involves checking weak convergence of reset kernels for any possible convergent sequence of states. We chose this condition because it is very general and it interfaces smoothly with the inductive proof technique that we use in the paper. However, in the following, we will briefly discuss several simpler conditions that can be checked more easily, and that should cover most practical cases.

We first start by observing that we can split the convergence condition in two parts, i.e. we can check that $\hat{\mathbf{r}}_\pi^{(N)}(\hat{\mathbf{x}}, \mathbf{w}) \to \hat{\mathbf{r}}_\pi(\hat{\mathbf{x}}, \mathbf{w})$ uniformly in $\hat{\mathbf{x}}$ and $\mathbf{w}$ and that $\mathbf{W}_\pi^{(N)}(\hat{\mathbf{x}}^{(N)}) \Rightarrow \mathbf{W}_\pi(\hat{\mathbf{x}})$ (weakly), for any $\hat{\mathbf{x}}^{(N)} \to \hat{\mathbf{x}}$.

We now focus attention on the weak convergence of random elements $\mathbf{W}^{(N)}(\hat{\mathbf{x}})$ to $\mathbf{W}(\hat{\mathbf{x}})$. First, note that the weak convergence condition is essentially equivalent to showing that

$$\sup_{\hat{\mathbf{x}} \in K} \| \int_E g(\hat{\mathbf{y}}) \mathbb{P}\{\mathbf{W}_\pi^{(N)}(\hat{\mathbf{x}}) = \hat{\mathbf{y}}\} \, \mathrm{d}\hat{\mathbf{y}} - \int_E g(\hat{\mathbf{y}}) \mathbb{P}\{\mathbf{W}_\pi(\hat{\mathbf{x}}) = \hat{\mathbf{y}}\} \, \mathrm{d}\hat{\mathbf{y}} \| \to 0,$$

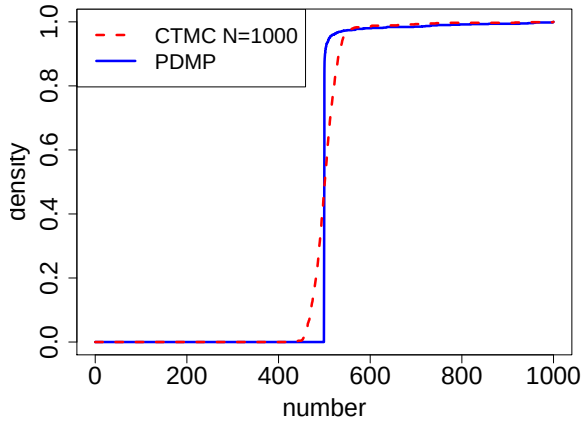
for any compact set $K \subseteq E$ and any uniformly continuous function $g : E \to \mathbb{R}$ and that $\int_E g(\hat{\mathbf{y}}) \mathbb{P}\{\mathbf{W}_\pi(\hat{\mathbf{x}}) = \hat{\mathbf{y}}\} \, \mathrm{d}\hat{\mathbf{y}}$ is a continuous function [40], which may be sometimes easier to check. Moreover, in practice we can expect $\mathbf{W}_\pi^{(N)}$ and $\mathbf{W}_\pi$ to have a simple structure, which should facilitate the task of verifying the scaling condition.

First of all, if $\mathbf{W}_\pi^{(N)}$ and $\mathbf{W}_\pi$ do not depend on $\hat{\mathbf{x}}$, then the condition reduces to $\mathbf{W}_\pi^{(N)} \Rightarrow \mathbf{W}_\pi$, which can be checked by showing one of the equivalent conditions of the Portmanteau theorem [9]. In particular, the condition is trivially true if $\mathbf{W}_\pi^{(N)}$ does not depend on N, i.e. if $\mathbf{W}_\pi^{(N)} = \mathbf{W}_\pi$.

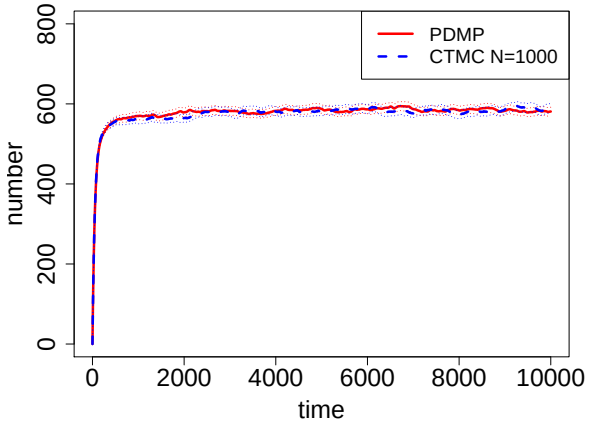
We consider now two examples, to illustrate the use of random resets and the hybrid convergence in this case.

Example 5.2. We consider a small variation of the client server model of Example 2.1. The difference is that we will assume different levels of severity of a breakdown, so that the repair time can be variable, depending on this level. For simplicity, we assume a single server, but a generalization to more than one server is straightforward. In order to model this situation in **sCCP**, we can either increase the number of internal states of the server (one for each level of damage) or use an additional (discrete) variable. We chose this second approach, introducing D, the damage-level variable. We assume that D takes values on the integers, and that each time a breakdown happens, its value is sampled from a geometric distribution with parameter 0.5, so that we have a probability $1/2^k$ to see a damage of level k. We therefore let $W \sim Geom(0.5)$, so that $\mathbb{P}\{W = k\} = (1 - 0.5)^{k-1} \cdot 0.5$. We further assume that the repair time is proportional to the damage level, so that the rate of repair is k_f/D. We therefore obtain the following **sCCP** code, where variables X_r, X_t, X_i, X_b are as in Example 2.1:

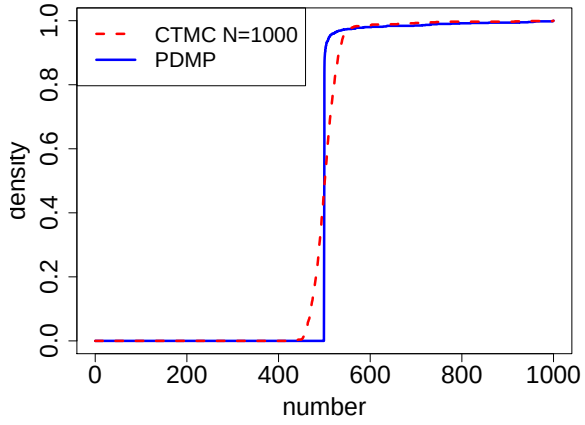
$$\begin{array}{lcl}
\texttt{client} & \stackrel{\text{def}}{=} & [* \to X_r' = X_r - 1 \wedge X_t' = X_t + 1]_{\min\{k_r X_r, \gamma_N k_s X_i\}} \texttt{.client +} \\
 & & [* \to X_r' = X_r + 1 \wedge X_t' = X_t - 1]_{k_t X_t} \texttt{.client} \\
\texttt{server} & \stackrel{\text{def}}{=} & [* \to X_i' = X_i - 1 \wedge X_b' = X_b + 1, D' = W]_{k_b X_i} \texttt{.server} \\
 & \texttt{+} & [* \to X_i' = X_i + 1 \wedge X_b' = X_b - 1]_{k_f / D \cdot X_b} \texttt{.server}
\end{array}$$



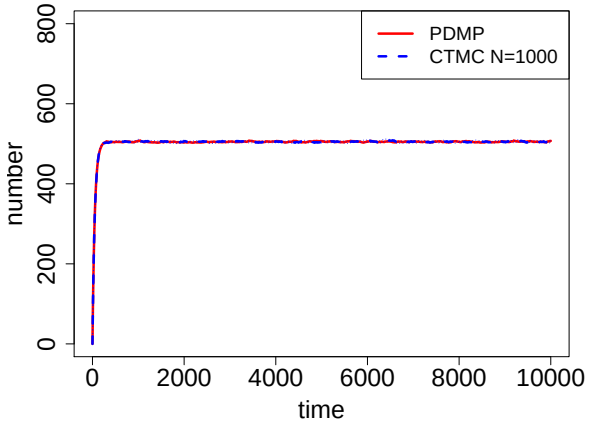
(a) Geometric Breakdown: $t = 10000$



(b) Geometric Breakdown: average



(c) Lognormal Breakdown: $t = 10000$



(d) Lognormal Breakdown: average

Figure 4: Empirical cumulative distribution of clients requesting service at time $t = 10000$ and average number of clients request service for the client server models of Example 5.2. The top row shows the model with severity level of breakdowns samples according to a geometric distribution with probability $p = 0.5$. Parameters are as in caption of Figure 3, a part from $k_f = 1/200$. The bottom row shows the model with fix rate lognormally distributed with mean -2.5 and standard deviation 1.0. Note that both histograms present a similar pattern. The bimodality of the distribution is captured for the geometric breakdown. Moreover, the hybrid model has less variability in the distribution (it has a sharper cumulative distribution function). The averages are almost indistinguishable.

In this case, we clearly have that X_i, X_b, and D are discrete variables, while X_r and X_t can be approximated continuously. D can also be seen as an environment variable, as it is used to modify a parameter of the model. Therefore, the transitions of the client agent become continuous transitions in the associated TDSHA, while the transitions of the server agent remain discrete and stochastic. Note that we made explicit the dependence on size in the rate functions. Clearly, all transitions satisfy the scalings of Theorem 5.1. This is true also for the breakdown transition, as W does not depend on the current state of the system. It follows that Theorem 5.1 applies also to this example, see also Figures 4(a) and 4(b).

A variation of this model is to replace the finite damage levels with a continuous level of damage, essentially sampling the repair rate from a continuous distribution. This can be done in **sCCP** by using a real-valued environment variable, call it K. For simplicity, here we assume that the fixing rate is sampled from a lognormal distribution with mean μ and standard deviation σ. We can obtain this variant of the model by replacing the `server` agent with the following one:

$$\begin{array}{rcl} \texttt{server} & \stackrel{\text{def}}{=} & [* \to X_i' = X_i - 1 \wedge X_b' = X_b + 1, K' = W]_{k_b X_i}\,.\,\texttt{server} \\ & + & [* \to X_i' = X_i + 1 \wedge X_b' = X_b - 1]_{K \cdot X_b}\,.\,\texttt{server} \end{array}$$

Also in this case, the hypotheses of Theorem 5.1 are satisfied, and convergence to the hybrid limit works (see Figures 4(c) and 4(d)).

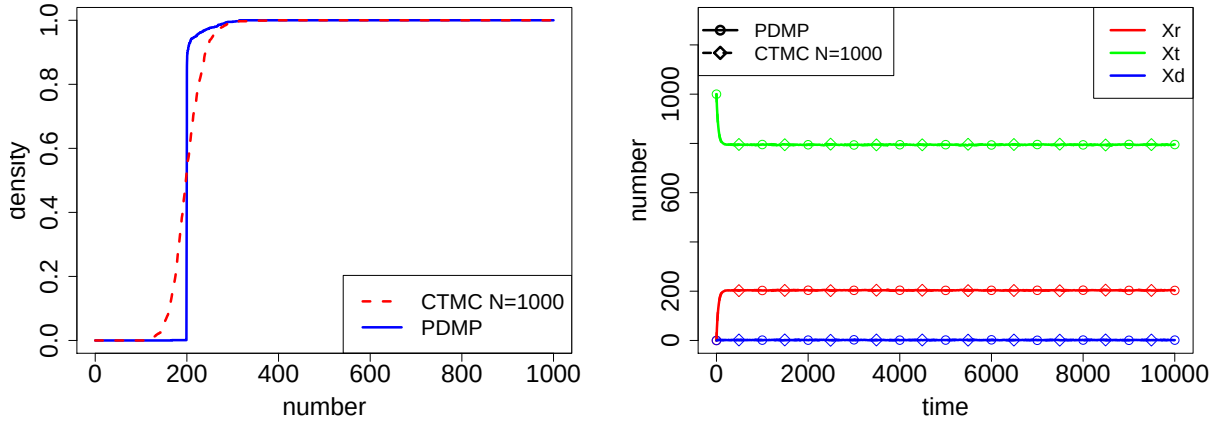
We turn now to discuss convergence of $\mathbf{W}_\pi^{(N)}(\hat{\mathbf{x}})$ to $\mathbf{W}_\pi(\hat{\mathbf{x}})$ when they depend on $\hat{\mathbf{x}}$. The situation is more delicate, as convergence has to be uniform. In the following, however, we list some sufficient conditions to guarantee convergence, that are of practical relevance.

1. $\mathbf{W}_\pi^{(N)}(\hat{\mathbf{x}}) = \mathbf{W}_\pi(\hat{\mathbf{x}})$ and $\mathbf{W}_\pi(\hat{\mathbf{x}})$ depends continuously on $\hat{\mathbf{x}}$;
2. $\mathbf{W}_\pi^{(N)}(\hat{\mathbf{x}})$ and $\mathbf{W}_\pi(\hat{\mathbf{x}})$ are discrete distributions with mass concentrated on points $\{\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_k, \ldots\}$, and $\mathbb{P}\{\mathbf{W}_\pi^{(N)}(\hat{\mathbf{x}}) = \hat{\mathbf{y}}_k\}$ converges to $\mathbb{P}\{\mathbf{W}_\pi(\hat{\mathbf{x}}) = \hat{\mathbf{y}}_k\}$ uniformly in any compact $K \subseteq E$;
3. $\mathbf{W}^{(N)}(\hat{\mathbf{x}})$ and $\mathbf{W}(\hat{\mathbf{x}})$ are unidimensional real random variables, with cumulative distribution functions $F^{(N)}(y, \hat{\mathbf{x}})$ and $F(y, \hat{\mathbf{x}})$, such that, for each $\hat{\mathbf{x}}^{(N)} \to \hat{\mathbf{x}}$, $F^{(N)}(y, \hat{\mathbf{x}}^{(N)}) \to F(y, \hat{\mathbf{x}})$ pointwise for any continuity point y of $F(y, \hat{\mathbf{x}})$.
4. $\mathbf{W}^{(N)}(\hat{\mathbf{x}})$ and $\mathbf{W}(\hat{\mathbf{x}})$ have values in $\mathbb{R}^h$, and they have continuous density functions $g^{(N)}(\mathbf{y}, \hat{\mathbf{x}})$, $g(\mathbf{y}, \hat{\mathbf{x}})$, and $\sup_{\mathbf{y} \in \mathbb{R}^k, \hat{\mathbf{x}} \in K} \|g^{(N)}(\mathbf{y}, \hat{\mathbf{x}}) - g(\mathbf{y}, \hat{\mathbf{x}})\| \to 0$, for each compact set $K \subseteq E$.
5. $\mathbf{W}^{(N)}(\hat{\mathbf{x}})$ and $\mathbf{W}(\hat{\mathbf{x}})$ can be decomposed into the product of marginal and conditional distributions that converge in the sense of Scaling 3, i.e. $\mathbf{W}(\hat{\mathbf{x}}) = \mathbf{W}_{i_1}(\hat{\mathbf{x}})\mathbf{W}_{i_2}(\hat{\mathbf{x}}, \mathbf{w}_{i_1}) \cdots \mathbf{W}_{i_k}(\hat{\mathbf{x}}, \mathbf{w}_{i_1}, \ldots, \mathbf{w}_{i_{n-1}})$, $\mathbf{W}^{(N)}(\hat{\mathbf{x}}) = \mathbf{W}_{i_1}^{(N)}(\hat{\mathbf{x}})\mathbf{W}_{i_2}^{(N)}(\hat{\mathbf{x}}, \mathbf{w}_{i_1}) \cdots \mathbf{W}_{i_k}^{(N)}(\hat{\mathbf{x}}, \mathbf{w}_{i_1}, \ldots, \mathbf{w}_{i_{n-1}})$, and $\mathbf{W}_{i_j}^{(N)}(\hat{\mathbf{x}}^{(N)}, \mathbf{w}^{(N)}) \Rightarrow \mathbf{W}_{i_j}(\hat{\mathbf{x}}, \mathbf{w})$, as $\hat{\mathbf{x}}^{(N)} \to \hat{\mathbf{x}}$ and $\mathbf{w}^{(N)} \to \mathbf{w}$.
6. $\mathbf{W}^{(N)}(\hat{\mathbf{x}})$ and $\mathbf{W}(\hat{\mathbf{x}})$ are mixtures of distributions $\mathbf{W}_j^{(N)}(\hat{\mathbf{x}})$ and $\mathbf{W}_j(\hat{\mathbf{x}})$ of one of the previous types, i.e. $\mathbf{W}^{(N)}(\hat{\mathbf{x}}) = \sum_j p_j^{(N)}(\hat{\mathbf{x}})\mathbf{W}_j^{(N)}(\hat{\mathbf{x}})$ and $\mathbf{W}(\hat{\mathbf{x}}) = \sum_j p_j(\hat{\mathbf{x}})\mathbf{W}_j(\hat{\mathbf{x}})$.

It is straightforward to show that each of these conditions implies that $\mathbf{W}^{(N)}(\hat{\mathbf{x}}^{(N)}) \Rightarrow \mathbf{W}(\hat{\mathbf{x}})$ as $\hat{\mathbf{x}}^{(N)} \to \hat{\mathbf{x}}$, hence they can be used whenever it is more appropriate.

Example 5.3. We consider again the client-server model of Example 2.1, but modify it by including the spread of a worm epidemic. We consider a situation in which a worm has spread on the network and activates on a specific date, sending all infected clients into a non-working state, called X_d, from which they need some time to recover. We abstract from the epidemic spreading and model the effect of the epidemics as an event that affects synchronously all clients and infects each of them with probability p. Let $W_i(X)$, $i = 1, 2$, be binomial distributions with success probability p and size given by X. For simplicity, we ignore the breakdown and repair of servers, so that we need four variables, X_r, X_t, X_d, and X_i, and initial network `client` $\|$ `worm`, where `client` is as in Example 2.1, while `worm` is given by the following code:

$$\begin{array}{rcl} \texttt{worm} & \stackrel{\text{def}}{=} & [* \to X_r' = X_r - W_1(X_r) \wedge X_t' = X_t - W_2(X_t) \wedge X_d' = X_d + W_1(X_r) + W_2(X_t)]_{k_w}\,.\,\texttt{worm} \\ & + & [* \to X_d' = X_d - 1 \wedge X_r' = X_r + 1]_{k_d \cdot X_d}\,.\,\texttt{worm} \end{array}$$



(a) Example 5.3, clients requesting at time $t = 10000$

(b) Example 5.3, average

Figure 5: Comparison of the empirical cumulative distribution of clients requesting service at $t = 10000$ (left) and average (right) of the limit PDMP and the CTMC at system size 1000, for the client-server model with worm infection of Example 5.3. There are 2 servers and $\gamma_N = N$ clients. Parameters are $k_s = 0.01\gamma_N$, $k_r = 2$, $k_t = 1/40$, $k_w = 1/2000$, $w_d = 0.1$, and the infection probability is $p = 0.33$.

In the limit TDSHA model, the infection action remains discrete and stochastic, while all others are approximated continuously (including the recovery). Here, the system size is clearly the number of clients (server dynamics are ignored, so the number of servers can be seen as a parameter). When looking at the normalized model for system size $\gamma_N = N$, then the reset of the infection transition, say for what concerns clients thinking, is $\hat{x}_t - \frac{1}{N}W(\lfloor N\hat{x}_t \rfloor)$, which can be also written as $\hat{x}_t - \frac{\lfloor N\hat{x}_t \rfloor}{N}\frac{1}{\lfloor N\hat{x}_t \rfloor}W(\lfloor N\hat{x}_t \rfloor)$, provided $\lfloor N\hat{x}_t \rfloor > 0$. By the law of large numbers, this expression converges to $\hat{x}_t - p\hat{x}_t$, so this should be the reset of the limit PDMP. However, to apply the limit results of this section to this model, we have to prove that $\frac{1}{N}W(\lfloor N\hat{x}^{(N)} \rfloor) \to \hat{x}p$ for any $\hat{x}^{(N)} \to \hat{x}$ and then apply point 3 of proposition above. To show this, observe that if $\hat{x} > 0$, then $\hat{x}^{(N)} > \hat{x}/2$ ultimately, hence $\lfloor N\hat{x}^{(N)} \rfloor \to \infty$, so that $\frac{1}{\lfloor N\hat{x}_t \rfloor}W(\lfloor N\hat{x}_t \rfloor) \to p$. When $\hat{x} = 0$, instead, observe that $\frac{1}{N}W(\lfloor N\hat{x}^{(N)} \rfloor) \leq \frac{\lfloor N\hat{x}^{(N)} \rfloor}{N} \to 0$, which shows the desired convergence.

*Remark** 5.3. The framework of population-**sCCP** programs forces the modeller to explicitly consider the notion of system size and to incorporate it in the rate functions. This requirement greatly simplifies the manual verification of the scaling conditions, at least for what concerns rate functions.

There are three kinds of conditions to check: convergence of rate functions, regularity of rate functions (local Lipschitzness), and convergence of reset kernels (or of increments).

Most of the time, these checks are easy to carry out: rates are often density dependent and differentiable and resets are constant increment updates. If rates depend on γ_N, usually this dependence is simple and verifying convergence poses no challenges. For instance, in a biochemical system, the (normalized) mass action rate when two molecules of the same kind react together has the form $k\hat{x}(\hat{x} - \frac{1}{\gamma_N})$, which is easily seen to converge uniformly in any compact set (i.e., whenever $\hat{x}$ is bounded). As for the regularity of rates, most of the time we will deal with functions constructed by algebraic operations, plus some other function like the exponential or the logarithm. All these functions are analytic [41], hence locally Lipschitz. Also the use of minimum or maximum preserves this property. What can be more challenging is the case in which resets have a stochastic part depending on the current state of the model. However, the conditions discussed in this section should cover most of the practical cases. Indeed, we can expect in most models the use within resets of simple discrete or continuous distributions, like Gaussian or uniform ones.

What is undoubtedly more challenging is to make this check automatic. This is partly due to the generality of **sCCP** as a modelling language, which allows a user to express very complex rates and updates.

Hence, a malicious user can construct models that are very complicated to check. However, in most practical cases it may be possible to set up automatic routines that verify the scaling, by clever use of computer algebra systems.

Another alternative is to identify a library of functions (for both rates and resets) which are guaranteed to satisfy the regularity and scaling conditions. This is what happens in the process algebra PEPA [60], where the syntactic-derived restrictions on the possible set of rate functions and updates guarantee that the conditions of the fluid approximation theorem (Theorem 4.1) are always satisfied. Constructing a library of "good" functions restricts the expressive power of the language, but should be enough to cover most practical modelling activity. Furthermore, libraries can be extended when needed, and the user can also use additional functions, if she also provides a "certificate of correctness". We will pursue this line of investigation in the future, with the implementation of the framework in mind.

6 Dealing with instantaneous transitions

In this section we discuss convergence to the hybrid limit in presence of instantaneous events. These events remain discrete also in the limit process and can introduce a discontinuity in the dynamics that is triggered as soon as their guard becomes true. The class of limit PDMP obtained in this way is more difficult to deal with than PDMP with just stochastic jumps. In fact, we cannot rely any more on the "smoothing" action in time of a continuous probability distribution like the exponential, but we need to track precisely the times at which instantaneous events happen. In particular, there can be time instants in which we can observe a jump in the limit process with probability greater than zero. This is particularly the case when the hybrid limit is a deterministic process, i.e. a process without discrete stochastic transitions and random resets.

From the point of view of weak convergence, dealing with instantaneous transitions requires us to prove that their firing times in the sequence of CTMC models converge to the firing time in the hybrid limit model. Furthermore, we need prove also convergence of the state after the reset. As we will see, both properties are not guaranteed to always hold. The problem resides in the intrinsic discontinuous nature of the exit times and resets on the activation region of guards. Thus, to prove convergence, we need to impose further regularity conditions on the PDMP, forcing its dynamics to avoid these discontinuous regions (with probability 1). We will start by discussing the issues with the exit time, then turn to reset kernels, and finally move to the limit theorem. After having discussed examples, we will consider a small extension of **sCCP**, allowing guards to depend on (simulation) time, and discuss limit theorems for this extended class of models.

Convergence of Exit Times

Convergence of exit times does not hold in general. Focussing on a deterministic trait of the PDMP dynamics, the problem is created by trajectories of the vector field that activate a guard by touching *tangentially* its boundary surface, as shown in Figure 6(a). In fact, for N large enough trajectories of the CTMC are contained in a small flow tube around this solution, hence some of them can cross the surface, while others may miss it. Another class of trajectories that creates problems is that of trajectories remaining in the boundary of the activation region of the guard (i.e., the discontinuity surface of the guard predicate) for a non-negligible amount of time, say for the time interval $[t_1, t_2]$. Here, the problem is that a CTMC trajectory can activate the guard in any time instant between t_1 and t_2.

However, convergence holds for trajectories of the vector field which transversally cross the discontinuity surface of a guard predicate, meaning that they intersect the surface at time t and enter in the interior of the region in which the guard is true just after t. Fortunately, this is the situation we are more likely to find in practice.

Now we prove a result about convergence of exit times that can be applied to the setting of Section 4. This result requires that almost surely only transversal crossings occur. We will then extend the hybrid limit theorem imposing this condition. We postpone discussion about how to check and/or enforce such a condition until later in the section.

We need some preliminary definitions. The first logical step that we need is to move from predicates to continuous functions in the definition of a guard.

Definition 6.1. Let $\hat{\mathbf{g}}(\hat{\mathbf{x}})$ be a guard predicate with closed activation region. A function $h : E \to \mathbb{R}$ is an *activation function* or a *guard function* for $\hat{\mathbf{g}}$ if it is a continuous function, and if the sets of points $\{\hat{\mathbf{x}} \mid h(\hat{\mathbf{x}}) \geq 0\}$ defines the activation region of $\hat{\mathbf{g}}$: $\hat{\mathbf{g}}(\hat{\mathbf{x}})$ is true if and only if $h(\hat{\mathbf{x}}) \geq 0$.
The function h is a *robust activation function* for $\hat{\mathbf{g}}$ if and only if $\partial\{\hat{\mathbf{x}} \mid h(\hat{\mathbf{x}}) < 0\} = \partial\{\hat{\mathbf{x}} \mid h(\hat{\mathbf{x}}) > 0\} = \{\hat{\mathbf{x}} \mid h(\hat{\mathbf{x}}) = 0\}$.

The notion of robust activation function essentially guarantees that the interior of the set in which $\hat{\mathbf{g}}$ is true is $\{\hat{\mathbf{x}} \mid h(\hat{\mathbf{x}}) > 0\}$, so that it makes sense to define a transversal crossing of the guard $\hat{\mathbf{g}}$ as a change of sign of the function h. The discontinuity surface of the guard is therefore $\mathcal{H} = \{\hat{\mathbf{x}} \mid h(\hat{\mathbf{x}}) = 0\}$.

The notion of robust activation is a very reasonable assumption to make. In practical cases, the function h will be a piecewise smooth function, usually piecewise linear, and the surface $\mathcal{H}$ will be a union of differentiable (or even analytic) manifolds of dimension $n-1$ or less. In particular, a function like $h(\hat{\mathbf{x}}) = x_1$ if $x_1 \leq 0$ and $h(\hat{\mathbf{x}}) = 0$ if $x_1 > 0$ is forbidden. This function is a bad activation function as its discontinuity surface $\mathcal{H}$ is the whole half-space $\{x_1 \geq 0\}$, and furthermore the sequence of functions $h^{(N)}(\hat{\mathbf{x}}) = h(\hat{\mathbf{x}}) - \frac{1}{N}$ converges uniformly to h but their activation set is empty for all N. This justifies the use of the term robust: we are forbidding functions which under any small perturbation induce a discontinuous change in the activation set. From now on, we restrict our attention to robust activation functions.

Coming back to PDMP derived from TDSHA, it is easy to see how to construct a guard function for the class of guards of instantaneous transitions. In fact, we are considering only positive boolean combinations of atoms of the form $h_i(\hat{\mathbf{x}}) \geq 0$, where h_i is continuous. Then, we just need to combine the functions h_i with maximum and minimum to take into account the structure of boolean combinators. Furthermore, if in the TDSHA we have k instantaneous transitions $\pi_1, \ldots, \pi_k$, with activation functions $h_1, \ldots, h_k$, we can combine them into a unique activation function h by taking their maximum: $h(\hat{\mathbf{x}}) = \max\{h_1(\hat{\mathbf{x}}), \ldots, h_k(\hat{\mathbf{x}})\}$. The function is a robust activation function which is greater than or equal to zero if and only if at least one guard is true. We call h the *activation function of the PDMP*.

Consider now a continuous trajectory $\hat{\mathbf{x}} : [0, \infty) \to E$ and let $h : E \to \mathbb{R}$ be a robust activation function, such that $h(\hat{\mathbf{x}}(0)) < 0$.

Definition 6.2. The robust activation function h (or the corresponding activation surface $\mathcal{H}$) is *transversal* for the trajectory $\hat{\mathbf{x}}(t)$ if and only if, letting $\zeta = \inf\{t \mid h(\hat{\mathbf{x}}(t)) \geq 0\}$, there is a $\delta > 0$ such that $h(t) > 0$ for $t \in (\zeta, \zeta + \delta]$.

Suppose now $\hat{\mathbf{X}}(t)$ is a stochastic process with almost surely continuous trajectories, like the fluid limit $\hat{\mathbf{x}}(t)$, with initial conditions (drawn from a distribution) $\hat{\mathbf{x}}_0$.

Definition 6.3. An activation function h is *robustly transversal* to $\hat{\mathbf{X}}(t)$ if and only if the set of trajectories for which it is transversal has probability 1.

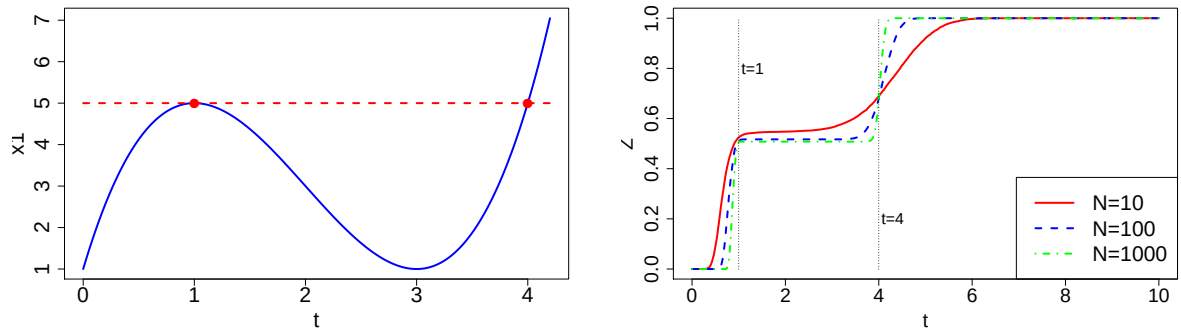
The notion of robustly transversal activation function can be lifted to PDMP, by requiring that all the guards of the PDMP are *robustly transversal* in each continuous trait of the dynamics,[10] i.e. that instantaneous transitions are activated transversally:

Definition 6.4. A PDMP $\hat{\mathbf{x}}(t)$ is robustly transversal if and only if with probability one its trajectories are *robustly transversal* in each continuous trait.

Consider now a sequence $\hat{\mathbf{X}}^{(N)}(t)$ of normalized CTMC associated with a sequence $(\mathcal{A}, \gamma_N)$ of population-**sCCP** models, and assume that $\hat{\mathbf{X}}^{(N)}(t)$ converges weakly to $\hat{\mathbf{X}}(t)$ as $N \to \infty$, where $\hat{\mathbf{X}}$ is a.s. continuous. Furthermore, let $h, h^{(N)} : E \to \mathbb{R}$ be the activation functions for $\hat{\mathbf{X}}$ and $\hat{\mathbf{X}}^{(N)}(t)$, respectively. Assume that $h^{(N)}$ converges to h uniformly for each compact set $K \subseteq E$ and call $\zeta^{(N)} = \inf\{t \mid h^{(N)}(\hat{\mathbf{X}}^{(N)}(t)) \geq 0\}$.

We can show the following lemma, whose proof is given in Appendix D.

[10] This means that, if the i-th jump of a PDMP trajectory $\hat{\mathbf{x}}(t)$, happening at time T_i, corresponds to an instantaneous transition, then there is a $\delta > 0$ such that by extending the continuous trajectory starting at $\hat{\mathbf{x}}(T_{i-1}^+ =$ up to $T_i + \delta$, the crossing is transversal, i.e. $h(\hat{\mathbf{x}}(t)) > 0$ for $t \in (T_i, T_i + \delta)$.



(a) $\hat{x}_1$ trajectory

(b) $\mathbb{P}\{Z^{(N)}(t) = 1\}$

Figure 6: Left: limit trajectory of $\hat{x}_1(t)$ for Example 6.1. Right: $\mathbb{P}\{Z^{(N)}(t) = 1\}$ as a function of t, for different values of N. Bi-modality of the distribution around $t = 1$ and $t = 4$ is manifest.

Lemma 6.1. *Let $(\mathcal{A}, \gamma_N)$ be a sequence of population-**sCCP** models for increasing population size $\gamma_N \to \infty$, as $N \to \infty$. Let $\hat{\mathbf{X}}^{(N)}(t)$ be the associated normalized sequence of CTMC, and suppose $\hat{\mathbf{X}}^{(N)} \Rightarrow \hat{\mathbf{X}}$, where $\hat{\mathbf{X}}$ has a.s. continuous sample paths. Let $h^{(N)}$, h be activation functions for $\hat{\mathbf{X}}^{(N)}$ and $\hat{\mathbf{X}}$, such that $h^{(N)} \to h$ uniformly, and suppose h is* robustly transversal *to $\hat{\mathbf{X}}$. Then $\zeta^{(N)} \Rightarrow \zeta$.* □

Example 6.1. We discuss now a hand-crafted example to demonstrate the need for the request of transversal activation of a guard. We consider a population-**sCCP** program $(\mathcal{A}, \gamma_N)$ with three continuous variables X_1, X_2, and X_3 taking values in $\mathbb{Z}$, and one discrete variable Z.

$$
\begin{array}{lcl}
\texttt{agent1} & \stackrel{\text{def}}{=} & [X_2 \geq 0 \to X_1' = X_1 + 1]_{X_2}\texttt{.agent1} \\
& + & [X_2 < 0 \to X_1' = X_1 - 1]_{|X_2|}\texttt{.agent1} \\
\texttt{agent2} & \stackrel{\text{def}}{=} & [* \to X_2' = X_2 + 1]_{X_3}\texttt{.agent2} \\
& + & [* \to X_2' = X_2 - 1]_{12\gamma_N}\texttt{.agent2} \\
\texttt{agent3} & \stackrel{\text{def}}{=} & [* \to X_3' = X_3 + 1]_{6\gamma_N}\texttt{.agent3} \\
\texttt{doom} & \stackrel{\text{def}}{=} & [X_1 \geq 5\gamma_N \to Z = 1]_{\infty:1}\texttt{.0}
\end{array}
$$

The initial network is $\texttt{agent1} \parallel \texttt{agent2} \parallel \texttt{agent3} \parallel \texttt{doom}$, with initial value of variables $X_1(0) = \gamma_N$, $X_2(0) = 9\gamma_N$, $X_3(0) = Z(0) = 0$.

Normalizing the model, we observe that all non-instantaneous transitions satisfy the continuous scaling. If we compute the drift, we obtain the following set of ODEs (as the guards in the transitions of `agent1` elicit with the modulus), with initial conditions $(1, 9, 0)$:

$$
\begin{cases}
\frac{\mathrm{d}x_1}{\mathrm{d}t} = x_2 + 9 \\
\frac{\mathrm{d}x_2}{\mathrm{d}t} = x_3 - 12 \\
\frac{\mathrm{d}x_3}{\mathrm{d}t} = 6
\end{cases}
$$

These equations can be integrated directly, obtaining $x_1(t) = t^3 - 6t^2 + 9t + 1$, whose trajectory can be seen in Figure 6(a). Notice that, for $t = 1$, the curve hits tangentially the line $x_1 = 5$, which is the activation surface associated with the robust activation function $h(\hat{\mathbf{x}}) = x_1 - 5$, while it transversally crosses such a line at $t = 4$. In Figure 6(b), we show the hitting time distribution for the sequence of CTMC for increasing size γ_N, by visualizing the passage-time distribution of the event $Z = 1$, i.e. $\mathbb{P}\{Z(t) = 1\}$ as a function of t. As we can see, the bimodal nature of the distribution persists also for large N, supporting the claim that tangential activation creates problems for convergence of exit times.

Convergence of Reset Kernels

There is a second source of discontinuity induced by instantaneous transitions, namely in the reset kernel of the PDMP on the activation surface of the guards. The problem lies in the fact that, if we have more than one instantaneous transition, a specific one will be active only in a subregion $\mathcal{H}_\pi$ of the activation surface $\mathcal{H}$, where $\mathcal{H} = \{\hat{\mathbf{x}} \mid h(\hat{\mathbf{x}}) = 0\}$ and $\mathcal{H}_\pi = \{\hat{\mathbf{x}} \in \mathcal{H} \mid h_\pi(\hat{\mathbf{x}}) = 0\}$. In particular, the reset kernel is not robust in the boundary $\partial_{\mathcal{H}}\mathcal{H}_\pi$ of $\mathcal{H}_\pi$ in $\mathcal{H}$. In fact, if a trajectory of the PDMP hits $\mathcal{H}$ in such a boundary, then a small perturbation can change the set of active guards (including or excluding π), and the fate of the system may be different. The same problem can manifest itself on the intersection between the activation surfaces of the guards of two instantaneous transitions π_1 and π_2: in any neighbourhood of (the boundary of) this region, we can find points in which only one of π_1 and π_2 is active. This lack of robustness reflects itself in a loss of continuity of the reset kernel. Hence we can no longer rely on this property to prove the convergence of the state after the reset (a fact used in the proof of Theorem 5.1).

Intuitively, convergence cannot hold for trajectories $\hat{\mathbf{x}}(t)$ of the PDMP that hit $\mathcal{H}$ in $\partial_{\mathcal{H}}\mathcal{H}_\pi$. In fact, trajectories of the CTMC that converge to $\hat{\mathbf{x}}(t)$ can hit either $\mathcal{H}_\pi$ or is complement in $\mathcal{H}$, implying that the CTMC can be reset to a different state from the PDMP. Furthermore, the probability of hitting $\mathcal{H}_\pi$ or its complement in $\mathcal{H}$ will depend on the geometry of $\mathcal{H}$ around the boundary $\partial_{\mathcal{H}}\mathcal{H}_\pi$, rather than on the priority functions governing the choice for the PDMP. We illustrate this point by the following simple example.

Example 6.2. We consider a model of a one-dimensional random walk in **sCCP**. More specifically, we consider a population-**sCCP** program $(\mathcal{A}, \gamma_N)$ with two variables to be approximated continuously, X and Y, and one variable Z to be kept discrete. In particular, X and Y will count how many times we go up and down, respectively. We have the following **sCCP** code:

$$\begin{array}{lcl} \texttt{up} & \stackrel{\text{def}}{=} & [* \to X' = X + 1]_{\gamma_N}\texttt{.up} \\ \texttt{down} & \stackrel{\text{def}}{=} & [* \to Y' = Y + 1]_{\gamma_N}\texttt{.down} \\ \texttt{doom1} & \stackrel{\text{def}}{=} & [X \geq \gamma_N \to Z = 1]_{\infty:99}\texttt{.0} \\ \texttt{doom2} & \stackrel{\text{def}}{=} & [Y \geq \gamma_N \to Z = -1]_{\infty:1}\texttt{.0} \end{array}$$

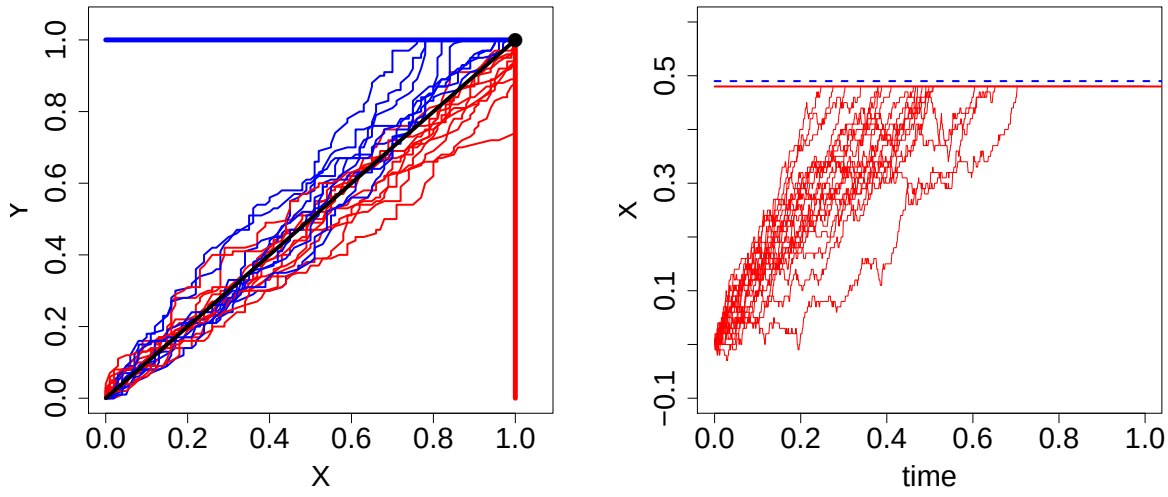
The initial network is $\texttt{up} \parallel \texttt{down} \parallel \texttt{doom1} \parallel \texttt{doom2}$, with initial value of variables $X(0) = Y(0) = Z(0) = 0$. This system is easily seen a to be a one-dimensional random walk for the variable $W = X - Y$. When visualized in the plane X, Y, the trajectories of the random walk are rotated by 45 degrees along the line defined by the equation $Y = X$ (see Figure 7(a)). In the normalized system, such trajectories will eventually hit one of the discontinuity surfaces at $\hat{x} = 1$ or $\hat{y} = 1$. The vector field of the PDMP is given by $F(x, y) = (1, 1)$, hence the solution from the point $\hat{x}(0) = \hat{y}(0) = 0$ is $\hat{x}(t) = \hat{y}(t) = t$, which corresponds to the line $\hat{x} = \hat{y}$. This line hits the activation surface in its corner point $(1, 1)$, where both transitions are active, hence after the reset $Z = 1$ with probability 0.99, and $Z = -1$ with probability 0.01. However, in the CTMC $\hat{X}$ and $\hat{Y}$ can only increment by $1/N$ asynchronously, meaning that each trajectory has to hit one of the two segments of the activation surface before the other. By a simple symmetry argument, we can see that for each N, the probability that $Z = 1$ after the reset is 0.5 (see again Figure 7(a)), hence convergence cannot hold for this model.

To have some hope to obtain convergence after a reset, we need to exclude trajectories of the PDMP $\hat{\mathbf{x}}(t)$ that are troublesome. Consider a **sCCP** model with m instantaneous transitions $\pi_1, \ldots, \pi_m$, and let $h_{\pi_1}, \ldots, h_{\pi_m}$ be the corresponding activation functions, and $h = \max\{h_{\pi_1}, \ldots, h_{\pi_m}\}$ be the activation function of the PDMP. Define the activation surface $\mathcal{H} = \{\hat{\mathbf{x}} \mid h(\hat{\mathbf{x}}) = 0\}$ and $\mathcal{H}_{\pi_j} = \{\hat{\mathbf{x}} \in \mathcal{H} \mid h_{\pi_j}(\hat{\mathbf{x}}) = 0\}$. Let $D_j = \partial_{\mathcal{H}}\mathcal{H}_{\pi_j}$ be the boundary of $\mathcal{H}_{\pi_j}$ in $\mathcal{H}$ and $D = \bigcup_j D_j$ be the union of such boundaries, called the discontinuity region of $\mathcal{H}$.

Definition 6.5. A PDMP $\hat{\mathbf{x}}(t)$ obtained from a TDSHA $\mathcal{T}$ has the *robust activation property* if and only if the set of trajectories hitting the discontinuity region D of $\mathcal{H}$ has probability zero.

This property essentially tells us that we can ignore the situations in which the PDMP activates instantaneous transitions in non-robust points.

However, this is not the only issue with reset kernels. There is another problem that can arise when we allow the activation functions h_π of instantaneous transitions to depend on the size γ_N, i.e. when



(a) Example 6.2

(b) Example 6.3

Figure 7: (left) Exemplification of the random walk model of Example 6.2. The activation surface of instantaneous transition `doom1` is shown in red, while the activation surface for `doom2` is visualized in blue. Trajectories are coloured according to the surface they hit. The black trajectory is the solution of the PDMP associated with the model. (right) Exemplification of the random walk model of Example 6.3. The trajectories coloured in red fire the `doom2` instantaneous transition, whose activation surface is also shown in red. The dotted blue line is the activation surface of `doom1`. Obviously, no trajectory is coloured in blue, as no trajectory can hit that surface.

$h_\pi^{(N)} \neq h_\pi$. The problem, in particular, manifests itself if the activation surfaces of two or more guards with size-dependent activation functions overlap (robustly) in the limit model.

Example 6.3. We consider a simple random walk model, with one variable X with values in $\mathbb{Z}$, that will be approximated continuously, and a variable Z that will remain discrete.

$$\begin{array}{lcl} \texttt{rw} & \stackrel{\text{def}}{=} & [* \to X' = X + 1]_{\gamma_N} \texttt{.rw} \\ & + & [* \to X' = X - 1]_{\gamma_N} \texttt{.rw} \\ \texttt{doom1} & \stackrel{\text{def}}{=} & [Z = 0 \wedge X \geq \gamma_N k - 1 \to Z = 1]_{\infty:1} \texttt{.0} \\ \texttt{doom2} & \stackrel{\text{def}}{=} & [Z = 0 \wedge X \geq \gamma_N k - 2 \to Z = -1]_{\infty:1} \texttt{.0} \end{array}$$

The initial network is $\texttt{rw} \parallel \texttt{doom1} \parallel \texttt{doom2}$, with initial value of variables $X(0) = Z(0) = 0$. The activation surface for $Z = 0$ of `doom1` in the normalized model is the hyperplane $\hat{X} = k - \frac{1}{\gamma_N}$, while that of `doom2` is the hyperplane $\hat{X} = k - \frac{2}{\gamma_N}$. As X is increased and decreased by one unit only (hence $\hat{X}$ is modified by $\frac{1}{N}$ units), for any N the system will always fire `doom2` (notice that the additional condition on Z forbids firing `doom1` once `doom2` has fired). Hence, $Z = -1$ eventually, for the CTMC models at any population level N (see also Figure 7(b), for $k = 0.5$). However, in the limit model both activation surfaces converge to the limit hyperplane $\hat{x} = k$, hence in the limit PDMP Z takes value -1 only with probability 0.5. Convergence again fails.

This example suggest that, in order to avoid such problems, we should either forbid N-dependent guards in instantaneous transitions of population-**sCCP** models, or try to forbid those situations in which more than one N-dependent guard can be robustly activated at the same time in the limit model. We state this in the following definition.

Definition 6.6. A set of activation functions of guards $h_1^{(N)}, \ldots, h_m^{(N)}$ of a population-**sCCP** model is *size-compatible* if and only if, for each j such that $h_j^{(N)}$ is size-dependent (i.e. $h_j^{(N)}$ converges uniformly to h_j in each compact set but $h_j^{(N)} \neq h_j$), then $int_{\mathcal{H}}(\mathcal{H}_j) \cap \mathcal{H}_i = \emptyset$, for each $i \neq j$ (i.e. in any point in which the limit activation function h_j is robustly zero in $\mathcal{H}$, no other h_i function is zero).

The limit PDMP $\hat{\mathbf{x}}(t)$ obtained from a population-**sCCP** model is *size-compatible* if and only if the set of activation function of guards of instantaneous transitions is size compatible.

Technically, Definitions 6.5 and 6.6 are the key properties that allow us to extend a lemma on the convergence of continuous reset kernels (Lemma C.1), into a more general result capable of dealing with discontinuous reset kernels of the form induced by instantaneous transitions. This will be formally discussed in Lemma D.3 in Appendix D.

Hybrid Convergence Theorem

The previous lemmas and hypothesis are the core argument for extending Theorem 5.1 in the presence of instantaneous transitions. Before proving it, we make explicit the scaling for instantaneous transitions.

Scaling 4 (Discrete Scaling for Instantaneous Transitions). A *normalized* instantaneous **sCCP** transition with *random reset* $\hat{\pi} = (g^{(N)}(\hat{\mathbf{X}}), \hat{\mathbf{X}}' = \hat{\mathbf{r}}^{(N)}(\hat{\mathbf{X}}, \mathbf{W}^{(N)}(\hat{\mathbf{X}})), \hat{\mathbf{p}}_\pi^{(N)}(\hat{\mathbf{X}}))$ of a population-**sCCP** program $(\mathcal{A}, \gamma_N)$ with variables partitioned into $\hat{\mathbf{X}} = (\mathbf{X}_d, \hat{\mathbf{X}}_c, \mathbf{X}_e)$, with $\hat{\mathbf{X}} \in E$, satisfies the *discrete scaling* if and only if:

1. The activation function $h^{(N)}(\hat{\mathbf{X}})$ of the guard $g^{(N)}(\hat{\mathbf{X}})$ converge uniformly in each compact $K \subset E$ to a continuous function $h(\hat{\mathbf{X}})$;
2. $\hat{\mathbf{p}}_\pi^{(N)}(\hat{\mathbf{X}}) = O(1)$, $\hat{\mathbf{p}}_\pi^{(N)}(\hat{\mathbf{X}})$ is continuous and it converges uniformly in each compact $K \subset E$ to the continuous function $\hat{\mathbf{p}}_\pi(\hat{\mathbf{X}})$;
3. Resets converge weakly (uniformly on compacts), i.e. for each $\hat{\mathbf{x}}^{(N)} \to \hat{\mathbf{x}}$ in E, $\hat{\mathbf{r}}_{(N)}(\hat{\mathbf{x}}^{(N)}, \mathbf{W}^{(N)}(\hat{\mathbf{x}}^{(N)}) \Rightarrow \hat{\mathbf{r}}(\hat{\mathbf{x}}, \mathbf{W}(\hat{\mathbf{x}}))$, as random elements on E.

If an instantaneous **sCCP**-transition of the form $\pi = [h_\pi^{(N)}(\hat{\mathbf{X}}) \geq 0 \to \hat{\mathbf{X}}' = \hat{\mathbf{r}}_\pi^{(N)}(\hat{\mathbf{X}}, \mathbf{W}^{(N)}(\hat{\mathbf{X}}))]_{\infty:\hat{\mathbf{p}}_\pi^{(N)}(\hat{\mathbf{X}})}$ satisfies the previous scaling, then the corresponding transition in the limit TDSHA is given by $(h_\pi(\hat{\mathbf{x}})) \geq 0, \hat{\mathbf{r}}_\pi(\hat{\mathbf{x}}, \mathbf{W}(\hat{\mathbf{x}})), \hat{\mathbf{p}}_\pi(\hat{\mathbf{x}}))$. Consider now the limit PDMP $\hat{\mathbf{x}}$ on E, associated with the normalized TDSHA $\hat{\mathcal{T}}(\mathcal{A})$ constructed from a sequence $(\mathcal{A}, \gamma_N)$ of population-**sCCP** models, in which all transitions satisfy Scalings 2, 3, or 4.

Theorem 6.1. *Let $(\mathcal{A}, \gamma_N)$ be a sequence of population-**sCCP** models for increasing system size $\gamma_N \to \infty$, as $N \to \infty$, with variables partitioned into $\mathbf{X} = (\mathbf{X}_d, \mathbf{X}_c, \mathbf{X}_e)$, with discrete stochastic actions satisfying scaling 3, instantaneous actions satisfying scaling 4, and continuous actions satisfying scaling 2. Let $\hat{\mathbf{X}}^{(N)}(t)$ be the associated sequence of normalized CTMC and $\hat{\mathbf{x}}(t)$ be the limit PDMP associated with the normalized limit TDSHA $\hat{\mathcal{T}}(\mathcal{A})$.*

If $\hat{\mathbf{x}}_0^{(N)} \Rightarrow \hat{\mathbf{x}}_0$ (weakly) and the PDMP is non-Zeno*,* robustly transversal*, has the* robust activation property *and it is* size-compatible*, then $\hat{\mathbf{X}}^{(N)}(t)$ converges weakly to $\hat{\mathbf{x}}(t)$, $\hat{\mathbf{X}}^{(N)} \Rightarrow \hat{\mathbf{x}}$, as random elements in the space of cadlag function with the Skorohod metric.*

Proof. The proof is only sketched here, see Appendix D for further details. The idea is to reason as in Theorem 5.1, just replacing the machinery about jump times by a more refined one taking into account also instantaneous jumps. Essentially, we have to take the minimum of the stochastic and instantaneous jump times, and choose which reset kernel to use according to which kind of event (stochastic or instantaneous) fires first. The weak convergence of these new jump times and reset kernels follows easily from the convergence of stochastic and instantaneous ones. □

*Remark** 6.1. Theorem 6.1 relies on three global properties of the PDMP associated with the population-**sCCP** model, namely the robust transversal, the robust activation, and the size-compatibility property.

The last requirement should be generally relatively easy to check, as it depends only on the activation functions of guards, and not on their interaction with the vector field. In fact, in most practical cases, guards are boolean combinations of linear predicates, hence if some of them depend on N, by computing the limit activation function (which should also be a combination of linear functions $h_{j,1}, \ldots, h_{j,k_j}$), one can discover if there is a robust overlapping of guards by solving a linear system of equations for each pair i, j of size-dependent guards (say $h_{i,1}(\hat{\mathbf{x}}) = 0, \ldots, h_{i,k_i}(\hat{\mathbf{x}}) = 0, h_{j,1}(\hat{\mathbf{x}}) = 0, \ldots, h_{j,k_j}(\hat{\mathbf{x}}) = 0$) and checking if the solution has dimension $n-2$ or less in each mode. Here we are implicitly assuming that the PDMP and the sequence of CTMC can evolve in an open subset of E of dimension n in each mode (if this is not the case, due to conservation laws, we can just use these laws to reduce the dimensionality of the system). If the activation functions are non-linear, then the previous approach can be still carried out, but checking the intrinsic dimensionality of a non-linear manifold is obviously more complicated.

On the other hand, the robustness conditions on the PDMP are more complex to check. The robust transversal property requires that the PDMP transversally crosses an activation surface with probability one. If the PDMP is deterministic (i.e., there is no discrete and stochastic transition), then this check can be carried out along the single trajectory starting from the given initial state. In case the number of firings of instantaneous transitions is finite, or these events are ultimately periodic, then it may be possible to set up a semi-decision procedure for this task. The problem, also in this simple case, is that checking if a trajectory has a tangential crossing is the same as looking for a non-simple zero[11] of the activation function. However, no root finding algorithm is able to properly deal with non-simple zeros, even for analytic functions, see e.g. [59]. In fact, we can only hope to compute a non-deterministic approximation of the trajectory, namely a flow tube around it, which for a tangential activation would intersect the surface but not cross it completely. This still does not prove that there is a tangential zero, just that we cannot ignore this possibility. Note that if a trajectory does not intersect the activation surface but a flow tube of small radius around it does, then the behaviour of the sequence of CTMC can diverge from that of the PDMP due to small fluctuations around the limit trajectory, which can lead to completely different behaviours. In those cases, even if convergence will hold in the limit, the speed of convergence can be very slow.

If the PDMP is a proper stochastic process, then checking the robust transversal property can be even more challenging. In fact, the condition requires us to show that non-transversal activations happen with

[11] A zero of a real valued differentiable function is non-simple if also the derivatives of the function up to order $k \geq 1$ are zero in the same point.

probability zero. One way to approach the problem is to exploit randomness to our advantage. Suppose that, in a given mode q, the continuous state space E_q has topological dimension n and that we can show that the activation surfaces have (topological) dimension $n-1$ and set of points B in the activation manifold corresponding to non-transversal crossing has (topological) dimension $n-2$ or less. Then, the set of points E_t of E_q such that $\hat{\mathbf{x}}(t) \in B$ if $\hat{\mathbf{x}}_0 \in E_t$ has dimension $n-2$ (it is the continuous image of B under the flow of the vector field for $-t$ units of time) so that the subset $E_B \subseteq E_q$ of initial points for which $\hat{\mathbf{x}}(t)$ hits B has dimension $n-1$. If we can further prove that the distribution at each time t of the PDMP is absolutely continuous with respect to the Lebesgue measure (i.e. $\mathbb{P}(A) = 0$ for each Borel set A of Lebesgue measure 0), it necessarily follows that the probability that $\hat{\mathbf{x}}(t) \in B$ is zero (as E_B has Lebesgue measure zero). This last property can be enforced by requiring that the initial conditions and the reset kernels are absolutely continuous probability distributions (e.g. n-dimensional multivariate Gaussian distributions). If the system satisfies some conservation law, so that we are interested in its dynamics in a manifold of dimension less than n, then we can reduce its dimensionality and analyse the reduced system in the way sketched above.

Proving that the set B has dimension $n-1$ or less, instead, is more challenging in general. If the activation function of guards are linear (or analytic), and the vector field is analytic, then one may exploit properties of analytic manifolds for this task, studying the set of zeros of the scalar product of the normal vector to the activation surface with the vector field (B, in fact, is contained in this zero set). We do not pursue this direction any further in this paper, leaving its investigation for future work, with the goal of providing (semi-)automatic static analysis procedures to check for the applicability of the hybrid approximation method, at least for a practically relevant subclass of population-**sCCP** models.

The property of robust activation can be dealt with along the lines sketched above, looking at the dimension of the intersection of activation surfaces. In this case, the task should be considerably simplified if all guards are linear.

Example 6.4. We consider now a different scenario, in which we model the spreading of a worm epidemic in a computer network. The class of models used for this circumstance is usually drawn from the well developed field of epidemiology, and we make no exception to this rule. We will consider a simple SIR model [2], in which each node of the network has three states: susceptible X_s, infected X_i and recovered X_r. Here the size of the system γ_N coincides with the total population N of nodes, which is assumed to be constant, i.e. $X_i + X_s + X_r = N$. We assume that infection happens by the malicious action of the worm in infected nodes, which try to send infected messages around the network. There is also a small chance that infection comes externally from the network. Recovery from an infection is obtained by patching an infected computer node. However, after some time new generations of worms appear, and we describe this by the loss of immunity of recovered nodes, that return to be susceptible. We assume that only infected nodes are patched. The **sCCP** code for this model is as follows:

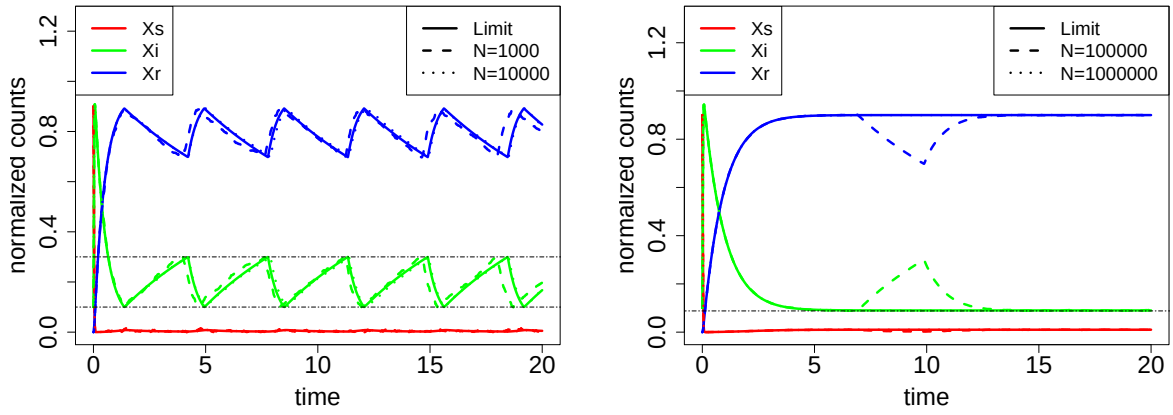
$$
\begin{array}{lcl}
\texttt{infection} & \stackrel{\text{def}}{=} & [* \to X_i' = X_i + 1 \wedge X_s' = X_s - 1]_{k_i X_s X_i/\gamma_N}.\texttt{infection} + \\
 & & [* \to X_i' = X_i + 1 \wedge X_s' = X_s - 1]_{\gamma_N k_e}.\texttt{infection} \\
\texttt{loss\_immunity} & \stackrel{\text{def}}{=} & [* \to X_r' = X_r - 1 \wedge X_s' = X_s + 1]_{k_s X_r}.\texttt{loss\_immunity} \\
\texttt{patching} & \stackrel{\text{def}}{=} & [* \to X_i' = X_i - 1 \wedge X_r' = X_r + 1]_{k_p X_i}.\texttt{patching}
\end{array}
$$

The patching rate k_p is the only controllable activity in the system, and we will use instantaneous transitions to model control policies. In particular, we consider here the following policy: if the fraction of infected computers is above a threshold α_1, we increase the patch rate from k_p^0 to k_p^1. If the fraction of infected fall below the threshold $\alpha_0 < \alpha_1$, we switch back to the normal patching rate. We model this in **sCCP** by introducing a new variable U, taking values 0 or 1, modifying the agent `patching` as

$$
\begin{array}{lcl}
\texttt{patching} & \stackrel{\text{def}}{=} & [U = 0 \to X_i' = X_i - 1 \wedge X_r' = X_r + 1]_{k_p^0 X_i}.\texttt{patching} \\
 & + & [U = 1 \to X_i' = X_i - 1 \wedge X_r' = X_r + 1]_{k_p^1 X_i}.\texttt{patching}
\end{array}
$$

and introducing the `control` agent

$$
\begin{array}{lcl}
\texttt{control} & \stackrel{\text{def}}{=} & [X_i/\gamma_N > \alpha_1 \to U' = 1]_{\infty:1}.\texttt{control} \\
 & + & [X_i/\gamma_N < \alpha_0 \to U' = 0]_{\infty:1}.\texttt{control}
\end{array}
$$



(a) $k_p^1 = 2$, $\alpha_1 = 0.1$

(b) $k_p^1 = 1$, $\alpha_1 = 0.088$

Figure 8: Comparison of hybrid and stochastic trajectories of the epidemics model of Example 6.4, for parameters $k_i = 100$, $k_e = 0.001$, $k_s = 0.1$, $k_p^0 = 0.1$, $k_p^1 = 2.0$ (left) or $k_p^1 = 1.0$ (right). Control thresholds are $\alpha_0 = 0.3$ and $\alpha_1 = 0.1$ (left) or $\alpha_1 = 0.088$ (right). Initial conditions are $\hat{x}_i(0) = 0.1$ and $\hat{x}_s(0) = 0.9$. Note that for N large the stochastic trajectory is indistinguishable from the hybrid limit one. In the figure on the right, the threshold α_1 is slightly smaller than the steady state of the ODE when $U = 1$. However, the stochastic system can hit the threshold and change mode, even for γ_N large. In any fixed time horizon, this event becomes less and less likely as N goes to infinity: here, we observe a spike for $N = 100000$, but not for N equal to one million.

First, note that this model satisfies the scaling assumptions of Theorem 6.1, when all variables except U and all stochastic **sCCP**-transitions are considered as continuous. As there is no stochastic transition, the limit model is deterministic. Hence, if we start from a given initial state, we need to check that the single trajectory of the limit model satisfies the assumptions. For a given set of parameters, shown in Figure 8, we can choose α_1 such that the steady state of the limit model with low patching rate is above α_1 and α_0 such that the steady state of the limit ODE model with high patching rate is below α_0, inducing oscillations in the limit hybrid model. This is confirmed in Figure 8(a), where we can also check visually that the crossing of the guard surfaces is always transversal. A formal proof can be given as well, by verifying that the projection of the vector field on the orthogonal direction to $\hat{X}_i = \alpha_1$ or $\hat{X}_i = \alpha_0$ is null in a single point, namely $(k_p^0/k_i, \alpha_1, 1 - k_p^0/k_i - \alpha_1)$ or $(k_p^1/k_i, \alpha_0, 1 - k_p^1/k_i - \alpha_0)$, respectively, and observing that the trajectory in Figure 8 never passes from these points. As the conditions of Theorem 6.1 are satisfied, we can conclude that the sequence of CTMC models obtained by the **sCCP**-program for increasing γ_N converges to the hybrid system.

In Figure 8(b), we show the same model for a different high patching rate and a different threshold α_1, such that the limit model in state $U = 1$ converges to a steady state slightly greater than α_1, thus never activating the instantaneous transition. We can see that the stochastic model behaves in the same way, but for N very large, because the proximity of α_1 induces an activation of the transition in stochastic trajectories with small (but non null) probability for any N. In particular, this implies that if we leave enough time, almost surely a stochastic trajectory will eventually cross the surface $\hat{x}_i = \alpha_1$, changing discrete mode. This shows that the notion of weak convergence is restricted to the transient behaviour, but does not bring in general information about the steady state.

6.1 Time-Dependent Guards

Guards depending on time on instantaneous transitions can be a valuable addition to the modelling language, as they allow us to describe global events that have a duration, which can be either deterministic or stochastic.

More precisely, we consider the extension of **sCCP** [19] with a reserved keyword *time*, referring to simulation time, whose usage is confined to guards and update functions of instantaneous transition, and to update functions of stochastic transitions, which have to be kept discrete. Moreover, the special variable *time* can never be updated. Specifically, we allow instantaneous transitions of the form $[\mathbf{g}_\pi(\mathbf{X}, time) \to \mathbf{X}' = \mathbf{r}_\pi(\mathbf{X}, \mathbf{W}, time)]_{\infty:w}$. In particular, $\mathbf{g}_\pi(\mathbf{X}, time)$ is required to be of the form $time \geq h_0(\mathbf{X}) \wedge \mathbf{g}_{\pi,1}(\mathbf{X})$, for some function h_0 and some standard guard predicate $\mathbf{g}_{\pi,1}(\mathbf{X})$ (with activation function $h_1(\hat{\mathbf{X}})$). We call a population-**sCCP** model $(\mathcal{A}, \gamma)$ with timed-guards a *time-guarded population-**sCCP** model*.

Translation of these transitions to TDSHA is straightforward and follows the same scheme as Section 2.3. The only difference is that in the TDSHA/PDMP setting it is more convenient to internalize the notion of time, by adding a dedicated clock variable keeping track of the global simulation time. This is done by adding a new continuous variable, $Time$, and a new automata, called *time-monitor*, in the parallel composition of TDSHA, with a single continuous transition of the form $(\mathbf{1}_{Time}, 1)$, where $\mathbf{1}_{Time}$ is the vector equal to one in the position of the variable $Time$, and zero elsewhere.

In the following, we restrict our attention to time-guarded transitions that satisfy the following scaling assumption with respect to the population size γ_N.

Scaling 5 (Discrete Scaling for Time-Guarded Instantaneous Transitions). A *normalized* time-guarded instantaneous **sCCP** transition with *random reset* $\hat{\pi} = (g^{(N)}(\hat{\mathbf{X}}, time), \hat{\mathbf{X}}' = \hat{\mathbf{r}}^{(N)}(\hat{\mathbf{X}}, \mathbf{W}^{(N)}(\hat{\mathbf{Y}}), time), \hat{\mathbf{p}}_\pi^{(N)}(\hat{\mathbf{X}}))$ of a population-**sCCP** program $(\mathcal{A}, \gamma_N)$ with variables partitioned into $(\mathbf{X}_d, \mathbf{X}_c, \mathbf{X}_e)$, $\hat{\mathbf{X}} \in E$, has *discrete scaling* if and only if:

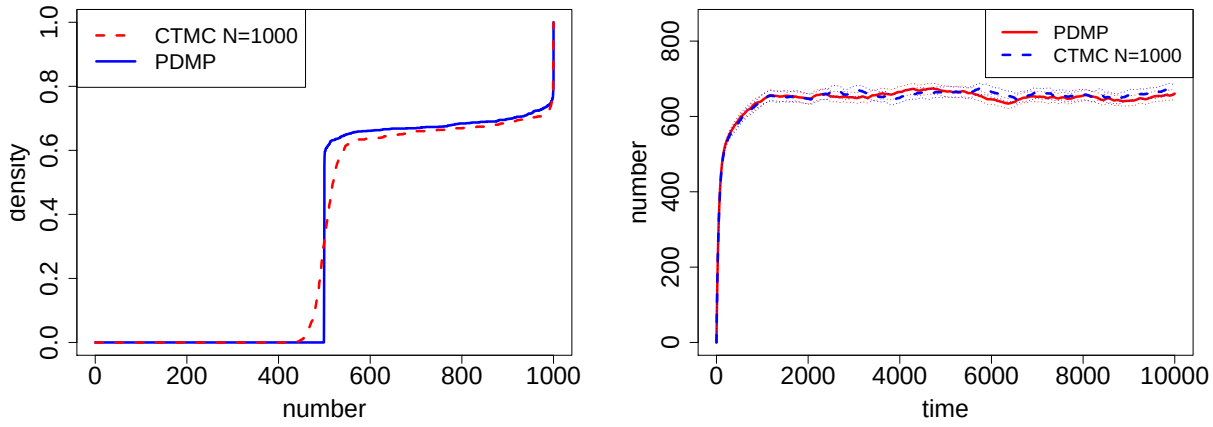
1. The activation function of the guard $g^{(N)}(\hat{\mathbf{X}})$, which is $\min\{time - h_0^{(N)}(\hat{\mathbf{X}}), h_1^{(N)}(\hat{\mathbf{X}})\}$, is such that $h_i^{(N)}(\hat{\mathbf{X}})$ converges uniformly in each compact $K \subset E$ to a continuous function $h_i(\hat{\mathbf{X}})$, $i = 0, 1$. Furthermore, $h_0^{(N)}$ and h_0 *do not* depend on the variables $\hat{\mathbf{X}}_c$ of $\hat{\mathbf{X}}$ that are modified continuously;
2. $\hat{\mathbf{p}}_\pi^{(N)}$ satisfies the same conditions as in Scaling 4;
3. Resets converge weakly (uniformly on compacts), i.e. for each $(\hat{\mathbf{x}}^{(N)}, t^{(N)}) \to (\hat{\mathbf{x}}, t)$ in $E \times \mathbb{R}_{\geq 0}$, $\hat{\mathbf{r}}_{(N)}(\hat{\mathbf{x}}^{(N)}, \mathbf{W}^{(N)}(\hat{\mathbf{x}}^{(N)}), t^{(N)}) \Rightarrow \hat{\mathbf{r}}(\hat{\mathbf{x}}, \mathbf{W}(\hat{\mathbf{x}}), t)$, as random elements on E.

Under the previous scaling, if we consider initial conditions (or the state after one jump) such that $\hat{\mathbf{X}}_0^{(N)} \Rightarrow \hat{\mathbf{X}}_0$, given the independence of the activation function from variables modified continuously, we easily obtain $h_0^{(N)}(\hat{\mathbf{X}}_0^{(N)}) \Rightarrow h_0(\hat{\mathbf{X}}_0)$ (reason as in Lemma 6.1). Recalling that in the limit PDMP, $Time$ is treated like a regular continuous variable, we can combine this observation with the discussion about exit times and reset kernels in the previous section to obtain convergence. Note, in particular, that the activation condition on *time* has always a robustly transversal activation function (as $Time$ is monotonically increasing). Hence, the only problems for convergence of a timed-transition can come from the other component of the guard (i.e. from the activation function h_1). Therefore, we obtain that the time $T^{(N)}$ in which $g^{(N)}$ becomes true converges weakly to the time T in which g becomes true. Then, a minor adaptation of the proof of Theorem 6.1 gives the following

Proposition 6.1. *Let $(\mathcal{A}, \gamma_N)$ be a sequence of time guarded population-**sCCP** models for increasing systems size, $\gamma_N \to \infty$, as $N \to \infty$, with variables partitioned into $\mathbf{X} = (\mathbf{X}_d, \mathbf{X}_c, \mathbf{X}_e)$, with discrete stochastic actions satisfying scaling 3, instantaneous actions satisfying scaling 4, time guarded actions satisfying scaling 5, and continuous actions satisfying scaling 2. Let $\hat{\mathbf{X}}^{(N)}(t)$ be the associated sequence of normalized CTMC and $\hat{\mathbf{x}}(t)$ be the limit PDMP associated with the sequence of normalized TDSHA $\hat{\mathcal{T}}(\mathcal{A}, \gamma_N)$.*

If $\hat{\mathbf{X}}_0^{(N)} \Rightarrow \hat{\mathbf{x}}_0$ (weakly) and the PDMP is non-Zeno*, robustly transversal, has the* robust activation property *and it is* size-compatible*, then $\hat{\mathbf{X}}^{(N)}(t)$ converges weakly to $\hat{\mathbf{x}}(t)$, $\hat{\mathbf{X}}^{(N)} \Rightarrow \hat{\mathbf{x}}$, as random elements in the space of cadlag function with the Skorokhod metric.* □

Example 6.5. As an example of time-dependent guards, we consider again the client-server model with breakdown, as in Example 5.2. In that example, we used random resets to model a variable level of damage, reflecting in the time needed to repair the system. Here, instead, we consider a single damage level, but with a generally distributed repair time. In terms of **sCCP** model, we need an environment variable, say K, representing the time in which the server repair will finish. It will be re-sampled from a given distribution



(a) Weibull Breakdown: $t = 10000$

(b) Weibull Breakdown: average

Figure 9: Empirical cumulative distribution of clients requesting service at time $t = 10000$ and average number of clients request service for the client-server model of Example 6.5. The model has a fixing time sampled from a Weibull distribution with shape 1.5 and rate 1/1000. Other parameters are as in the caption of Figure 3. The bimodality of the distribution is captured, with less variability in the the hybrid model. The average is almost indistinguishable.

each time a breakdown occurs. More specifically, let W be a random variable on the positive reals, with cumulative distribution function $F(t)$, independent of the current state. Each time the server breaks, we set K to $time + W$. The `server` agent now becomes

$$\begin{array}{rcl} \texttt{server} & \stackrel{\text{def}}{=} & [* \to X_i' = X_i - 1 \wedge X_b' = X_b + 1 \wedge K' = time + W]_{k_b X_i} \,.\texttt{server} \\ & + & [time = K \to X_i' = X_i + 1 \wedge X_b' = X_b - 1]_{\infty:1} \,.\texttt{server} \end{array}$$

It is easy to see that this modified model satisfies the scaling conditions of Proposition 6.1, as the guard of the times transition is independent of γ_N. It follows that convergence holds, as can be seen in Figure 9, where we consider a fixing time sampled according to a Weibull distribution.

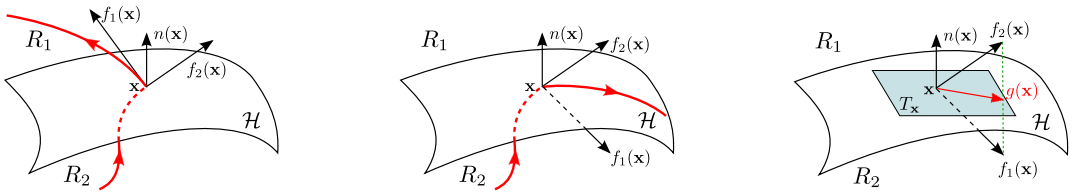
7 Dealing with Guards Depending on Continuous Variables

In this section, we look at what happens if we allow guards depending on continuous variables in **sCCP** transitions, which in the limit can either be approximated as continuous or be kept discrete and stochastic. This additional feature, which is straightforward from the point of view of the modelling language and which poses no problems in the definition of the CTMC semantics, has more complex consequences for what concerns the hybrid limits.

We will first focus on guards on continuous transitions, as these are somehow more delicate to deal with. Guards on discrete stochastic transitions, which create problems that are, in a certain sense, analogous to those with instantaneous transitions, will be discussed later on in Section 7.4.

7.1 Guards on Continuous Transitions

Guards on continuous transitions introduce discontinuities in the vector field. In fact, the rate function $\hat{\lambda}_\pi(\hat{\mathbf{x}})$ of a continuous transition π has to be multiplied by the indicator function of the guard predicate, which we assume to be of the form $h(\hat{\mathbf{x}}) \geq 0$, obtaining the discontinuous function $\hat{\mathbf{f}}_\pi(\hat{\mathbf{x}}) = \hat{\lambda}_\pi(\hat{\mathbf{x}}) \cdot \mathbf{I}\{h(\hat{\mathbf{x}}) \geq 0\}$. In doing this operation, we leave the world of differential equations, entering into the more intricate realm of discontinuous or piecewise-smooth dynamical systems (PWSS) [26, 33] or, more generally, of differential inclusions [3].



(a) Transversal motion (b) Sliding motion (c) Sliding vector field

Figure 10: Schematic representations of transversal motion across $\mathcal{H}$, of sliding motion along $\mathcal{H}$, and of the geometric construction of the sliding vector field $G(\hat{\mathbf{x}})$.

The problem, roughly speaking, is that existence and uniqueness of the solution of an ODE with discontinuous right-hand-side is not guaranteed even if all rate functions are regular (say Lipschitz continuous) and if guards are also described by smooth functions (say differentiable functions). Furthermore, solutions can exhibit strange behaviours, like *sliding motion* (sliding on a discontinuity surface) or *chattering* (Zeno behaviour in crossing discontinuity surfaces).

The lack of uniqueness, in particular, is problematic in our context, as it is a fundamental condition in the definition of the class of PDMP we consider here. Indeed, more general frameworks can be considered, like PDMP based on differential inclusions, but we leave the investigation of this direction for future work.

In this paper, we will follow the treatment of [20], in which the author discusses mean field limits in presence of guards, when the limit is a PWSS. A more general approach is that of [35], but we stick to the first one as we believe it is more intuitive. In the next subsection, we will briefly give an introduction to PWSS, in which we will discuss conditions for existence and uniqueness of solutions. Then, we will turn our attention to fluid approximation of those systems and plug these results into our framework.

7.1.1 Piecewise-Smooth Dynamical Systems

Consider an ordinary differential equation $\frac{d\hat{\mathbf{x}}}{dt} = F(\hat{\mathbf{x}})$. A solution in the classical sense is a (continuously) differentiable function $\hat{\mathbf{x}}(t)$ such that $\frac{d}{dt}\hat{\mathbf{x}}(t) = F(\hat{\mathbf{x}}(t))$, and $\hat{\mathbf{x}}(0) = \hat{\mathbf{x}}_0$. A classical result is the Picard Lindelöf theorem [53]: if F is (locally) Lipschitz on a set $E \subseteq \mathbb{R}^n$ and $\hat{\mathbf{x}}_0$ is in the interior of E, then there exists a unique global solution of the differential equation within E.

However, here we are interested in dynamical systems in which the right hand side of the ODE can be a discontinuous function, possibly undefined, on a set of points of measure zero. This is the setting studied in the theory of ordinary differential equations with discontinuous right-hand side [33]. In particular, we will consider the so-called switching systems or piecewise smooth (PWS) dynamical systems [26, 32]. Let $F : E \to \mathbb{R}^n$, with $E \subseteq \mathbb{R}^n$, and suppose there exist a finite set of domains $\mathcal{R}_i$, $i = 1, \ldots, s$, such that F is smooth (or at least Lipschitz) on $\bar{\mathcal{R}}_i$, the closure of $\mathcal{R}_i$, and $\bigcup \bar{\mathcal{R}}_i \supseteq \bar{E}$. Notice that F can be discontinuous only on the boundaries $\partial\mathcal{R}_i$ of the regions $\mathcal{R}_i$, so that the discontinuous set is $\mathcal{H} = \bigcup \partial\mathcal{R}_i$ and it has measure zero.

In the following, we will briefly sketch some basic notions of these systems, which we will need in the following, starting from the concept of a solution. In fact, given that the vector field is discontinuous, we cannot look anymore for solutions which are continuously differentiable functions. Therefore, we will look for solutions among *absolutely continuous* functions, i.e. continuous functions which are equal to the integral of another function [52] and are henceforth differentiable almost everywhere.

In order to define such solutions, we will lift the function F to a set valued function $\bar{F}$, $\bar{F}(\hat{\mathbf{x}}) \subseteq \mathbb{R}^n$, known as the *Filippov extension* of F. Then we define a *Filippov solution* as an absolutely continuous function $\hat{\mathbf{x}}(t)$ such that $\hat{\mathbf{x}}(0) = \hat{\mathbf{x}}_0$ and $\frac{d}{dt}\hat{\mathbf{x}}(t) \in \bar{F}(\hat{\mathbf{x}}(t))$ almost everywhere. That is to say, we replace the discontinuous differential equation by a *differential inclusion* [26, 3]. More specifically, we define $\bar{F}(\hat{\mathbf{x}})$ as $\mathtt{co}\{\lim_{k\to\infty} F(\hat{\mathbf{x}}_k) \mid \hat{\mathbf{x}}_k \to \hat{\mathbf{x}},\ \hat{\mathbf{x}}_k \notin \mathcal{H}\}$, where $\mathtt{co}$ denotes the convex closure of a set. Notice that for each continuity point $\hat{\mathbf{x}}$ of F, $\bar{F}(\hat{\mathbf{x}}) = \{F(\hat{\mathbf{x}})\}$, so that we have a proper differential inclusion only in the discontinuity region $\mathcal{H}$.

For simplicity, consider a PWS system constituted by only two regions $\mathcal{R}_1$ and $\mathcal{R}_2$, and let $\hat{\mathbf{x}} \in \mathcal{H}$ be a point of discontinuity of the vector field. Furthermore, suppose that F equals the function F_1 on $\bar{\mathcal{R}}_1$ and the function F_2 on $\bar{\mathcal{R}}_2$, and that $F_{1,j}(\hat{\mathbf{x}}) < F_{2,j}(\hat{\mathbf{x}})$. In this setting, $\bar{F}_j(\hat{\mathbf{x}}) = [F_{1,j}(\hat{\mathbf{x}}), F_{2,j}(\hat{\mathbf{x}})]$.

The existence of a solution, starting from a point $\mathbf{x_0}$, is guaranteed under mild conditions on the Filippov extension $\bar{F}$ of F [33]: $\bar{F}$ must be *(locally) bounded*[12] and *upper semicontinuous*[13]. Consider again a PWS system with two regions $\mathcal{R}_1$ and $\mathcal{R}_2$, as above. Then, existence is guaranteed if functions F_1 and F_2 are continuous on $\bar{\mathcal{R}}_1$ and $\bar{\mathcal{R}}_2$.

In order to understand the behaviour of a PWS dynamical system on a discontinuity point of the vector field, we restrict our attention to the two regions system (which is a good local model, unless a discontinuity point belongs to the boundary of more than two regions), further assuming that $\mathcal{R}_1$ and $\mathcal{R}_2$ are separated by a smooth surface $\mathcal{H}$. In particular, $\mathcal{H}$ is defined as $\mathcal{H} = \{\hat{\mathbf{x}} \mid h(\hat{\mathbf{x}}) = 0\}$, where h is a function with continuous second order derivatives, while $\mathcal{R}_1 = \{\hat{\mathbf{x}} \mid h(\hat{\mathbf{x}}) > 0\}$ and $\mathcal{R}_2 = \{\hat{\mathbf{x}} \mid h(\hat{\mathbf{x}}) < 0\}$. We further require that $\nabla h(\hat{\mathbf{x}}) \neq \mathbf{0}$ for each point $\hat{\mathbf{x}} \in \mathcal{H}$, so that the normal vector $n(\hat{\mathbf{x}}) = \nabla h(\hat{\mathbf{x}})/||\nabla h(\hat{\mathbf{x}})||$ is always defined for the surface $\mathcal{H}$, and always points into $\mathcal{R}_1$, see Figure 10.

To understand the behaviour of a trajectory when it hits the surface $\mathcal{H}$, consider a situation in which the solution is in the interior of $\mathcal{R}_2$ and hits $\mathcal{H}$ in $\hat{\mathbf{x}}$ at some time t. Then, two things can happen, depending on the relative orientation of the vectors $F_1(\hat{\mathbf{x}})$ and $F_2(\hat{\mathbf{x}})$ with respect to $\mathcal{H}$. In particular, as $\hat{\mathbf{x}}(t)$ hits $\mathcal{H}$ from $\mathcal{R}_2$, the vector $F_2(\hat{\mathbf{x}})$ must point towards $\mathcal{R}_1$. If also the vector $F_1(\hat{\mathbf{x}})$ points towards $\mathcal{R}_1$, then the trajectory $\hat{\mathbf{x}}(t)$ crosses the surface $\mathcal{H}$, possibly with a discontinuity in its derivative. This phenomenon is called *transversal motion*, see Figure 10(a). Alternatively, the vector $F_1(\hat{\mathbf{x}})$ may point towards $\mathcal{R}_2$. In this case, the trajectory cannot enter $\mathcal{R}_1$, as it will be pushed immediately back to $\mathcal{H}$, but, symmetrically, it cannot also remain in $\mathcal{R}_2$. Therefore, the motion is confined in the discontinuity surface $\mathcal{H}$. This kind of behaviour is known as *sliding motion*, see Figure 10(b). In particular, the trajectory $\hat{\mathbf{x}}(t)$ follows the solution of the vector field tangential to $\mathcal{H}$ obtained by selecting the only vector in $\bar{F}(\hat{\mathbf{x}})$ tangential to $\mathcal{H}$, see Figure 10(c). More precisely, the sliding motion is defined by the differential equation $\frac{d}{dt}\hat{\mathbf{x}} = G(\hat{\mathbf{x}})$, where G is the vector field $(\hat{\mathbf{x}}) = \alpha(\hat{\mathbf{x}})f_1(\hat{\mathbf{x}}) + (1 - \alpha(\hat{\mathbf{x}}))f_2(\hat{\mathbf{x}})$. The value of the weight coefficient $\alpha(\hat{\mathbf{x}})$ is obtained by requiring that $n^T(\hat{\mathbf{x}})G(\hat{\mathbf{x}}) = 0$ (i.e., that $G(\hat{\mathbf{x}})$ is tangential to $\mathcal{H}$), obtaining

$$\alpha(\hat{\mathbf{x}}) = \frac{n^T(\hat{\mathbf{x}})F_2(\hat{\mathbf{x}})}{n^T(\hat{\mathbf{x}})F_2(\hat{\mathbf{x}}) - n^T(\hat{\mathbf{x}})F_1(\hat{\mathbf{x}})},$$

where $n^T(\hat{\mathbf{x}})F_1(\hat{\mathbf{x}})$ is the projection of $F_1(\hat{\mathbf{x}})$ along the normal vector $n(\hat{\mathbf{x}})$ of $\mathcal{H}$ in $\hat{\mathbf{x}}$. Sliding motion continues until one (and only one) of the two vectors fields, say F_1, becomes tangential to $\mathcal{H}$. In this case, the motion continues in the region $\mathcal{R}_1$. The condition that only one vector out of $F_1(\hat{\mathbf{x}})$ and $F_2(\hat{\mathbf{x}})$ becomes tangential to $\mathcal{H}$ is known as the first order exit condition of the sliding motion [32]. If both F_1 and F_2 become tangential, then the motion continues on a submanifold of $\mathcal{H}$, but we do not consider these situations in this paper, which can, however, be treated similarly to the motion in the intersection of the boundary between three or more regions [32]. Hence, from now on we tacitly assume that sliding motion terminates with first order exit conditions.

In general, if we are in a point $\hat{\mathbf{x}}$ of $\mathcal{H}$, the behaviour of a solution starting in $\hat{\mathbf{x}}$ depends on the values of $n^T(\hat{\mathbf{x}})F_1(\hat{\mathbf{x}})$ and $n^T(\hat{\mathbf{x}})F_2(\hat{\mathbf{x}})$:

- If both $n^T(\hat{\mathbf{x}})F_1(\hat{\mathbf{x}})$ and $n^T(\hat{\mathbf{x}})F_2(\hat{\mathbf{x}})$ are non-zero and have the same sign, then there is a *transversal crossing* of the surface.
- If $n^T(\hat{\mathbf{x}})F_1(\hat{\mathbf{x}}) < 0$ and $n^T(\hat{\mathbf{x}})F_2(\hat{\mathbf{x}}) > 0$, we have a *stable sliding motion* along $\mathcal{H}$.
- If $n^T(\hat{\mathbf{x}})F_1(\hat{\mathbf{x}}) > 0$ and $n^T(\hat{\mathbf{x}})F_2(\hat{\mathbf{x}}) < 0$, we have an *unstable sliding motion* along $\mathcal{H}$.
- If only $n^T(\hat{\mathbf{x}})F_1(\hat{\mathbf{x}}) = 0$, then the trajectory continues in the region pointed by $F_2(\hat{\mathbf{x}})$, and similarly for $n^T(\hat{\mathbf{x}})F_2(\hat{\mathbf{x}}) = 0$ (*tangential crossing*).

[12] A set function $\bar{F}$ is locally bounded at $\hat{\mathbf{x}} \in \mathbb{R}^n$ if there exists $\varepsilon > 0$ and $M_{\hat{\mathbf{x}}} > 0$ such that $||\hat{\mathbf{z}}|| < M_{\hat{\mathbf{x}}}$ for each $\hat{\mathbf{z}} \in \bar{F}(\hat{\mathbf{y}})$ and $\hat{\mathbf{y}} \in B(\hat{\mathbf{x}}, \varepsilon)$.

[13] A set function $\bar{F}$ is upper semicontinuous at $\hat{\mathbf{x}} \in \mathbb{R}^n$ is for each $\varepsilon > 0$ there exists $\delta > 0$ such that $\bar{F}(\hat{\mathbf{y}}) \subset \bar{F}(\hat{\mathbf{x}}) + B(\mathbf{0}, \varepsilon)$ for each $\hat{\mathbf{y}} \in B(\hat{\mathbf{x}}, \delta)$.

For the scope of this paper, the uniqueness result for PSW systems plays a relevant role. More precisely, Filippov [33] proved that there is a *unique* solution starting in $\hat{\mathbf{x}} \in \mathcal{H}$, provided that at least one of $n^T(\hat{\mathbf{x}})F_1(\hat{\mathbf{x}}) < 0$ and $n^T(\hat{\mathbf{x}})F_2(\hat{\mathbf{x}}) > 0$ holds. Notice that this condition rules out unstable sliding motion. There is also a condition for existence and uniqueness expressed in terms of differential inclusions, which requires the set-valued function $\bar{F}$ to be one-sided Lipschitz continuous[14]. We remark that the global existence and uniqueness of a solution allows us to define a semiflow $\phi(t, \hat{\mathbf{x}})$ for the discontinuous vector field F, which is the condition required in the definition of a PDMP adopted here.

7.2 Deterministic Approximation for PWS Limits

We will now present the limit result of [20] in the framework of this paper, and then plug it in the proof of Theorem 5.1 in order to extend the hybrid convergence limit to this discontinuous setting. We start by expanding Scaling 1 to deal with guards.

Scaling 6 (Continuous Scaling with Guards). A normalized guarded **sCCP** transition $\hat{\pi} = (\hat{\mathbf{g}}_\pi^{(N)}(\hat{\mathbf{X}}), \hat{\mathbf{X}}' = \hat{\mathbf{X}} + \hat{\nu}_\pi^{(N)}, \hat{\lambda}_\pi^{(N)}(\hat{\mathbf{X}}))$ of a population-**sCCP** program $(\mathcal{A}, \gamma_N)$, with E the domain of normalised variables $\hat{\mathbf{X}}$, assumed to be all continuous, has *continuous scaling* if and only if:

1. The rate and update satisfy the conditions of Scaling 1;
2. $\hat{\mathbf{g}}_\pi^{(N)}(\hat{\mathbf{X}})$ is of the form $h_{\pi,1}(\hat{\mathbf{X}}) \geq 0 \wedge \ldots \wedge h_{\pi,k}(\hat{\mathbf{X}}) \geq 0$, where each $h_{\pi,j}$ is independent of N and has continuous second order derivatives, with $\nabla h_{\pi,j}(\hat{\mathbf{x}}) \neq 0$ for all $\hat{\mathbf{x}} \in \{\hat{\mathbf{x}} \mid h(\hat{\mathbf{x}}) = 0\}$.

Consider now a population-**sCCP** program $(\mathcal{A}, \gamma_N)$, with no discrete variables and no instantaneous transitions, and with stochastic actions satisfying either Scaling 1 or 6 (hence, all actions will be approximated continuously). Then, we can compute its drift according to equation 5, which defines a piecewise-smooth system:

$$F^{(N)}(\hat{\mathbf{X}}) = \sum_\pi \mathbb{E}[\hat{\nu}_\pi^{(N)}]\mathbf{I}\{\hat{\mathbf{g}}_\pi(\hat{\mathbf{X}})\}\hat{\mathbf{f}}_\pi^{(N)}(\hat{\mathbf{X}})$$

Notice, in particular that, as the guards are independent of N, it holds that $F^{(N)} \to F$ uniformly, where F is defined by

$$F(\hat{\mathbf{X}}) = \sum_\pi \mathbb{E}[\hat{\nu}_\pi]\mathbf{1}\{\hat{\mathbf{g}}_\pi(\hat{\mathbf{X}})\}\hat{\mathbf{f}}_\pi(\hat{\mathbf{X}}).$$

Let us take a closer look to the PWS systems $\frac{d}{dt}\hat{\mathbf{x}} = F(\hat{\mathbf{x}})$. Each transition of the model having a non-trivial guard, partitions the state space in several regions. In fact, if the predicate $\hat{\mathbf{g}}_\pi$ is a conjunction of inequalities defined by smooth functions $h_{\pi,j}$, then each such function partitions the state space in two regions: $\mathcal{R}_{\pi,j}^+$, where $h_{\pi,j}$ is positive, and $\mathcal{R}_{\pi,j}^-$, where $h_{\pi,j}$ is negative. Therefore, in order to define the PWS system, we have to consider all possible intersections of regions $\mathcal{R}_j^+$ and $\mathcal{R}_j^-$, for all distinct function h_j appearing in guards of transitions. If there are m_0 such functions, then we have 2^{m_0} distinct regions. In practice, however, many transitions usually have trivial guards, and there may be transitions sharing the same functions h_j, so that this number should be reasonably small. In the following, we indicate by $\mathcal{H}_j$ the manifold defined by h_j: $\mathcal{H}_j = \{\hat{\mathbf{x}} \in \mathbb{R}^n \mid h_j(\hat{\mathbf{x}}) = 0\}$.

In the rest of the paper, we require that the PWSS defined by F is globally regular, in the following sense:

1. solutions exist globally in E, and are unique, so that the PWSS admits a semi-flow on E;
2. sliding motion never happens on the intersection of more than one surface, and has first order exit condition;
3. the PWSS has no Zeno trajectories, i.e. the number of transversal crossings and traits of sliding motion is finite in each compact time interval $[0, T]$, for each trajectory of the PWSS.

[14] A set valued function F is one-sided Lipschitz if and only if for each $\mathbf{x}_1, \mathbf{x}_2 \in E$ and $\mathbf{y}_1 \in F(\mathbf{x}_1)$, $\mathbf{y}_2 \in F(\mathbf{x}_2)$, it holds that $(\mathbf{x}_1 - \mathbf{x}_2)^t \cdot (\mathbf{y}_1 - \mathbf{y}_2) \leq L\|\mathbf{x}_1 - \mathbf{x}_2\|$, for some $L > 0$.

These conditions are essentially those introduced in [20], just extended to the whole domain E.[15] If the PWSS is regular, then each of its trajectories is regular and therefore the following theorem holds:

Theorem 7.1. *Let $(\mathcal{A}, \gamma_N)$ be a sequence of **sCCP** models for increasing systems size, satisfying the conditions of Theorem 4.1, with all actions π satisfying either Scaling 1 or Scaling 6. Let $\hat{\mathbf{X}}^{(N)}(t)$ be the associated sequence of normalized CTMC, and assume $\hat{\mathbf{x}}^{(N)}(0) \to \hat{\mathbf{x}}_0$ (in probability/almost surely).*

Let $\hat{\mathbf{x}}(t)$ be the solution of the regular PWSS system $\frac{d}{dt}\hat{\mathbf{x}} = F(\hat{\mathbf{x}})$ starting in point $\hat{\mathbf{x}}(0) = \hat{\mathbf{x}}_0 \in E$. Fix a finite time horizon $T < \infty$, Then

$$\lim_{N\to\infty} \sup_{t\le T} \left\| \hat{\mathbf{X}}^{(N)}(t) - \hat{\mathbf{x}}(t) \right\| = 0 \quad \textit{in probability.}$$

□

As an immediate corollary, we get that if $\hat{\mathbf{x}}_0^{(N)} \Rightarrow \hat{\mathbf{x}}_0$, then $\hat{\mathbf{X}}^{(N)} \Rightarrow \hat{\mathbf{x}}$ as random elements in the space of cadlag functions.

The intuition behind the proof of the theorem is that each regular trajectory in $[0, T]$ of the PWSS can be sliced into a finite number of pieces, such that each piece is either the solution of a standard ODE within a continuity region of the vector field, or it is a sliding motion along a discontinuity surface. The idea is to prove the convergence of the sequence $\hat{\mathbf{X}}^{(N)}(t)$ to $\hat{\mathbf{x}}(t)$ in each piece separately, either using standard deterministic approximation (Theorem 4.1), or using a specialized version of such a result for sliding motion (Theorem IV.2 in [20]). Then, one simply proves convergence in $[0, T]$ by combining the convergence in each piece, exploiting convergence of exit times. Sliding motion is the difficult case, because the trajectory of the PWSS evolves according to the sliding vector field, which is different from the drift $F^{(N)}$ of the sequence of CTMC.

Differently from [20], we are not requiring that rate functions are globally bounded and Lipschitz on E, but they just satisfy these properties locally. However, the validity of the previous theorem depends only on a compact neighbourhood of the trajectory up to time T, and on the assumption of global existence of solutions. Furthermore, we are also allowing random increments, which can be dealt with exactly as in Theorem 4.1.

We also note that we could have relaxed the scaling condition on guards by making guard functions h_j depend on N, and assuming that they converge uniformly to a limit function, with the same properties. The previous theorem would still hold, but with some modification with respect to the proof of [20].[16]

7.3 Hybrid Limit with Guarded Continuous Transitions

The previous theorem can be easily plugged in the framework of Section 5, replacing Kurtz theorem in the proof device of Theorem 5.1. This can be done under the *regularity assumption* of the limit PWSS in each mode of the PDMP, i.e. for each possible combination of values of discrete variables. We call *PWS regular* such a PDMP. The reason is that the proof of Theorem 5.1 relies only on the weak convergence implied by Kurtz theorem and on the continuity of limit trajectories, which are also satisfied by the subset of PWSS considered here. Furthermore, as the treatment of instantaneous transitions or time-dependent guards in Section 6 is also independent of the fine-grained details of the continuous dynamics, we can also include those kinds of transitions. Notice that the notion of non-Zeno PDMP and those of robustly transversal PDMP, robust activation property and size-compatible PDMP, extend automatically to this PWS setting,

[15] This regularity condition can be simplified, if the problem is restated in terms of differential inclusions and we require that the set-values extension $\bar{F}$ of F is one-sided Lipschitz. Under this milder assumptions, the following theorem still holds. However, we stick here to the formulation in terms of PWSS, which is more natural in the context of PDMP.

[16] The proof becomes more involved because if guards are varying with N, $F^{(N)}$ does not converge uniformly to F any more. Essentially, one proves that the trajectories of the PWSS defined by the discontinuous vector field $F^{(N)}$ converge uniformly in each $[0, T]$ to the trajectories of F, and that $\hat{\mathbf{X}}^{(N)}$ converges in probability, uniformly in $[0, T]$, to $\hat{\mathbf{x}}^{(N)}$, the solution of $d/dt\ \hat{\mathbf{x}}^{(N)}(t) = F^{(N)}(\hat{\mathbf{x}}^{(N)}(t))$. To show that $\hat{\mathbf{x}}^{(N)}$ is regular, one relies on the regularity of the limit trajectory $\hat{\mathbf{x}}$, and on the uniform convergence of activation functions of guards and of the components $\hat{\mathbf{f}}_\pi^{(N)}$ of the vector field. To show convergence of $\hat{\mathbf{X}}^{(N)}$ to $\hat{\mathbf{x}}^{(N)}$, one either invokes an obvious modification of Theorem 4.1, or modifies the proof for sliding motion in [20] using again the uniform convergence of guard's functions and of rates. Alternatively, one could work with differential inclusions, as in [35].

as they only depend on the existence of the semi-flow of the continuous dynamics of the PDMP. Before stating the theorem, we need to extend Scaling 6 to the hybrid setting, as done for scaling 1.

Scaling 7 (Hybrid Continuous Scaling with Guards). A normalized guarded **sCCP** transition $\hat{\pi} = (\hat{\mathbf{g}}_{\pi}^{(N)}(\hat{\mathbf{X}}), \hat{\mathbf{X}}' = \hat{\mathbf{X}} + \hat{\nu}_{\pi}^{(N)}, \hat{\lambda}_{\pi}^{(N)}(\hat{\mathbf{X}}))$ of a population-**sCCP** program $(\mathcal{A}, \gamma_N)$, with variables $\hat{\mathbf{X}} \in E$ partitioned into $(\mathbf{X}_d, \hat{\mathbf{X}}_c, \mathbf{X}_e)$, has *hybrid continuous scaling* if and only if:

1. The rate and update satisfy the conditions of Scaling 6;
2. $\hat{\mathbf{g}}_{\pi}^{(N)}(\hat{\mathbf{X}}) = \hat{\mathbf{g}}_{\pi,d}^{(N)}(\mathbf{X}_d, \mathbf{X}_e) \wedge \hat{\mathbf{g}}_{\pi,c}(\hat{\mathbf{X}}_c)$, where $\hat{\mathbf{g}}_{\pi,c}(\hat{\mathbf{X}}_c)$ satisfies the condition of Scaling 6.

Therefore, we have the following:

Proposition 7.1. *Let $(\mathcal{A}, \gamma_N)$ be a sequence of time-guarded population-**sCCP** models for increasing systems size $\gamma_N \to \infty$, as $N \to \infty$, with variables partitioned into $\mathbf{X} = (\mathbf{X}_d, \mathbf{X}_c, \mathbf{X}_e)$, with discrete stochastic actions satisfying scaling 3, instantaneous actions satisfying scaling 4, time guarded actions satisfying scaling 5, and continuous actions satisfying either scaling 2 or scaling 7. Let $\hat{\mathbf{X}}^{(N)}(t)$ be the associated sequence of normalized CTMC and $\hat{\mathbf{x}}(t)$ be the limit PDMP associated with the limit normalized TDSHA $\hat{\mathcal{T}}(\mathcal{A})$.*

If $\hat{\mathbf{x}}_0^{(N)} \Rightarrow \hat{\mathbf{x}}_0$ (weakly) and the PDMP is non-Zeno*,* robustly transversal*, has the* robust activation property*, is* size-compatible *and* PWS regular*, then $\hat{\mathbf{X}}^{(N)}(t)$ converges weakly to $\hat{\mathbf{x}}(t)$, $\hat{\mathbf{X}}^{(N)} \Rightarrow \hat{\mathbf{x}}$, as random elements in the space of cadlag function with the Skorohod metric.* □

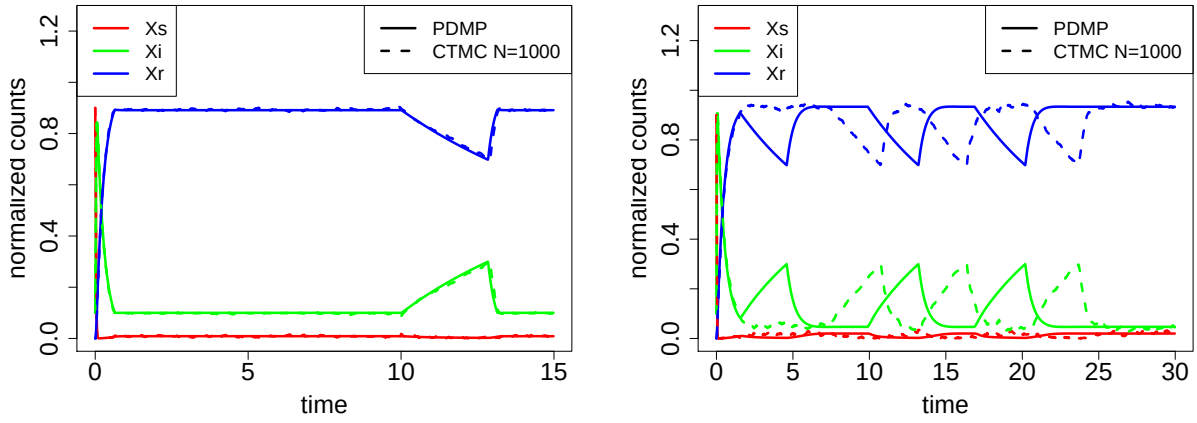
Example 7.1. We consider a variant of the computer network epidemic model of Example 6.4. In particular, we consider the following normal and emergency patching policies. Under the normal policy, patches are applied at constant rate k_p^1. Under the emergency policy, instead, computers in the network are patched with a rate $k_p^2 > k_p^1$ if the fraction $\hat{X}_i$ of infected nodes is above a threshold α, and at rate $k_p^3 < k_p^2$, $k_p^3 > k_p^1$, if the fraction of infected nodes is below α. The emergency policy is initiated as soon as the fraction of infected nodes becomes greater than a threshold $\beta > \alpha$, and is executed for $w \in \mathbb{R}^+$ units of time. We will use an environmental variable K to remember the next firing time of such a delayed transition. When the emergency policy is aborted, the normal policy is restored. We can model this policy in **sCCP** by suitably modifying the code of Example 6.4, with particular regard to the `patching` and the `control` agents. The variable U is the discrete variable modelling the patching policy: $U = 1$ indicates the normal policy, while $U = 2$ indicates the emergency one.

$$
\begin{array}{lcl}
\texttt{patching} & \stackrel{\text{def}}{=} & [U = 1 \to X_i' = X_i - 1 \wedge X_r' = X_r + 1]_{k_p^1 X_i}.\texttt{patching} \\
& + & [U = 2 \wedge \hat{X}_i \geq \alpha \to X_i' = X_i - 1 \wedge X_r' = X_r + 1]_{k_p^2 X_i}.\texttt{patching} \\
& + & [U = 2 \wedge \hat{X}_i < \alpha \to X_i' = X_i - 1 \wedge X_r' = X_r + 1]_{k_p^3 X_i}.\texttt{patching} \\
\texttt{control} & \stackrel{\text{def}}{=} & [\hat{X}_i > \beta \to U' = 2 \wedge K' = time + w]_{\infty:1}.\texttt{control} \\
& + & [time \geq K \to U' = 1]_{\infty:1}.\texttt{control}
\end{array}
$$

The limit model is a PWSS when $U = 2$. In Figure 11(a), we show a trajectory of the PWSS exhibiting sliding motion on the plane $\hat{X}_i = \alpha$. It is easy to check that the PWSS has a unique solution from any initial state. In fact, taking the scalar product of the two vector fields F_1, F_2 with the normal to $\hat{X}_i = \alpha$ on the two sides of the plane $\hat{X}_i = \alpha$, we obtain $k_i \alpha \hat{X}_s - k_p^2 \alpha$ and $k_i \alpha \hat{X}_s - k_p^3 \alpha$. Now, $k_i \alpha \hat{X}_s - k_p^3 \alpha > 0$ for $\hat{X}_s > k_p^3 / k_i$. But as $k_p^2 > k_p^3$, for $\hat{X}_s \leq k_p^3 / k_i$, we have that $k_i \alpha \hat{X}_s - k_p^2 \alpha \leq \alpha k_p^3 - k_p^2 \alpha < 0$, which shows that the uniqueness condition is verified. Also in this case, the hybrid limit model is deterministic, and we can see its trajectory from a fixed set of initial conditions in Figure 11. Inspecting this trajectory, we can easily convince ourselves that the crossing of activation surfaces is always transversal, so that we can apply Proposition 7.1.

7.4 Guards on discrete stochastic transitions

In this section, we consider a relaxation of Scaling 3, in which we allow guards on discrete stochastic transitions to depend on continuous variables. Similarly to continuous transitions, this extension introduces discontinuities in the rate functions of the PDMP. Intuitively, as the jump time distribution is obtained by



(a) Example 7.1

(b) Example 7.3

Figure 11: Comparison of single trajectories of the limit PDMP and the CTMC model for system size 1000, for the control policies of the network epidemic models of Example 7.1 (left) and Example 7.3 (right). The behaviour is essentially the same (modulo randomness of switch times in the figure on the right). Parameters are as in the caption of Figure 8. Additionally, in the model of Example 7.1, we have $k_p^1 = 0.05$, $k_p^2 = 4.0$, $k_p^3 = 0.5$, $\alpha = 0.1$, $\beta = 0.3$, and the duration of the emergency policy is 10 units of time. the rate of switch from emergency to normal patching policy in the model of Example 7.3 is 0.1.

the cumulative rate, i.e. by integrating the rate function, these discontinuities should not create problems, as far as the trajectories of the PDMP do not remain in a discontinuity surface of a rate for too long. Essentially, problems emerge if a continuous trajectory of the PDMP slides on the discontinuity surface of a guard for some time interval $[t_1, t_2]$. Suppose that on the surface the guard is false, hence the transition of the PDMP cannot fire. In this case, even if trajectories of the CTMC converge to the one of the PDMP, they may remain on the "wrong" side of the discontinuity surface, i.e. on the side in which the guard is true and the transition active, so that the event can fire in the sequence of CTMC. If this transition determines the fate of the system, than convergence to the PDMP can fail quite dramatically.

To explain better the problem, we consider the following example.

Example 7.2. Consider a simple **sCCP** model of a random walk in 1 dimension, for variable X, initially set to zero. Variable Z instead, can take values 0 and 1, and it is the fate variable. Initially it is set to zero, and it may become 1 by the firing of a stochastic transition with rate 1, but active only if $X > 0$. The **sCCP** program has initial configuration `random_walk` $\parallel$ `doom`, where

$$\begin{array}{lcl} \texttt{random\_walk} & \stackrel{\text{def}}{=} & [* \to X' = X + 1]_{\gamma_N} \texttt{.random\_walk} \\ & + & [* \to X' = X - 1]_{\gamma_N} \texttt{.random\_walk} \\ \texttt{doom} & \stackrel{\text{def}}{=} & [X > 0 \to Z = 1]_1 \texttt{.0} \end{array}$$

Notice that the rate of the transitions of the `random_walk` agent grow with γ_N, hence they can be approximated continuously, and, once normalized, induce the drift $F^{(N)}(\hat{\mathbf{X}}) = F(\hat{\mathbf{X}}) = 0$. Hence, the limit PDMP model has quite boring continuous dynamics, in fact a constant one on the discontinuity surface $\hat{x} = 0$. It follows that the discrete stochastic transition will never fire in the PDMP model, and Z will remain 0. However, for any N, the CTMC model will spend half of its time on the subspace $\hat{X} > 0$, meaning that the `doom` agent will eventually fire its transition, on average in 2 time units. Hence Z will be equal to 1 with probability going to 1 as T increases. Hence convergence does not hold for this model. Notice, however, that if we set the initial value of $\hat{x}$ to $-\varepsilon$, for $\varepsilon > 0$, then convergence will hold. In particular, by the Kurtz theorem, for any time $T < \infty$, the CTMC $\hat{\mathbf{X}}^{(N)}(t)$ will be smaller than $-\varepsilon/2$, in $[0, T]$, for N large enough, hence the `doom` transition will not fire in $[0, T]$. However, it will eventually fire for any N, as with probability

one, $\hat{\mathbf{X}}^{(N)}(t)$ will get eventually above zero and remain there for a long enough time. In addition, the time when this event will happen is pushed further and further into the future as N grows. This does not create problems for weak convergence, as the Skorohod metric on which weak convergence is based discounts the future and will give a smaller and smaller weight to the difference of Z-values as N grows (see Appendix C for details on the Skorohod metric).

From the previous discussion, it should be clear that the problems in introducing guards for discrete stochastic transitions are somehow related to the way the flow of the vector field interacts with the discontinuity surface of the guards. This suggests the following definition:

Definition 7.1. A cadlag function $\hat{\mathbf{x}}(t)$ taking values in E is *robustly compatible* with the activation function $h(\hat{\mathbf{x}})$ of a guard predicate $\hat{\mathbf{g}}(\hat{\mathbf{x}})$ if and only if the set $\{t \geq 0 \mid h(\hat{\mathbf{x}}(t)) = 0\}$ has Lebesgue measure zero.
A PDMP is *robustly compatible* with a guard $\hat{\mathbf{g}}_\pi(\hat{\mathbf{x}})$ if almost surely its trajectories are robustly compatible with the activation function of $\hat{\mathbf{g}}_\pi(\hat{\mathbf{x}})$.
A PDMP derived from a TDSHA $\mathcal{T}$ is *robustly compatible* if and only if it is *robustly compatible* with all guards of discrete stochastic transitions of $\mathcal{T}$.

We can now introduce the scaling condition for guarded discrete stochastic transitions.

Scaling 8 (Discrete Scaling for Guarded Stochastic Transitions). A *guarded normalized* **sCCP** transition with *random reset*, $\hat{\pi} = (\hat{\mathbf{g}}_{\langle N \rangle}\pi(\hat{\mathbf{X}}), \hat{\mathbf{X}}' = \hat{\mathbf{r}}^{(N)}(\hat{\mathbf{X}}, \mathbf{W}^{(N)}(\hat{\mathbf{X}})), \hat{\lambda}_\pi^{(N)}(\hat{\mathbf{X}}))$ of a population-**sCCP** program $(\mathcal{A}, \gamma_N)$ with variables $\hat{\mathbf{X}} \in E$ partitioned into $(\mathbf{X}_d, \hat{\mathbf{X}}_c, \mathbf{X}_e)$, has *discrete scaling* if and only if:

1. $\hat{\mathbf{g}}_\pi^{(N)}(\hat{\mathbf{X}})$ has activation function $h^{(N)}(\hat{\mathbf{X}})$, converging uniformly in every compact subset K of E to the continuous function $h(\hat{\mathbf{X}})$;

2. $\hat{\lambda}_\pi^{(N)}$ and $\hat{\mathbf{r}}_\pi^{(N)}$ satisfy the same conditions as Scaling 3.

Technically, the condition of robust compatibility of a PDMP is needed to prove the convergence of the stochastic jump times and of the states after the reset. The problem here lies in the fact that, on the surface $\{h(\hat{\mathbf{x}}) = 0\}$, the reset kernel R of the PDMP is discontinuous, and a sequence $\hat{\mathbf{x}}^{(N)}$ of points approaching $\hat{\mathbf{x}} \in \{h(\hat{\mathbf{x}}) = 0\}$ may activate a different subset of guards for each $R^{(N)}$. This may lead to radically different behaviours and compromise convergence. This problem is essentially the same as we had with the discontinuity of the reset kernel for instantaneous transitions. The robust compatibility condition permits us to ignore such points, as there is probability zero of jumping from them. With these assumptions in force, we get the following proposition, whose proof can be found in Appendix D.

Proposition 7.2. *Let $(\mathcal{A}, \gamma_N)$ be a sequence of population-**sCCP** models for increasing systems size $\gamma_N \to \infty$, as $N \to \infty$, with variables partitioned into $\mathbf{X} = (\mathbf{X}_d, \mathbf{X}_c, \mathbf{X}_e)$, with discrete stochastic actions satisfying either scaling 3 or scaling 8, no instantaneous actions, and continuous actions satisfying scaling 2. Let $\hat{\mathbf{X}}^{(N)}(t)$ be the associated sequence of normalized CTMC and $\hat{\mathbf{x}}(t)$ be the limit PDMP associated with the normalized limit TDSHA $\hat{\mathcal{T}}(\mathcal{A})$.*

If $\hat{\mathbf{x}}_0^{(N)} \Rightarrow \hat{\mathbf{x}}_0$ (weakly) and the PDMP is non-Zeno *and* robustly compatible*, then $\hat{\mathbf{X}}^{(N)}(t)$ converges weakly to $\hat{\mathbf{x}}(t)$, $\hat{\mathbf{X}}^{(N)} \Rightarrow \hat{\mathbf{x}}$, as random elements in the space of cadlag function with the Skorohod metric.* □

Example 7.3. We consider again the computer network epidemics model of Examples 6.4 and 7.1. In particular, we modify the control policy with respect to Example 7.1 by assuming that the emergency policy is dropped in favour of the normal one in an exponentially distributed time, with rate k_d, provided $\hat{X}_i$ is below $\beta_2 < \beta$. The control agent then becomes

$$\begin{array}{rcl} \texttt{control} & \stackrel{\text{def}}{=} & [U = 1 \wedge \hat{X}_i > \beta \to U' = 2]_{\infty:1}\,.\,\texttt{control} \\ & + & [U = 2 \wedge \hat{X}_i < \beta_2 \to U' = 1]_{k_d}\,.\,\texttt{control} \end{array}$$

Furthermore, we assume that there is a single emergency patch rate $k_p^2 > k_p^1$, so that the patch agent is

$$\begin{array}{rcl} \texttt{patching} & \stackrel{\text{def}}{=} & [U = 1 \to X_i' = X_i - 1 \wedge X_r' = X_r + 1]_{k_p^1 X_i}\,.\,\texttt{patching} \\ & + & [U = 2 \to X_i' = X_i - 1 \wedge X_r' = X_r + 1]_{k_p^2 X_i}\,.\,\texttt{patching} \end{array}$$

If we take the scalar product between the vector field and the normal $\mathbf{1}_i$ to the plane $\hat{X}_i - \beta_2 = 0$ and set it to zero in the plane, we get the equation $k_i \hat{X}_s - k_p^2 = 0$, which has only one solution $(k_p^2/k_i, \beta_2, 1 - k_p^2/k_i - \beta_2)$ if $k_p^2/k_i + \beta_2 \leq 1$, and no solution otherwise. In particular, it follows that the trajectories of the vector field do not slide on $\hat{X}_i = \beta_2$, hence the PDMP is robustly compatible. Thus, the model satisfies the hypothesis of Proposition 7.2, and convergence to the hybrid limit holds.

*Remark** 7.1. In order to check the robust compatibility of a PDMP with respect to a guard of a stochastic transition, we can proceed similarly to Remark 6.1, by applying the randomization trick. In particular, the property holds if we can show that the set of trajectories of the vector field sliding on the discontinuous surface has dimension $n-1$ or less,[17] where n is the number of continuous variables, and if initial conditions and resets are absolutely continuous with respect to the Lebesgue measure. If guards are linear, this check should be relatively easy to carry out, see Example 7.3.

7.5 Collecting all results together.

In this subsection, we collect in a unique statement all the approximation results spread throughout the paper.

Theorem 7.2. *Let $(\mathcal{A}, \gamma_N)$ be a sequence of time-guarded population-**sCCP** models for increasing system size $\gamma_N \to \infty$, as $N \to \infty$, with variables partitioned into $\mathbf{X} = (\mathbf{X}_d, \mathbf{X}_c, \mathbf{X}_e)$, satisfying the following scaling conditions:*

1. *discrete stochastic actions with guards* not *depending on continuous variables satisfy scaling 3;*
2. *discrete stochastic actions with guards depending on continuous variables satisfy scaling 8;*
3. *instantaneous actions satisfy scaling 4;*
4. *time-guarded actions satisfy scaling 5;*
5. *continuous actions with guards* not *depending on continuous variables satisfy scaling 2;*
6. *continuous actions with guards depending on continuous variables satisfy scaling 7;*

Let $\hat{\mathbf{X}}^{(N)}(t)$ be the associated sequence of normalized CTMC and $\hat{\mathbf{x}}(t)$ be the limit PDMP associated with the limit normalized TDSHA $\hat{\mathcal{T}}(\mathcal{A})$. If

1. *the PDMP is* non-Zeno*;*
2. *the PDMP is* robustly transversal*, has the* robust activation property*, and is* size-compatible *(for instantaneous transitions);*
3. *the PDMP is* robustly compatible *with all guards of discrete stochastic actions;*
4. *the PDMP is* PWS regular *(if continuous transitions guarded by continuous variables are present);*
5. $\hat{\mathbf{x}}_0^{(N)} \Rightarrow \hat{\mathbf{x}}_0$ *(weakly)*

then $\hat{\mathbf{X}}^{(N)}(t)$ converges weakly to $\hat{\mathbf{x}}(t)$, $\hat{\mathbf{X}} \Rightarrow \hat{\mathbf{x}}$, as random elements in the space of cadlag function with the Skorohod metric. □

[17] This holds, for instance, if the set of zeros of the scalar product of the normal to the discontinuity surface of the guard with the vector field has dimension $n-2$ or less.

8 Conclusions

In this paper, we discussed hybrid limits of Markov population processes, described as stochastic concurrent constraint programs. We considered the limit behaviour when only a subset of system variables, corresponding to populations of growing size, is approximated continuously, while the other variables remain discrete. We proved that the sequence of CTMC for increasing population size converges to a stochastic hybrid system, described as a Piecewise-Deterministic Markov Process.

We first considered the simplest case, in which continuous transitions, i.e. those becoming fluxes of the limit vector field, satisfy the standard scaling *à la Kurtz*, while discrete transitions are stochastic and have continuous rates and reset kernels.

We then extended these results by including several sources of discontinuity in the evolution of the system: instantaneous transitions in the **sCCP** program, which induce forced transitions at the PDMP level, and guards depending on continuous variables in continuous and stochastic transitions. In all these cases, discontinuities create potential problems in the interactions between the deterministic vector field in each mode of the PDMP and the discontinuity surfaces of the involved functions. Essentially, the reset kernels become non-Feller (i.e. non-continuous), and one has to impose additional conditions to enforce convergence of the times at which transitions fire and the states of the system after a jump. In general, the conditions required can be quite difficult to check. However, in practical cases the geometry of discontinuous surfaces should be reasonably simple (mainly linear hyperplanes), hence checking the required conditions should not be too hard.

Nevertheless, it is unlikely that one can find general algorithmic procedures to check automatically if a sequence of models is amenable of hybrid approximation, unless there is no source of discontinuity and rate functions and resets satisfy further constraints, see Remark 5.3.

The moral is that one should avoid introducing too many discontinuities in a model, if approximation results are needed to perform more efficient analysis.

In this direction, we are investigating some relaxation techniques. In most of the cases, boolean conditions may be replaced by smooth counterparts without altering significantly the model behaviour. For instance, instantaneous events may be replaced by stochastic ones with a rate that changes continuously from zero to a very large value in the proximity of the activation surface, as done in [1]. Moreover, guards in discrete and stochastic transitions may be replaced by sharp sigmoid functions modulating the rates. However, this operation is not always possible without introducing spurious behaviours. In this case, one has to verify additional regularity conditions before using hybrid approximation.

Another strategy to simplify the conditions required for the hybrid limits, similar in spirit to the previous one, is to introduce randomness in the continuous evolution. In particular, we could replace the vector field of the PDMP by a stochastic differential equation, obtaining a Stochastic Hybrid System in the sense of [22]. The simplest possibility is to perturb the trajectories of the vector field with Gaussian noise, obtaining the so-called *central-limit approximation*, for which a limit result analogous to Theorem 4.1 exists (see [42], Chapter 11). This fact guarantees that convergence proofs presented in this paper extend straightforwardly to this new setting. Furthermore, in doing this, we get the advantage of removing the bad behaviours happening at the discontinuity boundary. Intuitively, in the central-limit regime, the probability of a trajectory to slide on a surface of dimension $n-1$ is zero. Similarly, the probability that a trajectory tangentially hits a surface is zero. It follows that in this setting, most of the additional conditions on PDMP required for convergence hold almost automatically. On the downside, simulating a SDE is more expensive from a computational point of view, although the regularity of Gaussian Processes may be exploited to improve efficiency. We are currently investigating this direction.

We also plan to investigate the definition of algorithms to check the conditions for convergence and to suggest a partition of variables into discrete and continuous (also for models in which the dependency on system size is not explicit).

An important question related to approximation theorems is if the weak convergence result, which are limited to the transient dynamics, can be extended to steady state. In the deterministic case, this can be done only in a limited number of cases. Essentially, convergence of steady state depends on the phase space properties of the limit ODE, and it is guaranteed only in presence of a unique globally attracting steady state, see e.g. [5, 7]. In the future, we would like to investigate if similar results can be found for the hybrid limit case. The situation in this case is more complex. On the one hand, if the limit process

is deterministic, then we can exploit recent results [6], provided we can characterise invariant measures for deterministic hybrid systems. On the other hand, if the limit process is stochastic, one has to prove that it does indeed have a steady state, and extend the result on invariant measures from the deterministic case to the stochastic one. In this setting, computing the invariant measure of the PDMP can be quite challenging in itself [27].

Finally, we want to understand what happens if we include non-determinism in the framework, especially in terms of uncertainty on parameters. This would require consideration of stochastic hybrid systems combining differential inclusions [3] with imprecise probabilities [63].

Acknowledgements. Work partially supported by "FRA-UniTS" grant.

References

[1] A. Abate, M. Prandini, J. Lygeros, and S. Sastry. Approximation of general stochastic hybrid systems by switching diffusions with random hybrid jumps. In Magnus Egerstedt and Bud Mishra, editors, *Proceedings of Hybrid Systems: Computation and Control, HSCC 2008*, volume 4981 of *Lecture Notes in Computer Science*, pages 598–601. Springer, 2008.

[2] H. Andersson and T. Britton. *Stochastic Epidemic Models and Their Statistical Analysis*. Springer-Verlag, 2000.

[3] J. Aubin and A. Cellina. *Differential Inclusions*. Springer-Verlag, 1984.

[4] K. Ball, T. G. Kurtz, L. Popovic, and G. Rempala. Asymptotic analysis of multiscale approximations to reaction networks. *Ann. Appl. Probab.*, 16(4):1925–1961, 2006.

[5] M. Benaïm and J. Le Boudec. A class of mean field interaction models for computer and communication systems. *Performance Evaluation*, 2008.

[6] M. Benaïm and J.Y. Le Boudec. On mean field convergence and stationary regime. *CoRR*, abs/1111.5710, 2011.

[7] M. Benaïm and J. Weibull. Deterministic approximation of stochastic evolution in games. *Econometrica*, 2003.

[8] P. Billingsley. *Probability and Measure*. John Wiley and Sons, 1979.

[9] P. Billingsley. *Convergence of Probability Measures, 2nd Edition*. Wiley, 1999.

[10] L. Bortolussi. Stochastic concurrent constraint programming. In *Proceedings of 4th International Workshop on Quantitative Aspects of Programming Languages (QAPL 2006)*, volume 164 of *ENTCS*, pages 65–80, 2006.

[11] L. Bortolussi. Limit behavior of the hybrid approximation of stochastic process algebras. In *Proceedings of 17th International Conference on Analytical and Stochastic Modeling Techniques and Applications, ASMTA 2010*, volume 6148 of *Lecture Notes in Computer Science*, pages 367–381. Springer, 2010.

[12] L. Bortolussi, V. Galpin, J. Hillston, and M. Tribastone. Hybrid semantics for PEPA. In *Proceedings of 7th International Conference on the Quantitative Evaluation of Systems, QEST 2010*, pages 181–190. IEEE Computer Society, 2010.

[13] L. Bortolussi and J. Hillston. Fluid model checking. In *Proceedings of CONCUR 2012*, 2012.

[14] L. Bortolussi, J. Hillston, D. Latella, and M. Massink. Continuous approximation of collective systems behaviour: a tutorial. *Performance Evaluation*, submitted. Preliminary version available as a CNR/ISTI technical report.

[15] L. Bortolussi and A. Policriti. Modeling biological systems in concurrent constraint programming. *Constraints*, 13(1-2):66–90, 2008.

[16] L. Bortolussi and A. Policriti. Hybrid dynamics of stochastic π-calculus. *Mathematics in Computer Science*, 2(3):465–491, 2009.

[17] L. Bortolussi and A. Policriti. Hybrid dynamics of stochastic programs. *Theoretical Computer Science*, 411(20):2052–2077, 2010.

[18] L. Bortolussi and A. Policriti. (Hybrid) automata and (stochastic) programs. The hybrid automata lattice of a stochastic program. *Journal of Logic and Computation*, in print.

[19] L. Bortolussi and A. Policriti. Studying cancer-cell populations by programmable models of networks. *Network Modeling and Analysis in Health Informatics and Bioinformatics*, pages 1–17, in print.

[20] Luca Bortolussi. Hybrid limits of continuous time Markov chains. In *Proceedings of Eighth International Conference on the Quantitative Evaluation of Systems, QEST 2011*, pages 3–12. IEEE Computer Society, 2011.

[21] Luca Bortolussi and Alberto Policriti. Dynamical systems and stochastic programming: To ordinary differential equations and back. In Corrado Priami, Ralph-Johan Back, and Ion Petre, editors, *Transactions on Computational Systems Biology XI*, volume 5750 of *Lecture Notes in Computer Science*, pages 216–267. Springer Berlin / Heidelberg, 2009. 10.1007/978-3-642-04186-0_11.

[22] M.L. Bujorianu and J. Lygeros. General stochastic hybrid systems: Modeling and optimal control. In *Proceedings of 43rd IEEE Conference on Decision and Control (CDC 2004)*, pages 182–187, 2004.

[23] L. Cardelli. From processes to ODEs by chemistry. *downloadable from* `http://lucacardelli.name/`, 2006.

[24] F. Ciocchetta and J. Hillston. Bio-PEPA: A framework for the modelling and analysis of biological systems. *Theoretical Computer Science*, 410(33-34):3065–3084, 2009.

[25] J. Colom and M. Silva. Convex geometry and semiflows in P/T nets. a comparative study of algorithms for computation of minimal p-semiflows. In G. Rozenberg, editor, *Advances in Petri Nets 1990*, volume 483 of *LNCS*, pages 79–112. Springer Berlin / Heidelberg, 1991.

[26] J. Cortes. Discontinuous dynamical systems: A tutorial on solutions, nonsmooth analysis, and stability. *IEEE Control Systems Magazine*, pages 36–73, 2008.

[27] O. Costa and F. Dufour. Stability and ergodicity of piecewise deterministic Markov processes. *SIAM Journal on Control and Optimization*, 47(2):1053–1077, 2008.

[28] A. Crudu, A.Debussche, A.Muller, and O.Radulescu. Convergence of stochastic gene networks to hybrid piecewise deterministic processes. *Annals of Applied Probability*, in press.

[29] R.W.R. Darling. Fluid limits of pure jump Markov processes: A practical guide. `arXiv.org`, 2002.

[30] R.W.R. Darling and J.R. Norris. Differential equation approximations for Markov chains. *Probability Surveys*, 5, 2008.

[31] M.H.A. Davis. *Markov Models and Optimization*. Chapman & Hall, 1993.

[32] L. Dieci and L. Lopez. Sliding motion in Filippov differential systems: Theoretical results and a computational approach. *SIAM J. Numer. Anal.*, 47:2023–2051, 2009.

[33] A.F. Filippov. *Differential Equations with discontinuous right-hand sides*. Mathematics and Its Applications. Kluwer Academic, 1988.

[34] V. Galpin, L. Bortolussi, and J. Hillston. Hybrid modelling by composition of flows. *Formal Aspects of Computing*, in print.

[35] N. Gast and B. Gaujal. Mean field limit of non-smooth systems and differential inclusions. *SIGMETRICS Perform. Eval. Rev.*, 38:30–32, October 2010.

[36] R.A. Hayden, A. Stefanek, and J.T. Bradley. Fluid computation of passage-time distributions in large Markov models. *Theor. Comput. Sci.*, 413(1):106–141, 2012.

[37] Richard A. Hayden and Jeremy T. Bradley. A fluid analysis framework for a Markovian process algebra. *Theor. Comput. Sci.*, 411(22-24):2260–2297, 2010.

[38] J. Hillston. *A Compositional Approach to Performance Modelling*. Cambridge University Press, 1996.

[39] J. Hillston. Fluid flow approximation of PEPA models. pages 33 – 42, sept. 2005.

[40] A.F. Karr. Weak convergence of a sequence of Markov chains. *Probability Theory and Related Fields*, 33:41–48, 1975.

[41] S. Krantz and P.R. Harold. *A Primer of Real Analytic Functions (Second ed.)*. Birkhäuser, 2002.

[42] T. Kurtz and S. Ethier. *Markov Processes - Characterisation and Convergence*. Wiley, 1986.

[43] T. G. Kurtz. Solutions of ordinary differential equations as limits of pure jump Markov processes. *Journal of Applied Probability*, 7:49–58, 1970.

[44] T.G. Kurtz. *Approximation of population processes*. SIAM, 1981.

[45] M. Ajmone Marsan, G. Balbo, G. Conte S. Donatelli, and G. Franceschinis. *Modelling with Generalized Stochastic Petri Nets*. Wiley, 1995.

[46] M. Massink, D. Latella, A. Bracciali, M. D. Harrison, and J. Hillston. Scalable context-dependent analysis of emergency egress models. *Formal Asp. Comput.*, 24(2):267–302, 2012.

[47] M. Massink, D. Latella, A. Bracciali, and J. Hillston. Modelling non-linear crowd dynamics in Bio-PEPA. In Dimitra Giannakopoulou and Fernando Orejas, editors, *FASE*, volume 6603 of *Lecture Notes in Computer Science*, pages 96–110. Springer, 2011.

[48] S. Menz, J.C. Latorre, Ch. Schtte, and W. Huisinga. Hybrid stochasticñdeterministic solution of the chemical master equation. *SIAM Interdisciplinary Journal Multiscale Modeling and Simulation*, In press.

[49] J. R. Norris. *Markov Chains*. Cambridge University Press, 1997.

[50] J. Pahle. Biochemical simulations: stochastic, approximate stochastic and hybrid approaches. *Brief Bioinform.*, 10(1):53–64, 2009.

[51] D. Pollard. *Convergence of stochastic processes*. Springer-Verlag, 1984.

[52] W. Rudin. *Functional Analysis*. McGraw-Hill, 1973.

[53] W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, 1976.

[54] V. A. Saraswat. *Concurrent Constraint Programming*. MIT press, 1993.

[55] A. Shwartz and A. Weiss. *Large Deviations for Performance Analysis*. Chapman & Hall, 1995.

[56] A. Singh and J.P. Hespanha. Lognormal moment closures for biochemical reactions. In *Proceedings of 45th IEEE Conference on Decision and Control*, 2006.

[57] S. Soliman. Invariants and other structural properties of biochemical models as a constraint satisfaction problem. *Algorithms for Molecular Biology*, 7(15), 2012.

[58] D. J. T. Sumpter and D. S. Broomhead. Relating individual behaviour to population dynamics. *Proceedings of the Royal Society B*, 2001.

[59] P. Taylor. A lambda calculus for real analysis. *Journal of Logic and Analysis*, 2(5):1–115, 2010.

[60] Mirco Tribastone, Stephen Gilmore, and Jane Hillston. Scalable differential analysis of process algebra models. *IEEE Trans. Software Eng.*, 38(1):205–219, 2012.

[61] K. S. Trivedi and V. G. Kulkarni. Fspns: Fluid stochastic petri nets. In M. Ajmone Marsan, editor, *Application and Theory of Petri Nets*, volume 691 of *Lecture Notes in Computer Science*, pages 24–31. Springer, 1993.

[62] E. O. Voit. *Computational Analysis of Biochemical Systems.* Cambridge University Press, 2000.

[63] P. Walley. *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, 1991.

[64] D. J. Wilkinson. *Stochastic Modelling for Systems Biology.* Chapman & Hall, 2006.

A Notation

Here are some notational conventions followed throughout the paper.

- We denote the closure of a set A by $\bar{A}$ and its boundary by ∂A.
- We denote with $\mathbb{R}$ the reals, $\mathbb{Z}$ the integers, $\mathbb{N}$ the natural numbers, and $\mathbb{R}_{\geq 0}$ the non negative reals.
- Q indicates a countable subset of $\mathbb{R}^k$, the set of modes. E is the state space of the fluid limit (in this case $E \subseteq \mathbb{R}^n$) or of the hybrid limit, and in this case $E \subseteq Q \times \mathbb{R}^n$.
- n is the dimension of continuous variables, k is the dimension of discrete variables, m is the dimension of the full vector of variables.
- $\mathbf{X}$ denotes the non-normalized vector of variables. We assume an ordering of variables, so we can interchange sets and vectors of variables freely. $\mathbf{Y}$ or $\mathbf{X}_c$ are vectors of non-normalized continuous variables, while $\mathbf{Z}$ or $\mathbf{X}_d$ are vectors of discrete variables.
- $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$ denote points in the non-normalized space $\mathbb{R}^m$.
- $\hat{\mathbf{X}}$ denotes the normalized vector of variables, taking values in E. $\hat{\mathbf{Y}}$ or $\hat{\mathbf{X}}_c$ are vectors of normalized continuous variables.
- $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, $\hat{\mathbf{z}}$ denote points in the normalized space E.
- Given a vector function $f : \mathbb{R}^n \to \mathbb{R}^n$, with $f(\mathbf{X})$ we indicate the variables it depends to, while with $f[X](\mathbf{X})$ we indicate the function corresponding to variable X.
- γ and γ_N denote the system size.
- $(\mathcal{A}, \gamma_N)$ is a population-**sCCP** program. $\mathcal{T}(\mathcal{A})$ is the TDSHA associated with **sCCP** program $\mathcal{A}$. $\hat{\mathcal{T}}(\mathcal{A}, \gamma_N)$ is the TDSHA associated with the normalized population-**sCCP** program $(\mathcal{A}, \gamma_N)$ and $\hat{\mathcal{T}}(\mathcal{A})$ is the limit normalized TDSHA.
- π denotes a **sCCP** action, while η denotes a TDSHA transition.
- $\mathbf{X}^{(N)}(t)$ denotes a CTMC associated with a population **sCCP** model with system size γ_N, while $\hat{\mathbf{X}}^{(N)}(t)$ denotes the corresponding normalised CTMC.
- $\hat{\mathbf{x}}(t)$ denotes either the (normalised) limit PDMP or the fluid limit.
- With $B_\varepsilon(\hat{\mathbf{x}})$ we indicate the ball of radius ε centred in $\hat{\mathbf{x}}$, while $B_\varepsilon(\hat{\mathbf{x}}([0,T]) = \bigcup_{t\in[0,T]} B_\varepsilon(\hat{\mathbf{x}}(t))$.
- τ_i is the jump time of the i-th stochastic event in the PDMP. ζ_i is the jump time of the i-th instantaneous event. T_i is the jump time of the i-th event.

B Additional material on sCCP

In this appendix, we provide a few additional details about **sCCP**. We start by formalising the construction transforming a generic **sCCP** program $\mathcal{A}$ into a flat **sCCP**model $flat(\mathcal{A})$.

Definition B.1. Let $\mathcal{A} = (A, \mathsf{Def}, \mathbf{X}, \mathcal{D}, \mathbf{x_0})$ be a **sCCP** program. Its *flattened* version $flat(\mathcal{A}) = (A^+, \mathsf{Def}^+, \mathbf{X}^+, \mathcal{D}^+, \mathbf{x_0}^+)$ is constructed as follows:

- We add a new variable for each component $C \in \mathsf{Def}$: $\mathbf{X}^+ = \mathbf{X} \cup \mathbf{P}$, with $\mathbf{P} = \{P_C \mid C \in \mathsf{Def}\}$ taking values in $\mathbb{N}$.
- Each component C is replaced by a component C^+. If $C = \pi.A + M$, with $\pi = [g(\mathbf{X}), u(\mathbf{X}, \mathbf{X}', \mu)]_{\lambda(\mathbf{X}}$, then $C^+ = \pi^+.C^+ + M^+$, with $\pi^+ = [g^+(\mathbf{X}), u^+(\mathbf{X}, \mathbf{X}', \mu)]_{\lambda^+(\mathbf{X})}$.
- The guard of π^+ is $g^+(\mathbf{X}) = g(\mathbf{X}) \wedge P_C > 0$.

- The update of π^+ is $u^+(\mathbf{X}, \mathbf{X}', \mu) = u(\mathbf{X}, \mathbf{X}', \mu) \wedge P'_C = P_C - 1 + \#(C, A) \wedge \bigwedge_{C_1 \neq C} P'_{C_1} = P_{C_1} + \#(C_1, A)$, where $\#(C, A)$ is the number of occurrences of component C in the parallel composition A.
- The rate of π^+ is $\lambda^+(\mathbf{X}) = P_C \cdot \lambda(\mathbf{X})$.
- If π is an instantaneous action, then $\infty : p(\mathbf{x})$ becomes $\infty : P_C \cdot p(\mathbf{x})$ in π^+.
- The initial value of variables $\mathbf{x_0}^+$ equals $\mathbf{x_0}$ for variables in $\mathbf{X}$ and $\#(C, A)$ for each variable P_C.
- the initial network of $flat(\mathcal{A})$ is $A^+ = \|_{C^+ \in \mathsf{Def}^+} C^+$.

The notion of flattening has been previously defined in [21], but was called the extended version of a **sCCP** program. In [18] we also showed that a **sCCP** program and its flattened version have isomorphic labelled transitions systems, hence they are stochastically equivalent.

Example B.1. We illustrate this transformation with a simple example. Consider a simple model of a population of bacteria, in which each bacterium can consume a source of food and reproduce, or die. Both actions happen after some exponentially delayed time. We can model this in **sCCP** by using a single integer variable F, representing the available food, initially set to f_0, and by describing each bacterium as an agent as follows:

$$\texttt{bacterium} \stackrel{\text{def}}{=} [F > 0 \to F' = F - 1]_{k_r} . (\texttt{bacterium} \parallel \texttt{bacterium}) + [* \to *]_{k_d} . \mathbf{0}$$

The initial network $\texttt{bacterium} \parallel \ldots \parallel \texttt{bacterium}$ consists of m copies of the agent. This model can be flattened by introducing a new variable, B, counting the number of bacteria in the system, initially set to m, and replacing the agent `bacterium` by

$$\begin{array}{rcl} \texttt{bacterium\_flat} & \stackrel{\text{def}}{=} & [F > 0 \wedge B > 0 \to F' = F - 1 \wedge B = B + 1]_{k_r \cdot B} . \texttt{bacterium\_flat} \\ & + & [B > 0 \to B' = B - 1]_{k_d \cdot B} . \texttt{bacterium\_flat} \end{array}$$

The new initial network will contain only the agent `bacterium_flat`. Notice how the rates are updated to take into account the shift of perspective from the single agent to the population view.

C Background Material

In this section we briefly recall some notions that are needed in the proofs.

Hybrid state space. Let Q be a countable subset of $\mathbb{R}^k$, and consider $Q \times \mathbb{R}^n$, the hybrid space. A point $\mathbf{x} \in Q \times \mathbb{R}^n$ is a pair $\mathbf{x} = (q, \mathbf{y})$, $\mathbf{y} \in \mathbb{R}^n$.

In $Q \times \mathbb{R}^n$, we define a metric $\bar{d}$ for which $Q \times \mathbb{R}^n$ is a complete and separable metric space. This metric is derived from the euclidean metric d in $\mathbb{R}^n$ by

$$\bar{d}(\mathbf{x}_1, \mathbf{x}_2) = \begin{cases} d(\mathbf{y}_1, \mathbf{y}_2)/(1 + d(\mathbf{y}_1, \mathbf{y}_2)) & if\ \mathbf{x}_i = (q_i, \mathbf{y}_i)\ and\ q_1 = q_2, \\ 1 & if\ \mathbf{x}_i = (q_i, \mathbf{y}_i)\ and\ q_1 \neq q_2 \end{cases}$$

In particular, $\bar{d}(\mathbf{x}_1, \mathbf{x}_2) < 1$ if and only if $\mathbf{x}_1$ and $\mathbf{x}_2$ have the same Q-coordinate. Hence, a sequence converges with respect to $\bar{d}$, $\mathbf{x}_N \to \mathbf{x}$, if and only if $\mathbf{x} = (q, y)$, $\mathbf{x}_N = (q, \mathbf{y}_N)$ for $N \geq N_0$, and $\mathbf{y}_N \to \mathbf{y}$ in $\mathbb{R}^n$.

Each subset A of $Q \times \mathbb{R}^n$ is of the form $A = \bigcup_{q \in Q} \{q\} \times A_q$, $A_q \subset \mathbb{R}^n$. A sub-base for the topology of $Q \times \mathbb{R}^n$ is given by the open balls of the form $\{q\} \times B_\varepsilon(\mathbf{y})$. The boundary of a set A is denoted by ∂A and the closure by $\overline{A}$. Borel sets $\mathcal{B}$ for $Q \times \mathbb{R}^n$ are defined from the Borel sets $\mathcal{B}_q$ of $\mathbb{R}^n$ as $\mathcal{B} = \bigcup_{q \in Q} \{q\} \times \mathcal{B}_q$. See [31] for further details.

Skorohod metric. Continuous Time Markov Chains and Piecewise Deterministic Markov Processes considered in this paper can be seen as random variables on the space of cadlag functions $D([0,\infty),E)$ with values in $E \subseteq Q \times \mathbb{R}^n$. A cadlag function $f : [0,\infty) \to E$ is right continuous and has left limits for any $t \in [0,\infty)$.

The space $D([0,\infty),E)$ is given the structure of a metric space by the Skorohod metric. The Skorohod metric is first defined on compact time intervals $[0,T]$ and then extended over the whole positive time axis $[0,\infty)$.

Consider the uniform metric on the space $D([0,T],E)$, i.e. $\sup_{0\leq t\leq T} \|\mathbf{x}^{(N)}(t) - \mathbf{x}(t)\|$. If we have a sequence $\mathbf{x}^{(N)}$ of cadlag functions, then they will converge to $\mathbf{x}$ in the uniform norm if and only if the discontinuous jumps of $\mathbf{x}^{(N)}$ happen precisely at the same times as those of $\mathbf{x}$ (for $N \geq N_0$). The idea behind the Skorohod metric is to allow a small difference in these jump times by re-synchronizing them. That is to say, if the uniform metric allows one to wiggle space a bit, the Skorohod metric allows us also to wiggle time. To formalize this statement, let $\omega(t) : [0,T] \to [0,T]$ be a time-wiggle function, i.e a strictly increasing continuous function. Call $\mathcal{I}_T$ the set of such functions. Then, the Skorohod distance between $\mathbf{x},\mathbf{y} \in D([0,T],E)$ is

$$d_T(\mathbf{x},\mathbf{y}) = \inf_{\omega\in\mathcal{I}_T} \max\{ \sup_{t\in[0,T]} \|\omega(t) - t\|, \sup_{t\in[0,T]} \|\mathbf{x}(t) - \mathbf{y}(\omega(t))\|\}.$$

The metric d_T is extended to a metric on $D([0,\infty),E)$ by discounting large times as follows:

$$d(\mathbf{x},\mathbf{y}) = \sum_{K\in\mathbb{N}} 2^{-K} \min\{1, d_K(\mathbf{x},\mathbf{y})\}.$$

The Skorohod metric defines a topology for which $D([0,\infty),E)$ is complete and separable, i.e. it is a Polish space. See [51, 9] for a detailed introduction to the metric and its properties.

We note here that a sequence of functions $\mathbf{x}^{(N)} \in D([0,\infty),E)$ converges to $\mathbf{x} \in D([0,\infty),E)$ if and only if for each $T > 0$ there is a sequence of time-wiggle functions $\omega^{(N)} \in \mathcal{I}_T$ satisfying $\sup_{t\in[0,T]} \|\omega^{(N)}(t) - t\| \to 0$ and $\sup_{t\in[0,T]} \|\mathbf{x}^{(N)}(t) - \mathbf{x}(\omega^{(N)}(t))\| \to 0$.

Weak Convergence. The notion of weak convergence of a sequence of random variables $\mathbf{X}^{(N)}$ on a Polish space E to a random variable $\mathbf{X}$ is essentially the convergence of the induced probability measures on E. Weak convergence of probability measures is defined as convergence in the weak* topology [52]. More specifically, denote by $\mathcal{C}_b(E)$ the set of bounded continuous functions $f : E \to \mathbb{R}$ (note that we can have $E = D([0,\infty),E_0)$; in this case f is usually called a functional), and let $P, P^{(N)}$ be probability measures on E. We refer the reader to [51, 9] for an introduction to the subject.

Definition C.1. $P^{(N)}$ *converges weakly* to P, $P^{(N)} \Rightarrow P$, if and only if, for each $f \in \mathcal{C}(E)$,

$$\lim_{N\to\infty} \int_E f(\mathbf{x}) P^{(N)}(\mathbf{x}) d\mathbf{x} = \int_E f(\mathbf{x}) P(\mathbf{x}) d\mathbf{x}.$$

In case we have random variables $\mathbf{X}, \mathbf{X}^{(N)}$, then the previous condition can be written as

$$\lim_{N\to\infty} \mathbb{E}[f(\mathbf{X}^{(N)})] = \mathbb{E}[f(\mathbf{X})].$$

In this case, we write $\mathbf{X}^{(N)} \Rightarrow \mathbf{X}$.

The Portmanteau theorem provides a set of equivalent conditions for weak convergence of $\mathbf{X}^{(N)}$ to $\mathbf{X}$:

1. $\mathbf{X}^{(N)} \Rightarrow \mathbf{X}$;
2. $\lim_{N\to\infty} \mathbb{E}[f(\mathbf{X}^{(N)})] = \mathbb{E}[f(\mathbf{X})]$ for all bounded, uniformly continuous functions $f : E \to \mathbb{R}$;
3. $\limsup_{N\to\infty} \mathbb{P}\{\mathbf{X}^{(N)} \in F\} \leq \mathbb{P}\{\mathbf{X} \in F\}$ for all closed sets F;
4. $\liminf_{N\to\infty} \mathbb{P}\{\mathbf{X}^{(N)} \in G\} \geq \mathbb{P}\{\mathbf{X} \in G\}$ for all open sets G;

5. $\lim_{N\to\infty} \mathbb{P}\{\mathbf{X}^{(N)} \in A\} = \mathbb{P}\{\mathbf{X} \in A\}$ for all $\mathbf{X}$-continuity sets A (i.e., such that $\mathbb{P}\{\mathbf{X} \in \partial A\} = 0$).

Recall that there are other modes of convergence of random variables, among which *almost sure convergence* and *convergence in probability*. These two notions, differently from weak convergence, require to have fixed the sample space in which random variables are defined. More precisely, let $\mathbf{X}, \mathbf{X}^{(N)}$ be random variables on E, defined on the sample space Ω, with σ-algebra $\mathcal{A}$ and probability measure $\mathbb{P}$. (i.e. $\mathbf{X} : \Omega \to E$ is a $\mathcal{A}, \mathcal{B}$ measurable function). Then $\mathbf{X}^{(N)}$ converges to $\mathbf{X}$ almost surely if and only if $\mathbb{P}\{\lim_{N\to\infty} \mathbf{X}^{(N)}(\omega) = \mathbf{X}(\omega)\} = 1$, while it converges in probability if and only if, for each $\delta > 0$, $\lim_{N\to\infty} \mathbb{P}\{\|\mathbf{X}^{(N)} - \mathbf{X}\| > \delta\} = 0$.

These three notions are linked in several ways. Almost sure convergence implies convergence in probability, which in turn implies weak convergence. Furthermore, the Skorohod representation theorem states that, if $\mathbf{X}^{(N)} \Rightarrow \mathbf{X}$, then there is a sample space $(\Omega, \mathcal{A}, \mathbb{P})$, and realizations $\tilde{\mathbf{X}}, \tilde{\mathbf{X}}^{(N)}$ of $\mathbf{X}, \mathbf{X}^{(N)}$ on Ω (i.e. $\tilde{\mathbf{X}} : \Omega \to E$ induces the same probability on E as $\mathbf{X}$) such that $\tilde{\mathbf{X}}^{(N)}$ converges to $\tilde{\mathbf{X}}$ almost surely. Furthermore Ω can be taken as the unit interval with the Lebesgue measure. In particular, for real-valued random variables $X^{(N)}, X$ on $E \subseteq \mathbb{R}$, the Skorohod representation can be constructed using the quantile function $F^{\leftarrow}$, i.e. the pseudo-inverse of the cumulative distribution function $F(t) = \mathbb{P}\{X \leq t\}$: $\tilde{X^{(N)}} = (F^{(N)})^{\leftarrow}(U)$ and $\tilde{X} = F^{\leftarrow}(U)$, with U uniform in $[0, 1]$ [51].

Finally, weak convergence to a deterministic limit, i.e. a random variable concentrating all the probability mass to a point of E, implies convergence in probability.

In the following, we also need the notion of *tight* probability measure.

Definition C.2. A probability measure P on E is *tight* if and only if, for each $\varepsilon > 0$, there is a compact set $K_\varepsilon \subset E$ such that $P(K_\varepsilon) > 1 - \varepsilon$.

A sequence $P^{(N)}$ of probability measure s on E is *uniformly tight* if and only if for each $\varepsilon > 0$, there is a compact set $K_\varepsilon \subset E$ such that $P^{(N)}(K_\varepsilon) > 1 - \varepsilon$ for each $N \geq 0$.

If the space E is Polish, i.e. a complete and separable metric space, then each probability measure on E is tight. Furthermore, if $P^{(N)} \Rightarrow P$, then $P^{(N)}$ is uniformly tight.

Tightness is the right notion to characterize weak convergence in the space $D([0, \infty), E)$ equipped with the Skorohod topology. Let $\pi_{t_1,\dots,t_k} : D([0, \infty), E) \to E^k$ be the projection at fixed times $t_1, \dots, t_k \in [0, \infty)$. If $\mathbf{X}$ is a random variable in $D([0, \infty), E)$, then $\pi_{t_1,\dots,t_k}(\mathbf{X})$ is called a finite dimensional distribution. Now, if $\mathbf{X}, \mathbf{X}^{(N)}$ are random variables in $D([0, \infty), E)$, $\mathbf{X}^{(N)}$ is uniformly tight and $\pi_{t_1,\dots,t_k}(\mathbf{X}^{(N)}) \Rightarrow \pi_{t_1,\dots,t_k}(\mathbf{X})$, for t_j taken from a subset $\Gamma \subseteq [0, \infty)$ whose complement is at most countable (convergence of finite dimensional distributions), then $\mathbf{X}^{(N)} \Rightarrow \mathbf{X}$. Uniform tightness of $\mathbf{X}^{(N)}$ can be checked using some criteria based on the modulus of continuity, see [51, 9] for further details.

Finally, we need the *continuous mapping theorem*: Let $\mathbf{X}^{(N)}, \mathbf{X}$ be random variables on E and let $h : E \to E_1$. If h is $\mathbf{X}$-almost surely continuous (i.e., $\mathbb{P}\{\mathbf{X} \in C_h\} = 1$, where $C_h \subseteq E$ is the set of continuity points of h), and $\mathbf{X}^{(N)} \Rightarrow \mathbf{X}$, then $h(\mathbf{X}^{(N)}) \Rightarrow h(\mathbf{X})$.

Markov Kernels. A Markov kernel or Markov transition kernel on E, with σ-algebra $\mathcal{B}$, is a function $R : E \times \mathcal{B} \to [0, 1]$ such that

1. $R(\cdot, A)$ is a measurable function of for each $A \in \mathcal{B}$.
2. $R(\mathbf{y}, \cdot)$ is a probability measure for each $\mathbf{y} \in E$.

We now prove a Lemma that will be used in many proofs in next section, that allows us to propagate weak convergence by Markov kernels, under a suitable notion of continuity of the kernel.

Lemma C.1. *Let $R^{(N)}(\hat{\mathbf{y}}) = R^{(N)}(\hat{\mathbf{y}}, \cdot)$ and $R(\hat{\mathbf{y}}) = R(\hat{\mathbf{y}}, \cdot)$ be Markov transition kernels on some Polish space E such that $R^{(N)}(\hat{\mathbf{y}}^{(N)}) \Rightarrow R(\hat{\mathbf{y}})$, whenever $\hat{\mathbf{y}}^{(N)} \to \hat{\mathbf{y}}$. If $\hat{\mathbf{Y}}^{(N)} \Rightarrow \hat{\mathbf{Y}}$, where $\hat{\mathbf{Y}}^{(N)}$, $\hat{\mathbf{Y}}$ are random elements in E, then $R^{(N)}(\hat{\mathbf{Y}}^{(N)}) \Rightarrow R(\hat{\mathbf{Y}})$ and $(\hat{\mathbf{Y}}^{(N)}, R^{(N)}(\hat{\mathbf{Y}}^{(N)})) \Rightarrow (\hat{\mathbf{Y}}, R(\hat{\mathbf{Y}}))$.*

Proof. The proof is essentially the same as that of the core argument of Theorem 1 in [40]. We reproduce it here just for completeness. Fix a bounded and uniformly continuous function $g : E \to \mathbb{R}$. We need to

show that $|\mathbb{E}[g(R^{(N)}(\hat{\mathbf{Y}}^{(N)}))] - \mathbb{E}[g(R(\hat{\mathbf{Y}}))]| \to 0$. For simplicity, call $R^{(N)}g(\hat{\mathbf{y}}) = \int_E g(\hat{\mathbf{x}})R^{(N)}(\hat{\mathbf{y}}, \hat{\mathbf{x}})\ \mathrm{d}\hat{\mathbf{x}}$ and similarly $Rg(\hat{\mathbf{y}})$, and further let $P^{(N)}(\hat{\mathbf{y}}) = \mathbb{P}\{\hat{\mathbf{Y}}^{(N)} = \hat{\mathbf{y}}\}$, and similarly for $P(\hat{\mathbf{y}})$. Then

$$\begin{aligned}
|\mathbb{E}[g(R^{(N)}(\hat{\mathbf{Y}}^{(N)}))] - \mathbb{E}[g(R(\hat{\mathbf{Y}}))]| &= \left| \int_E R^{(N)}g(\hat{\mathbf{y}})P^{(N)}(\hat{\mathbf{y}})\ \mathrm{d}\hat{\mathbf{y}} - \int_E Rg(\hat{\mathbf{y}})P(\hat{\mathbf{y}})\ \mathrm{d}\hat{\mathbf{y}} \right| \\
&\leq \underbrace{\int_E |R^{(N)}g(\hat{\mathbf{y}}) - Rg(\hat{\mathbf{y}})|P^{(N)}(\hat{\mathbf{y}})\ \mathrm{d}\hat{\mathbf{y}}}_{(a)} \\
&+ \underbrace{\left| \int_E Rg(\hat{\mathbf{y}})P^{(N)}(\hat{\mathbf{y}})\ \mathrm{d}\hat{\mathbf{y}} - \int_E Rg(\hat{\mathbf{y}})P(\hat{\mathbf{y}})\ \mathrm{d}\hat{\mathbf{y}} \right|}_{(b)}.
\end{aligned}$$

The term (b) in the previous inequality goes to zero as the hypothesis imply that $Rg(\hat{\mathbf{y}})$ is a continuous function (see [40]), and so we just need to focus on (a). As E is a Polish space, i.e. a separable complete metric space, it follows from $P^{(N)} \Rightarrow P$ and P tight that $P^{(N)}$ is uniformly tight. Then, we find a compact set E_ε such that $P^{(N)}(E_\varepsilon) > 1 - \varepsilon/2\|g\|_\infty$, and so

$$\begin{aligned}
\int_E |R^{(N)}g(\hat{\mathbf{y}}) - Rg(\hat{\mathbf{y}})|P^{(N)}(\hat{\mathbf{y}})\ \mathrm{d}\hat{\mathbf{y}} &\leq \int_{E_\varepsilon} |R^{(N)}g(\hat{\mathbf{y}}) - Rg(\hat{\mathbf{y}})|P^{(N)}(\hat{\mathbf{y}})\ \mathrm{d}\hat{\mathbf{y}} + \varepsilon \\
&\leq \sup_{\hat{\mathbf{y}} \in E_\varepsilon} |R^{(N)}g(\hat{\mathbf{y}}) - Rg(\hat{\mathbf{y}})| + \varepsilon \to \varepsilon,
\end{aligned}$$

and therefore

$$\limsup_{N\to\infty} \int_E |R^{(N)}g(\hat{\mathbf{y}}) - Rg(\hat{\mathbf{y}})|P^{(N)}(\hat{\mathbf{y}})\ \mathrm{d}\hat{\mathbf{y}} \leq \varepsilon,$$

as the hypothesis of the lemma imply that $R^{(N)}g$ converges to Rg uniformly on compact sets [40]. As the previous inequality holds for each $\varepsilon > 0$, then $R^{(N)}(\hat{\mathbf{Y}}^{(N)}) \Rightarrow R(\hat{\mathbf{Y}})$. To prove the second part of the theorem, observe that in the previous computation we can always restrict the expectation to a closed set $F_1 \subset E$, showing that $\mathbb{E}[g(R^{(N)}(\hat{\mathbf{Y}}^{(N)}))\mathbf{1}_{F_1}(\hat{\mathbf{Y}}^{(N)})] \to \mathbb{E}[g(R(\hat{\mathbf{Y}}))\mathbf{1}_{F_1}(\hat{\mathbf{Y}})]$. Now, if we fix another closed set $F_2 \subset E$, and choose bounded uniformly continuous functions $g_\rho \downarrow \mathbf{1}_{F_2}$, approximating from above the indicator function of F_2, we then have

$$\mathbb{P}\{(\hat{\mathbf{Y}}^{(N)}, R^{(N)}(\hat{\mathbf{Y}}^{(N)})) \in F_1 \times F_2\} \leq \mathbb{E}[g_\rho(R^{(N)}(\hat{\mathbf{Y}}^{(N)}))\mathbf{1}_{F_1}(\hat{\mathbf{Y}}^{(N)})],$$

from which, fixing ρ,

$$\limsup_{N\to\infty} \mathbb{P}\{(\hat{\mathbf{Y}}^{(N)}, R^{(N)}(\hat{\mathbf{Y}}^{(N)})) \in F_1 \times F_2\} \leq \mathbb{E}[g_\rho(R(\hat{\mathbf{Y}}))\mathbf{1}_{F_1}(\hat{\mathbf{Y}})],$$

by letting $\rho \to 0$ and invoking the bounded convergence theorem, as g_ρ converges to $\mathbf{1}_{F_2}$, we have

$$\limsup_{N\to\infty} \mathbb{P}\{(\hat{\mathbf{Y}}^{(N)}, R^{(N)}(\hat{\mathbf{Y}}^{(N)})) \in F_1 \times F_2\} \leq \mathbb{P}\{(\hat{\mathbf{Y}}, R(\hat{\mathbf{Y}})) \in F_1 \times F_2\},$$

which by the Portmanteau theorem [9], implies that $(\hat{\mathbf{Y}}^{(N)}, R^{(N)}(\hat{\mathbf{Y}}^{(N)})) \Rightarrow (\hat{\mathbf{Y}}, R(\hat{\mathbf{Y}}))$. □

Functional Analysis. In the following, we will also need the famous Gronwall inequality, which we recall here for convenience.

Proposition C.1. *For any real valued integrable function f on the interval $[0, T]$, if*

$$f(t) \leq C + D\int_0^t f(s)\ \mathrm{d}s,$$

then

$$f(T) \leq Ce^{DT}.$$

D Proof of Main Lemmas and Theorems

We will start by providing a quick proof of Theorem 4.1, the classic fluid theorem. This will be helpful in proving subsequent lemmas and theorems.

Theorem (4.1)**.** *Let $(\mathcal{A}, \gamma_N)$ be a sequence of population-**sCCP** models for increasing system size $\gamma_N \to \infty$, satisfying the conditions of this section, and with all **sCCP**-actions π satisfying the continuous scaling condition. Let $\hat{\mathbf{X}}^{(N)}(t)$ be the sequence of normalized CTMC associated with the **sCCP**-program and $\hat{\mathbf{x}}(t)$ be the solution of the fluid ODE.*
If $\hat{\mathbf{x}}_0^{(N)} \to \hat{\mathbf{x}}_0$ almost surely, then for any $T < \infty$, $\sup_{t \le T} \|\hat{\mathbf{X}}^{(N)}(t) - \hat{\mathbf{x}}(t)\| \to 0$ as $N \to \infty$, almost surely.

Proof. Intuitively, the result of the theorem is a limit result, hence it depends only on what happens in a neighbourhood $B_\varepsilon(\hat{\mathbf{x}}([0,T])$ of the solution of the ODE. Thus, by restricting our attention to a compact set $K \subset E$ containing $B_\varepsilon(\hat{\mathbf{x}}([0,T]) \cap E$ for some ε, we can assume that all functions g_η defining the rate of normalized transitions are bounded, say by B_η, and Lipschitz, say with Lipschitz constant L_η. This assumption is not restrictive, as we can always extend the functions g_η on the whole E so that they are globally bounded and Lipschitz continuous. Clearly, $\hat{\mathbf{x}}(t)$ will remain unchanged by this operation, as it depends only on the value of g_η in K.

We consider now the representation of the CTMC $\hat{\mathbf{X}}^{(N)}(t)$ in terms of Poisson processes [42]. We will use one Poisson process for each transition η and each possible value of ν_η (which is a random element with bounded first and second moments). In the following, we indicate with $p_\eta(\mathbf{w})$ the probability that $\nu_\eta = \mathbf{w}$.

$$\hat{\mathbf{X}}^{(N)}(t) = \hat{\mathbf{X}}_0^{(N)} + \sum_{\eta \in \mathcal{T}} \sum_{\mathbf{w} \in \mathbb{Z}^n} \frac{\mathbf{w}}{N} N_\eta \left(N p_\eta(\mathbf{w}) \int_0^t g_\eta^{(N)}(\hat{\mathbf{X}}^{(N)}(s))\, \mathrm{d}s \right) \tag{7}$$

Furthermore, $\hat{\mathbf{x}}(t)$ in integral form is:

$$\hat{\mathbf{x}}(t) = \hat{\mathbf{x}}_0 + \int_0^t \sum_{\eta \in \mathcal{T}} \mathbb{E}[\nu_\eta] g_\eta^{(N)}(\hat{\mathbf{x}}(s))\, \mathrm{d}s \tag{8}$$

In the following, we need the notion of centred Poisson process [42], defined by $\hat{N}(\lambda t) = N(\lambda t) - \lambda t$, for which the following law of large numbers holds: $\sup_{t \le T} \frac{1}{N} \hat{N}(N\lambda t) \to 0$ almost surely.

Now, we define

$$\begin{aligned} \varepsilon^{(N)}(t) &= \hat{\mathbf{X}}^{(N)}(t) - \hat{\mathbf{X}}^{(N)}(0) - \int_0^t \sum_{\eta \in \mathcal{T}} \mathbb{E}[\nu_\eta] g_\eta^{(N)}(\hat{\mathbf{x}}(s))\, \mathrm{d}s \\ &= \sum_{\eta \in \mathcal{T}} \sum_{\mathbf{w} \in \mathbb{Z}^n} \frac{\|\mathbf{w}\|}{N} \hat{N}_\eta \left(N p_\eta(\mathbf{w}) \int_0^t g_\eta^{(N)}(\hat{\mathbf{X}}^{(N)}(s))\, \mathrm{d}s \right), \end{aligned}$$

so that

$$\|\varepsilon^{(N)}(t)\| \le \sum_{\eta \in \mathcal{T}} \sum_{\mathbf{w} \in \mathbb{Z}^n} \frac{\mathbf{w}}{N} \hat{N}_\eta \left(N p_\eta(\mathbf{w}) B_\eta t \right).$$

By the finiteness of second order moments for ν_η, the previous equation is summable and we can further exchange limit and summation over $\mathbf{w}$ [42], to conclude that $\sup_{t \le T} \|\varepsilon^{(N)}(t)\| \to 0$ by the law of large numbers for centred Poisson processes. Therefore, we have that

$$\begin{aligned} \sup_{t \le T} \|\hat{\mathbf{X}}^{(N)}(t) - \hat{\mathbf{x}}(t)\| &\le \underbrace{\|\hat{\mathbf{X}}_0^{(N)} - \hat{\mathbf{x}}_0\| + \sup_{t \le T} \|\varepsilon^{(N)}(t)\| + \sup_{t \le T} \|F^{(N)}(\hat{\mathbf{X}}^{(N)}(t)) - F(\hat{\mathbf{X}}^{(N)}(t))\|}_{= \delta^{(N)}(T) \to 0 \ a.s.} \\ &+ \int_0^t \|F(\hat{\mathbf{X}}^{(N)}(t)) - F(\hat{\mathbf{x}}(t))\|\, \mathrm{d}s. \end{aligned}$$

Calling $\beta^{(N)}(T) = \sup_{t \leq T} \|\hat{\mathbf{X}}^{(N)}(t) - \hat{\mathbf{x}}(t)\|$ and applying Lipschitz condition to the last term, we have that

$$\beta^{(N)}(T) \leq \delta^{(N)}(T) + L \int_0^T \beta^{(N)}(t) \, \mathrm{d}t.$$

By applying Gronwall's inequality (see Proposition C.1), we finally obtain

$$\beta^{(N)}(T) \leq \delta^{(N)}(T) e^{LT} \to 0 \text{ almost surely.}$$

□

We will turn now to prove Theorem 5.1. We will need some auxiliary lemmas. The first one is a straightforward generalization of the Kurtz theorem, to the case in which the initial condition of the limit process is sampled from a distribution on the space E.

Lemma D.1. *Let $\hat{\mathbf{X}}(t)$ and $\hat{\mathbf{x}}(t)$ be as in Theorem 4.1. Furthermore, assume that $\|\hat{\mathbf{X}}_0 - \hat{\mathbf{x}}_0\|$ converges to zero almost surely. Then, for any $T < \infty$, $\sup_{t \leq T} \|\hat{\mathbf{X}}^{(N)}(t) - \hat{\mathbf{x}}(t)\| \to 0$ as $N \to \infty$ almost surely.*

Proof. To begin with, suppose $\hat{\mathbf{x}}_0$ is supported on a compact set K_0. Then the proof proceeds as in Theorem 4.1, with the only caveat that we need to consider a compact set K containing an ε-neighbourhood of all trajectories starting in K_0 up to time T (which is a compact set, by continuity of the ODE flow). In fact, the argument of Theorem 4.1 does not require that the initial condition of ODE is deterministic, but just the convergence in probability of $\hat{\mathbf{X}}_0$ to $\hat{\mathbf{x}}_0$.

Now, as $\hat{\mathbf{x}}_0$ is *tight*, for each $\varepsilon > 0$ there is a compact set K_ε such that $\mathbb{P}\{\hat{\mathbf{x}}_0 \notin K_\varepsilon\} < \varepsilon$. Conditional on $\hat{\mathbf{x}}_0 \in K_\varepsilon$, the convergence of $\hat{\mathbf{X}}^{(N)}(t)$ to $\hat{\mathbf{x}}(t)$ is then almost sure. Hence, fix a sequence $\varepsilon_k \downarrow 0$, such that the corresponding $K_{\varepsilon_k} \uparrow E$, and with $\sum_k \varepsilon_k < \infty$. By discarding a set of measure 0, we can assume that convergence in each K_{ε_k} is sure. Then, any other point u of the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ on which processes are defined that makes convergence fail has to belong to the complement of K_{ε_k} infinitely often. By the Borel-Cantelli lemma [8], the set of all such u has probability zero. □

We will now turn to consider convergence of stochastic jump times, focussing attention on a single jump time, given convergent initial conditions of the stochastic and the piecewise-deterministic system. As we consider the first stochastic jump time, we can focus our attention on deterministic systems (with random initial conditions). In order to do this, we combine simple properties of the space of cadlag functions with the Skorohod representation theorem (see Appendix C).

Proposition D.1. *Let $\mathbf{x}^{(N)}$, $\mathbf{x}$ be elements of $D([0,\infty), E)$, such that $\mathbf{x}^{(N)} \to \mathbf{x}$ (with respect to the Skorohod metrics). Then, for each $T > 0$, $\int_0^T \mathbf{x}^{(N)}(s) \, \mathrm{d}s \to \int_0^T \mathbf{x}(s) \, \mathrm{d}s$.*

Proof. By the definition of the Skorohod metrics, let $\omega^{(N)}(t)$ be a sequence of time-wiggle functions such that $\sup_{t \leq T} \|\omega^{(N)}(t) - t\| \to 0$, and $\sup_{t \leq T} \|\mathbf{x}^{(N)}(t) - \mathbf{x}(\omega^{(N)}(t))\| \to 0$. Now,

$$\int_0^T \|\mathbf{x}^{(N)}(s) - \mathbf{x}(s)\| \, \mathrm{d}s \leq \underbrace{\int_0^T \|\mathbf{x}^{(N)}(s) - \mathbf{x}(\omega^{(N)}(s))\| \, \mathrm{d}s}_{(a)} + \underbrace{\int_0^T \|\mathbf{x}(\omega^{(N)}(s)) - \mathbf{x}(s)\| \, \mathrm{d}s}_{(b)}.$$

The term (a) goes to zero by the uniform convergence of $\mathbf{x}^{(N)}(t)$ to $\mathbf{x}(\omega^{(N)}(t))$, while for (b), observe that the function $g^{(N)}(t) = \|\mathbf{x}(\omega^{(N)}(t)) - \mathbf{x}(t)\|$ goes to zero in every continuity point of $\mathbf{x}$, thus almost everywhere. Furthermore, in $[0,T]$ the function $\mathbf{x}$ is bounded by a compactness argument (see [9]), and so is $g^{(N)}$, so that we can apply the bounded convergence theorem [8] to conclude that (b) converges to zero. □

Jump times. In order to show convergence of jump times of discrete stochastic transitions, we need the notion of *cumulative rate* for the PDMP $\Lambda(t)$ and for the CTMCs.

Consider discrete stochastic actions $\pi \in \mathsf{disc}(\mathcal{A})$, and the rate of firing of discrete actions in the PDMP, $\hat{\lambda}(\hat{\mathbf{x}}) = \sum_{\pi \in \mathsf{disc}(\mathcal{A})} \hat{\lambda}_\pi(\hat{\mathbf{x}})$, and in the CTMC at level N, $\hat{\lambda}^{(N)}(\hat{\mathbf{x}}) = \sum_{\pi \in \mathsf{disc}(\mathcal{A})} \hat{\lambda}_\pi^{(N)}(\hat{\mathbf{x}})$. Then, the *cumulative rate* for the PDMP $\hat{\mathbf{x}}$ is

$$\Lambda(t) = \int_0^t \hat{\lambda}(\hat{\mathbf{x}}(s)) \, \mathrm{d}s, \tag{9}$$

while for the CTMC $\hat{\mathbf{X}}^{(N)}(t)$ is

$$\Lambda^{(N)}(t) = \int_0^t \hat{\lambda}^{(N)}(\hat{\mathbf{X}}^{(N)}(s))\ \mathrm{d}s. \tag{10}$$

These are the cumulative rates of non-homogeneous Poisson processes. We are interested in the first firing time, whose cumulative distribution function is given by $1 - e^{-\Lambda(t)}$. By a standard inversion method [8, 64], the first firing time for the PDMP is given by $\tau = \inf\{t \geq 0 \mid 1 - e^{-\Lambda(t)} \geq U\} = \inf\{t \geq 0 \mid \Lambda(t) \geq \xi\}$, where U is a uniform random variable and ξ is an exponentially distributed random variable with rate 1 (it holds that $\xi = -\log U$). We assume $\inf \emptyset = \infty$. Similarly, we can define $\tau^{(N)} = \inf\{t \geq 0 \mid \Lambda^{(N)}(t) \geq \xi\}$, the first firing time of a discrete transition $\pi \in \mathsf{disc}(\mathcal{A})$ in $\hat{\mathbf{X}}^{(N)}(t)$. By the Skorohod representation theorem for unidimensional random variables (see Section C or, for instance, [51]), if the pointwise convergence of $\Lambda^{(N)}(t)$ to $\Lambda(t)$ holds, then $\tau^{(N)} \to \tau$ almost surely. We can combine these facts with Proposition D.1, to prove the following lemma.

Lemma D.2. *Let $\hat{\mathbf{X}}^{(N)}(t) \Rightarrow \hat{\mathbf{x}}(t)$, and $\tau^{(N)}$, τ be defined as above. If $\hat{\lambda}$ is continuous and $\hat{\lambda}^{(N)} \to \hat{\lambda}$ uniformly in each compact set $K \subseteq E$, then $\tau^{(N)} \Rightarrow \tau$, as $N \to \infty$.*

Proof. The first step of the proof is to use the Skorohod representation theorem to construct realizations $\tilde{\mathbf{X}}^{(N)}$ of $\hat{\mathbf{X}}^{(N)}$ and $\tilde{\mathbf{X}}$ of $\hat{\mathbf{X}}$ on some probability space $\mathcal{P}$ such that $\tilde{\mathbf{X}}^{(N)} \to \tilde{\mathbf{X}}$ almost surely, as random elements in the space of cadlag functions.

It then follows that $\hat{\lambda}^{(N)}(\tilde{\mathbf{X}}^{(N)}) \to \hat{\lambda}(\tilde{\mathbf{X}})$ almost surely. In fact, consider the time-wiggle functions $\omega^{(N)}$ (depending also on the sample space $(\Omega, \mathcal{A}, \mathbb{P})$, i.e. $\omega^{(N)} = \omega^{(N)}(u,t)$, for $u \in \Omega$) such that $\omega^{(N)} \to id$ and $\tilde{\mathbf{Y}}^{(N)} = \tilde{\mathbf{X}}^{(N)} \circ \omega^{(N)} \to \tilde{\mathbf{X}}$ uniformly in $[0,T]$. Then

$$\begin{aligned}\sup_{t \leq T} \|\hat{\lambda}^{(N)}(\tilde{\mathbf{X}}^{(N)}(\omega^{(N)}(t))) - \hat{\lambda}(\tilde{\mathbf{X}}(t))\| \leq & \underbrace{\sup_{t \leq T} \|\hat{\lambda}^{(N)}(\tilde{\mathbf{Y}}^{(N)}(t)) - \hat{\lambda}(\tilde{\mathbf{Y}}^{(N)}(t))\|}_{(a)} \\ & + \underbrace{\sup_{t \leq T} \|\hat{\lambda}(\tilde{\mathbf{Y}}^{(N)}(t)) - \hat{\lambda}(\tilde{\mathbf{X}}(t))\|}_{(b)}.\end{aligned}$$

Term (a) goes to zero by uniform convergence of $\hat{\lambda}^{(N)}$ to $\hat{\lambda}$ (as in $[0,T]$ $\tilde{\mathbf{X}}^{(N)}$ and $\tilde{\mathbf{X}}$ are contained in a compact set), while term (b) goes to zero due to uniform convergence of $\tilde{\mathbf{Y}}^{(N)}$ to $\tilde{\mathbf{X}}$ in $[0,T]$ and uniform continuity of $\hat{\lambda}$ in $[0,T]$.

Now, we can apply Proposition D.1 to $\hat{\lambda}^{(N)}(\tilde{\mathbf{X}}^{(N)}) \to \hat{\lambda}(\tilde{\mathbf{X}})$, to conclude that $\Lambda^{(N)}(T) \to \Lambda(T)$ almost surely for each $T > 0$. Combining this with the Skorohod representation theorem for real random variables,[18] we get $\tilde{\tau}^{(N)} \to \tilde{\tau}$ almost surely, where $\tilde{\tau}^{(N)}$ and $\tilde{\tau}^{(N)}$ are the jump times obtained from the realizations of the original processes. It then follows that $\tau^{(N)} \Rightarrow \tau$. □

We finally consider the convergence of states at times $\tau^{(N)}$ and τ.

Proposition D.2. *Let $\hat{\mathbf{X}}^{(N)} \Rightarrow \hat{\mathbf{x}}$ and let $\tau^{(N)}, \tau$ be stopping times satisfy conditions of the previous lemma. Then, conditional on $\tau < \infty$, $\hat{\mathbf{X}}^{(N)}(\tau^{(N)}) \Rightarrow \hat{\mathbf{x}}(\tau)$.*

Proof. In fact, let $t^{(N)} \to t < \infty$ (here we use implicitly the fact that $\tau < \infty$). Then, use Skorohod representation theorem and take representations $\tilde{\mathbf{X}}^{(N)}$ and $\tilde{\mathbf{X}}$ of $\hat{\mathbf{X}}^{(N)}$ and $\hat{\mathbf{X}}$ such that $\tilde{\mathbf{X}}^{(N)} \to \tilde{\mathbf{X}}$ almost surely. By continuity of $\tilde{\mathbf{X}}$ and uniform convergence of $\tilde{\mathbf{X}}^{(N)}$ and $\tilde{\mathbf{X}}$ in $[0,T]$ (the Skorohod metrics and the uniform metrics on compact sets are the same when the limit function is continuous), it follows that $\tilde{\mathbf{X}}^{(N)}(t^{(N)}) \to \tilde{\mathbf{X}}(t)$ almost surely, hence $\hat{\mathbf{X}}^{(N)}(t^{(N)}) \Rightarrow \hat{\mathbf{x}}(t)$. Hence, by seeing $\hat{\mathbf{X}}^{(N)}$ and $\hat{\mathbf{X}}$ as Markov kernels, we can apply Lemma C.1 to conclude. □

[18] We are effectively coupling $\hat{\mathbf{X}}^{(N)}$, $\hat{\mathbf{x}}$, $\tau^{(N)}$ and τ on the probability space $\Omega \times [0,1]$. Note in particular that we allow τ and $\tau^{(N)}$ to take the value ∞. This can happen with non-null probability if and only if $\Lambda(T)$ does not diverge as $T \to \infty$.

In order to prove Theorem 5.1, we will use an inductive argument whose core is the following corollary, which combines the previous results in the light of weak convergence.

Corollary D.1. *Let $(\mathcal{A}, \gamma_N)$ be a sequence of **sCCP** models for increasing systems size, satisfying the conditions of this section, and with all actions π satisfying the continuous scaling condition. Let $\hat{\mathbf{X}}^{(N)}(t)$ be the associated sequence of normalized CTMC and $\hat{\mathbf{x}}(t)$ be the solution of the fluid ODE.*

If $\hat{\mathbf{X}}_0^{(N)} \Rightarrow \hat{\mathbf{x}}_0$, then

1. *$\hat{\mathbf{X}}^{(N)} \Rightarrow \hat{\mathbf{x}}$, as random elements in the space of cadlag function on E, with the Skorohod metrics.*
2. *If $\tau^{(N)}$, τ, are the jump times of a stochastic event with rate $\lambda^{(N)}$ and λ, respectively, then $\tau^{(N)} \Rightarrow \tau$;*
3. *$\hat{\mathbf{X}}^{(N)}(\tau^{(N)}) \Rightarrow \hat{\mathbf{x}}(\tau)$;*
4. *If $R^{(N)}(\hat{\mathbf{y}})$ and $R(\hat{\mathbf{y}})$ are reset kernels satisfying $R^{(N)}(\hat{\mathbf{y}}^{(N)}) \Rightarrow R(\hat{\mathbf{y}})$ whenever $\hat{\mathbf{y}}^{(N)} \to \hat{\mathbf{y}}$, then $R^{(N)}(\hat{\mathbf{X}}^{(N)}(\tau^{(N)})) \Rightarrow R(\hat{\mathbf{x}}(\tau))$;*
5. *Under the previous conditions, $(\hat{\mathbf{X}}_0^{(N)}, \hat{\mathbf{X}}^{(N)}, \tau^{(N)}, \hat{\mathbf{X}}^{(N)}(\tau^{(N)}), R^{(N)}(\hat{\mathbf{X}}^{(N)}(\tau^{(N)}))) \Rightarrow (\hat{\mathbf{x}}_0, \hat{\mathbf{x}}, \tau, \hat{\mathbf{x}}(\tau), R(\hat{\mathbf{x}}(\tau)))$*

Proof. The proof works simply by constructing an a.s. convergent realization of the initial conditions. Then, by Lemma D.1, we obtain point 1. Point 2 follows from Lemma D.2 and point 1, while point 3 from Proposition D.2 and point 2. Point 4 follows from Lemma C.1 and point 3. Point 5, instead, follows again from Lemma C.1, observing that each element on the vector is defined conditionally on the previous one (e.g. $\hat{\mathbf{X}}^{(N)}$ depends conditionally on $\hat{\mathbf{X}}_0^{(N)}$, $\tau^{(N)}$ depends conditionally on $\hat{\mathbf{X}}^{(N)}$, and so on), and this dependence satisfies the assumptions of the Lemma. For instance, if $\hat{\mathbf{y}}^{(N)} \to \hat{\mathbf{y}}$, then $(\hat{\mathbf{X}}^{(N)} | \hat{\mathbf{X}}_0^{(N)} = \hat{\mathbf{y}}^{(N)}) \Rightarrow (\hat{\mathbf{x}} | \hat{\mathbf{x}}_0 = \hat{\mathbf{y}})$, and the dependency is measurable (in fact, continuous) on $\hat{\mathbf{y}}^{(N)}$, $\hat{\mathbf{y}}$. Similar observations hold for the other elements of the vector. Then an iterated application of Lemma C.1 is enough to conclude. □

We can now prove Theorem 5.1.

Theorem (5.1)**.** *Let $(\mathcal{A}, \gamma_N)$ be a sequence of population-**sCCP** models for increasing system size $\gamma_N \to \infty$, satisfying the conditions of this section, with variables partitioned into discrete $\mathbf{X}_d$, continuous $\mathbf{X}_c$, and environment ones $\mathbf{X}_e$. Assume that discrete actions satisfy scaling 3 and continuous actions satisfy scaling 2. Let $\hat{\mathbf{X}}^{(N)}(t)$ be the sequence of normalized CTMC associated with the **sCCP** program and $\hat{\mathbf{x}}(t)$ be the PDMP associated with the limit normalized TDSHA $\hat{\mathcal{T}}(\mathcal{A})$.*

If $\hat{\mathbf{x}}_0^{(N)} \Rightarrow \hat{\mathbf{x}}_0$ (weakly) and the PDMP is non-Zeno, then $\hat{\mathbf{X}}(t)$ converges weakly to $\hat{\mathbf{x}}(t)$, $\hat{\mathbf{X}} \Rightarrow \hat{\mathbf{x}}$, as random elements in the space of cadlag function with the Skorohod metric.

Proof. The basic idea of the proof to show weak convergence is to apply inductively the previous corollary, to show weak convergence of $\hat{\mathbf{X}}_m^{(N)}(t) = \hat{\mathbf{X}}^{(N)}(t \wedge \tau_{m+1}^{(N)})$ to $\hat{\mathbf{x}}_m(t) = \hat{\mathbf{x}}(t \wedge \tau_{m+1})$, i.e. to processes stopped after $m+1$ jumps, and then lift this to the full weak convergence.

Step 1: weak convergence conditional on m jumps or less. Consider the sequence $\tau_1, \tau_2, \ldots$ of jump times of discrete stochastic transitions in the PDMP $\hat{\mathbf{x}}(t)$ and the sequence $\tau_1^{(N)}, \tau_2^{(N)}, \ldots$ of jump times of discrete transitions $\pi \in \mathsf{disc}(\mathcal{A})$ in $\hat{\mathbf{X}}^{(N)}(t)$. In the following, we need to take care also of the fact that τ_m may be infinite with probability greater than zero. Note that, conditional on τ_m being infinite, all τ_{m+j} will be infinite, too.

In order to be more concise, let us introduce some additional local notation. First, denote $\hat{\mathbf{Z}}_m^{(N)} = \hat{\mathbf{X}}^{(N)}(\tau_m^{(N)+})$ and $\hat{\mathbf{z}}_m = \hat{\mathbf{x}}(\tau_m^+)$ the states of the CTMC at level N and of the PDMP after the m-th discrete jump. If τ_m or $\tau_m^{(N)}$ are infinite, we assume $\hat{\mathbf{z}}_m$ or $\hat{\mathbf{Z}}_m^{(N)}$ be equal to a special value $(q_\Delta, \mathbf{0})$, where q_Δ is a special state of Q where nothing happens: the vector field and the jump rate are null (i.e. it is a cemetery point). Note that $(q_\Delta, \mathbf{0})$ has distance 1 from any point $(q, \mathbf{x})$ in E.
Let also $\hat{\mathbf{Z}}_0^{(N)} = \hat{\mathbf{X}}_0^{(N)}$, $\hat{\mathbf{z}}_0 = \hat{\mathbf{x}}_0$, $\tau_0^{(N)} = \tau_0 = 0$. We define $\hat{\mathbf{Y}}_m^{(N)}(t)$ to be the CTMC starting from $\hat{\mathbf{Z}}_m^{(N)}$ with no discrete jumps, and $\hat{\mathbf{y}}_m(t)$ the PDMP starting in $\hat{\mathbf{z}}_m$ with no discrete jumps (in fact, an ODE with random initial conditions). Notice that, if $\tau_m^{(N)}$ (resp. τ_m) is finite, then $\hat{\mathbf{Y}}_m^{(N)}(t)$ (resp. $\hat{\mathbf{y}}_m(t)$) coincides

with $\hat{\mathbf{X}}^{(N)}(\tau_m^{(N)}+t)$ (resp. $\hat{\mathbf{x}}(\tau_m+t)$) for $\tau_m^{(N)} \le t < \tau_{m+1}^{(N)}$ (resp. $\tau_m \le t < \tau_{m+1}$), by the strong Markov property of CTMC [49] and of PDMP [31].

We will now prove that, for each $m > 0$, conditional on $\tau_m < \infty$, $\hat{\mathbf{Y}}_m^{(N)} \Rightarrow \hat{\mathbf{y}}_m$ and $\tau_{m+1}^{(N)} \Rightarrow \tau_{m+1}$. Moreover, if $\tau_{m+1} < \infty$, then also $\hat{\mathbf{Z}}_{m+1} \Rightarrow \hat{\mathbf{z}}_{m+1}$. Finally, we will also show that, conditional on $\tau_{m+1} < \infty$, $(\hat{\mathbf{Z}}_0^{(N)}, \hat{\mathbf{Y}}_0^{(N)}, \tau_1^{(N)}, \hat{\mathbf{Z}}_1^{(N)}, \dots, \hat{\mathbf{Y}}_m^{(N)}, \tau_{m+1}^{(N)}, \hat{\mathbf{Z}}_{m+1}^{(N)}) \Rightarrow (\hat{\mathbf{z}}_0, \hat{\mathbf{y}}_0, \tau_1, \hat{\mathbf{z}}_1, \dots, \hat{\mathbf{y}}_m, \tau_{m+1}, \hat{\mathbf{z}}_{m+1})$. The argument is a simple induction. In particular, the induction hypothesis is that $(\hat{\mathbf{Z}}_0^{(N)}, \hat{\mathbf{Y}}_0^{(N)}, \tau_1^{(N)}, \hat{\mathbf{Z}}_1^{(N)}, \dots, \hat{\mathbf{Z}}_m^{(N)}) \Rightarrow (\hat{\mathbf{z}}_0, \hat{\mathbf{y}}_0, \tau_1, \hat{\mathbf{z}}_1, \dots, \hat{\mathbf{z}}_m)$ conditional on $\tau_m < \infty$.[19] From this, $\hat{\mathbf{Y}}_m^{(N)} \Rightarrow \hat{\mathbf{y}}_m$ is immediate from Lemma D.1, and $\tau_{m+1}^{(N)} \Rightarrow \tau_{m+1}$ follows from Lemma D.2. Now, conditional on $\tau_{m+1} < \infty$, we can apply Proposition D.2 to conclude that $\hat{\mathbf{Z}}_{m+1} \Rightarrow \hat{\mathbf{z}}_{m+1}$. As $\tau_{m+1} < \infty$ implies $\tau_m < \infty$, reasoning as in Corollary D.1 (using the same argument there to extend inductively its length), we obtain the weak convergence of vectors of the random elements.

Consider now $\hat{\mathbf{X}}_m^{(N)}(t) = \hat{\mathbf{X}}^{(N)}(t \wedge \tau_{m+1}^{(N)})$ and $\hat{\mathbf{x}}_m(t) = \hat{\mathbf{x}}(t \wedge \tau_{m+1})$. We can write $\hat{\mathbf{X}}_m^{(N)}(t) = \sum_{i=0}^m \hat{\mathbf{Y}}_i^{(N)}(t - \tau_i^{(N)})\mathbf{I}\{\tau_i^{(N)} \le t < \tau_{i+1}^{(N)}\} + \mathbf{I}\{\tau_{m+1}^{(N)} \le t\}\hat{\mathbf{Z}}_{m+1}^{(N)}$ and $\hat{\mathbf{x}}_m(t) = \sum_{i=0}^m \hat{\mathbf{y}}_i(t-\tau_i)\mathbf{I}\{\tau_i \le t < \tau_{i+1}\} + \mathbf{I}\{\tau_{m+1} \le t\}\hat{\mathbf{z}}_{m+1}$. Now, the functional that associates with T the cadlag element $\mathbf{I}\{t \le T\}$ is continuous with respect to Skorohod metrics, and if we consider the previous definitions of $\hat{\mathbf{X}}_m^{(N)}$ and $\hat{\mathbf{x}}_m$ as a function of $(\hat{\mathbf{Z}}_0^{(N)}, \hat{\mathbf{Y}}_0^{(N)}, \tau_1^{(N)}, \hat{\mathbf{Z}}_1^{(N)}, \dots, \hat{\mathbf{Z}}_{m+1}^{(N)})$ and $(\hat{\mathbf{z}}_0, \hat{\mathbf{y}}_0, \tau_1, \hat{\mathbf{z}}_1, \dots, \hat{\mathbf{z}}_{m+1})$, respectively, then this function is continuous. Hence, conditional on $\tau_{m+1} < \infty$, $\hat{\mathbf{X}}_m^{(N)} \Rightarrow \hat{\mathbf{x}}_m$ by the continuous mapping theorem. The same property holds also when $\tau_{m+1} = \infty$. In fact, conditioning on $\tau_j < \infty$ and $\tau_{j+1} = \infty$, for $j \le m$, we observe that the process $\mathbf{w}_j(t) = \sum_{i=0}^{j-1} \hat{\mathbf{y}}_i(t-\tau_i)\mathbf{I}\{\tau_i \le t < \tau_{i+1}\} + \mathbf{I}\{t \ge \tau_j\}\hat{\mathbf{y}}_j(t-\tau_j)$ coincides with $\hat{\mathbf{x}}_m(t)$, and by the same argument above, applied to vectors $(\hat{\mathbf{Z}}_0^{(N)}, \hat{\mathbf{Y}}_0^{(N)}, \tau_1^{(N)}, \hat{\mathbf{Z}}_1^{(N)}, \dots, \tau_{j+1}^{(N)})$ and $(\hat{\mathbf{z}}_0, \hat{\mathbf{y}}_0, \tau_1, \hat{\mathbf{z}}_1, \dots, \tau_{j+1})$, the processes $\mathbf{W}_j^{(N)}(t) = \sum_{i=0}^{j-1} \hat{\mathbf{Y}}_i^{(N)}(t - \tau_i^{(N)})\mathbf{I}\{\tau_i^{(N)} \le t < \tau_{i+1}^{(N)}\} + \mathbf{I}\{t \ge \tau_j^{(N)}\}\hat{\mathbf{Y}}_j^{(N)}(t - \tau_j^{(N)})$ converge weakly to $\mathbf{w}_j$. Now, the processes $\mathbf{W}_j^{(N)}(t)$ and $\hat{\mathbf{X}}_m^{(N)}(t)$ are the same up to time $\tau_{j+1}^{(N)}$, which is a divergent sequence under the event $\{\tau_j < \infty, \tau_{j+1} = \infty\}$[20]. This implies that $\hat{\mathbf{X}}_m^{(N)}$ converges weakly to $\mathbf{W}_j^{(N)}$ (in fact, their Skorokhod distance converges weakly to zero, in fact a.s. under any a.s. realisation of $\tau_{j+1}^{(N)} \to \infty$), and hence, by uniqueness of the limit, to $\mathbf{W}_j = \hat{\mathbf{x}}_m$. Now, as the events $\{\tau_j < \infty, \tau_{j+1} = \infty\}$, for $j = 0, \dots, m$ and $\{\tau_{m+1} < \infty\}$ are disjoint and their union has probability one, we can remove the conditioning and conclude $\hat{\mathbf{X}}_m^{(N)} \Rightarrow \hat{\mathbf{x}}_m$.[21]

Step 2: Weak convergence. We now lift the weak convergence $\hat{\mathbf{X}}_m^{(N)} \Rightarrow \hat{\mathbf{x}}_m$ to weak convergence of the full processes $\hat{\mathbf{X}}^{(N)} \Rightarrow \hat{\mathbf{x}}$. Consider a bounded uniformly continuous function f from the space $D = D([0,\infty), E)$ of cadlag functions with values in E to $\mathbb{R}$. By the definition of the Skorohod metric d in D, for each $\rho > 0$, there is a $T > 0$ such that $d(x,y) < d_T(x,y) + \rho$, where d_T is the metric restricted to the compact time interval $[0,T]$ (see Section C). By the uniform continuity of f, given $\varepsilon > 0$, we fix a $\rho > 0$ such that $|f(x) - f(y)| < \varepsilon/4$ whenever $d(x,y) < \rho$.

Now, fix $T > 0$ according to the previous condition on ρ, and choose m such that $\mathbb{P}\{\tau_{m+1} > T\} > 1 -$

[19] In particular, this implies that $\hat{\mathbf{z}}_m \neq (q_\Delta, \mathbf{0})$, and as $\hat{\mathbf{Z}}_m^{(N)} \Rightarrow \hat{\mathbf{z}}_m$, ultimately also $\hat{\mathbf{Z}}_m^{(N)} \neq (q_\Delta, \mathbf{0})$.

[20] In fact, $\tau_{j+1}^{(N)} \Rightarrow \tau_{j+1}$ conditional on $\tau_j < \infty$. Moreover, if $\tau_{j+1} = \infty$, it also holds that $\tau_{j+k}^{(N)} \Rightarrow \tau_{j+k}$ for any $k > 0$, as $\tau_{j+k} = \infty$ and $\tau_{j+k}^{(N)} \ge \tau_{j+1} \to \infty$. This means that by induction we can conclude $\tau_j^{(N)} \Rightarrow \tau_j$ for any j.

[21] To see this more precisely, couple all processes $\mathbf{W}_j^{(N)}$, $\mathbf{w}_j$, $\hat{\mathbf{x}}_m$, and $\hat{\mathbf{X}}_m^{(N)}$, and the exit times τ_j on a common space Ω, and let Ω_j the subset corresponding to the event $\{\tau_j = \infty, \tau_{j-1} < \infty\}$, for $j = 1, \dots, m$, and Ω_0 be the subset corresponding to the event $\{\tau_m < \infty\}$. Clearly Ω_j, for $j = 0, \dots, m$, form a partition of Ω. Let $\mathbb{P}$ be the probability measure on Ω, let μ_j the push-forward measure on the space of cadlag functions on E of $\mathbf{w}_j$, i.e. of $\hat{\mathbf{x}}_m$ conditioned on Ω_j, and let $\mu_j^{(N)}$ be the push-forward measure of $\hat{\mathbf{X}}_m^{(N)}$ conditioned on Ω_j, for j such that $p_j = \mathbb{P}(\Omega_j) > 0$ (call J such a set of indices). We know $\mu_j^{(N)} \Rightarrow \mu_j$. Then μ, the push-forward measure of $\hat{\mathbf{x}}_m$, coincides with $\sum_{j\in J} p_j \mu_j$, and similarly $\mu^{(N)}$, the push-forward measure of $\hat{\mathbf{X}}_m^{(N)}$, is $\sum_{j\in J} p_j \mu_j^{(N)}$. Now, let F be a closed set of the cadlag space $D([0,\infty), E$. Then

$$\limsup_N \mu_N(F) \le \sum_{j\in J} p_j \limsup_N \mu_j^{(N)}(F) \le \sum_{j\in J} p_j \mu_j(F) = \mu(F),$$

which implies $\mu^{(N)} \Rightarrow \mu$, and hence $\hat{\mathbf{X}}_m^{(N)} \Rightarrow \hat{\mathbf{x}}_m$, by the Portmanteau theorem.

$\varepsilon/(16\|f\|)$, which can be found since the expected number of discrete transitions fired by the PDMP at time T is finite. As $\tau_{m+1}^{(N)} \Rightarrow \tau_{m+1}$, we can also find an N_0 such that, for all $N \geq N_0$, $\mathbb{P}\{\tau_{m+1}^{(N)} > T\} > 1 - \varepsilon/(8\|f\|)$ (using the liminf condition in the Portmanteau theorem).

Now, conditioning on $\tau_{m+1} > T$, we have that $\hat{\mathbf{x}}(t \wedge T) = \hat{\mathbf{x}}_m(t \wedge T)$, and so $d(\hat{\mathbf{x}}, \hat{\mathbf{x}}_m) \leq \rho$. Similarly, conditioning on $\tau_{m+1}^{(N)} > T$, we have $\hat{\mathbf{X}}^{(N)}(t \wedge T) = \hat{\mathbf{X}}_m^{(N)}(t \wedge T)$ and $d(\hat{\mathbf{X}}^{(N)}, \hat{\mathbf{X}}_m^{(N)}) \leq \rho$. Now

$$\begin{aligned}\left|\mathbb{E}[f(\hat{\mathbf{X}}^{(N)})] - \mathbb{E}[f(\hat{\mathbf{x}})]\right| &\leq \underbrace{\mathbb{E}[|f(\hat{\mathbf{X}}^{(N)}) - f(\hat{\mathbf{X}}_m^{(N)})|]}_{(a)} + \underbrace{\mathbb{E}[|f(\hat{\mathbf{x}}) - f(\hat{\mathbf{x}}_m)|]}_{(b)} \\ &+ \underbrace{\left|\mathbb{E}[f(\hat{\mathbf{X}}_m^{(N)})] - \mathbb{E}[f(\hat{\mathbf{x}}_m)]\right|}_{(c)}\end{aligned}$$

Now, term (c) goes to zero as $\hat{\mathbf{X}}_m^{(N)} \Rightarrow \hat{\mathbf{x}}_m$. To bound (b), instead, using properties of conditional expectation, we have

$$\begin{aligned}\mathbb{E}[|f(\hat{\mathbf{x}}) - f(\hat{\mathbf{x}}_m)|] &= \mathbb{E}[\mathbb{E}[|f(\hat{\mathbf{x}}) - f(\hat{\mathbf{x}}_m)| \;\big|\; \mathbf{I}\{\tau_{m+1} > T\}]] \\ &\leq \mathbb{E}[|f(\hat{\mathbf{x}}) - f(\hat{\mathbf{x}}_m)| \;\big|\; \mathbf{I}\{\tau_{m+1} > T\} = 1] \cdot \mathbb{P}\{\tau_{m+1} > T\} + 2\|f\|\mathbb{P}\{\tau_{m+1} \leq T\} \\ &\leq \mathbb{E}[|f(\hat{\mathbf{x}}) - f(\hat{\mathbf{x}}_m)| \;\big|\; \mathbf{I}\{\tau_{m+1} > T\} = 1] + \varepsilon/4 \leq \varepsilon/2,\end{aligned}$$

where the last inequality follows from the choice of ρ and the fact that $d(\hat{\mathbf{x}}, \hat{\mathbf{x}}_m) \leq \rho$. A similar argument can be used for term (a), allowing us to conclude that

$$\limsup_{N\to\infty} \mathbb{E}[|f(\hat{\mathbf{X}}^{(N)}) - f(\hat{\mathbf{X}}_m^{(N)})|] \leq \varepsilon/2,$$

from which we have

$$\limsup_{N\to\infty} \left|\mathbb{E}[f(\hat{\mathbf{X}}^{(N)})] - \mathbb{E}[f(\hat{\mathbf{x}})]\right| \leq \varepsilon.$$

By the arbitrariness of $\varepsilon > 0$, we can finally conclude that $\left|\mathbb{E}[f(\hat{\mathbf{X}}^{(N)})] - \mathbb{E}[f(\hat{\mathbf{x}})]\right| \to 0$. □

D.1 Instantaneous transitions

In this subsection, we give the proofs of results contained in Section 6 of the paper.

Lemma (6.1)**.** *Let $(\mathcal{A}, \gamma_N)$ be a sequence of population-**sCCP** models for increasing population size. Let $\hat{\mathbf{X}}^{(N)}(t)$ be the associated sequence of normalized CTMC, and suppose $\hat{\mathbf{X}}^{(N)} \Rightarrow \hat{\mathbf{X}}$, where $\hat{\mathbf{X}}$ has a.s. continuous sample paths. Let $h^{(N)}$, h be activation functions for $\hat{\mathbf{X}}^{(N)}$ and $\hat{\mathbf{X}}$, such that $h^{(N)} \to h$ uniformly, and suppose h is transversal to $\hat{\mathbf{X}}$. Then $\zeta^{(N)} \Rightarrow \zeta$.*

Proof. First of all, use the Skorohod representation theorem to construct representations $\tilde{\mathbf{X}}^{(N)}$ of $\hat{\mathbf{X}}^{(N)}$ and $\tilde{\mathbf{X}}$ of $\hat{\mathbf{X}}$ such that $\tilde{\mathbf{X}}^{(N)} \to \tilde{\mathbf{X}}$ almost surely. Now fix sample trajectories $\tilde{\mathbf{x}}^{(N)} \to \tilde{\mathbf{x}}$ in the Skorohod metrics. As $\tilde{\mathbf{X}}$ is almost surely continuous, we can assume $\tilde{\mathbf{x}}$ continuous. In this case, the Skorohod metric is the same as the uniform metric on compact sets $[0, T]$. In particular, we can take T larger than $\tilde{\zeta} + \delta$, for any $\delta > 0$, where $\tilde{\zeta}$ is the exit time for $\tilde{\mathbf{x}}$. Now, since h is transversal for $\hat{\mathbf{X}}$, there is a $\delta > 0$ such that $h(\tilde{\mathbf{x}}(t)) > 0$ for $t \in (\tilde{\zeta}, \tilde{\zeta} + \delta]$. Let $\bar{h} = \min\{\max\{-h(\tilde{\mathbf{x}}(t)) \mid t \in [\tilde{\zeta} - \delta, \tilde{\zeta}]\}, \max\{h(\tilde{\mathbf{x}}(t)) \mid t \in [\tilde{\zeta}, \tilde{\zeta} + \delta]\}\}$. Fix $\varepsilon > 0$, $\varepsilon < \bar{h}$, and let $\tilde{\zeta}_\varepsilon^- = \sup\{t \leq \tilde{\zeta} \mid h(\tilde{\mathbf{x}}(t)) \leq -\varepsilon\}$ and $\tilde{\zeta}_\varepsilon^+ = \inf\{t \geq \tilde{\zeta} \mid h(\tilde{\mathbf{x}}(t)) \geq \varepsilon\}$. By continuity of $\tilde{\mathbf{x}}$, it follows that $\|h(\tilde{\mathbf{x}}(t)) - h(\tilde{\mathbf{x}}(\tilde{\zeta}))\| < \varepsilon$ for any $t \in (\tilde{\zeta}_\varepsilon^-, \tilde{\zeta}_\varepsilon^+)$, and that $\tilde{\zeta}_\varepsilon^-, \tilde{\zeta}_\varepsilon^+ \to \tilde{\zeta}$ as $\varepsilon \to 0$.

Now, choose a compact set K in E that contains the ε-neighbourhood of $\tilde{\mathbf{x}}$ in $[0, T]$, for $T > \tilde{\zeta}_\varepsilon^+$. As h is uniformly continuous in K, pick a $\rho > 0$ such that $\|h(\hat{\mathbf{x}}_1) - h(\hat{\mathbf{x}}_2)\| < \varepsilon/4$ whenever $\|\hat{\mathbf{x}}_1 - \hat{\mathbf{x}}_2\| < \rho$, and fix $N_0 > 0$ such that $\tilde{\mathbf{x}}^{(N)}(t)$ is ρ-close to $\tilde{\mathbf{x}}(t)$ for $N \geq N_0$, uniformly in $[0, T]$. Furthermore, find N_1 such that, for $N \geq N_1$, $\sup_{\hat{\mathbf{x}} \in K} \|h^{(N)}(\hat{\mathbf{x}}) - h(\hat{\mathbf{x}})\| < \varepsilon/4$. Let $\bar{N} = \max\{N_0, N_1\}$. It follows that, for $N \geq \bar{N}$, $\|h^{(N)}(\tilde{\mathbf{x}}^{(N)}(t)) - h(\tilde{\mathbf{x}}(t))\| \leq \|h^{(N)}(\tilde{\mathbf{x}}^{(N)}(t)) - h(\tilde{\mathbf{x}}^{(N)}(t))\| + \|h(\tilde{\mathbf{x}}^{(N)}(t)) - h(\tilde{\mathbf{x}}(t))\| < \varepsilon/2$, and so $h^{(N)}(\tilde{\mathbf{x}}^{(N)}(t)) < 0$ for $t \in [0, \tilde{\zeta}_\varepsilon^-]$, hence $\tilde{\zeta}^{(N)} > \tilde{\zeta}_\varepsilon^-$. Furthermore, $h^{(N)}(\tilde{\mathbf{x}}^{(N)}(\tilde{\zeta}_\varepsilon^+)) > 0$, and so $\tilde{\zeta}^{(N)} < \tilde{\zeta}_\varepsilon^+$. It follows $\tilde{\zeta}^{(N)} \to \tilde{\zeta}$ a.s., and therefore $\zeta^{(N)} \Rightarrow \zeta$. □

We now turn the attention to resets of instantaneous guards, proving a version of Lemma C.1 dealing with the discontinuities in the reset kernels under some regularity assumptions. We first recall some notation. Let $\hat{\mathbf{p}}_i^{(N)}$, $\hat{\mathbf{p}}_i$ be the weight functions, with $\hat{\mathbf{p}}_i$ continuous and $\hat{\mathbf{p}}_i^{(N)}$ uniformly convergent to $\hat{\mathbf{p}}_i$ on each compact set $K \subseteq E$. Furthermore, let $\hat{\mathbf{p}}(\hat{\mathbf{x}}) = \sum_i \hat{\mathbf{p}}_i(\hat{\mathbf{x}})$, and similarly $\hat{\mathbf{p}}^{(N)}(\hat{\mathbf{x}}) = \sum_i \hat{\mathbf{p}}_i^{(N)}(\hat{\mathbf{x}})$. Let $R_i^{(N)}$ and R_i be the reset kernels associated with the instantaneous transitions satisfying $R_i^{(N)}(\hat{\mathbf{x}}^{(N)}) \Rightarrow R_i(\hat{\mathbf{x}})$ whenever $\hat{\mathbf{x}}^{(N)} \to \hat{\mathbf{x}}$. Finally, let $h_i^{(N)}$, h_i be the activation functions of guards, with h_i continuous and $h_i^{(N)}$ converging uniformly to h_i. The further properties that are required for the activation functions are the following:

- Each h_i is a robust activation function, according to Definition 6.1.
- The PDMP is robustly transversal, see Definition 6.3.
- The set of activation functions h_i enjoys the size-compatibility property, see Definition 6.6.
- The PDMP has the robust activation property, as stated in Definition 6.5.

Consider now the activation function for the PDMP, defined by $h(\hat{\mathbf{x}}) = \max\{h_1(\hat{\mathbf{x}}), \dots, h_m(\hat{\mathbf{x}})\}$, and let $\mathcal{H} = \{\hat{\mathbf{x}} \mid h(\hat{\mathbf{x}}) = 0\}$ be the activation surface of instantaneous transitions. $\mathcal{H}^{(N)}$ is defined similarly. Furthermore, let $\mathcal{H}_i = \mathcal{H} \cap \{\hat{\mathbf{x}} \mid h_i(\hat{\mathbf{x}}) = 0\}$ be the portion of $\mathcal{H}$ in which transition i is active. Call $D_i = \partial_{\mathcal{H}} \mathcal{H}_i$ the boundary of $\mathcal{H}_i$ in $\mathcal{H}$ and $D = \bigcup_{i=1}^m D_i$. The robust activation property implies that the probability of jumping from D is zero. Furthermore, let I_{dep} be the index of size-dependent activation functions, i.e. such that $h^{(N)} \neq h$, and I_{ind} the index set of size-independent activation functions. The size-compatibility condition states that, for each $i \in I_{dep}$ and $\hat{\mathbf{x}} \in int_{\mathcal{H}}(\mathcal{H}_i)$, $h_j(\hat{\mathbf{x}}) \neq 0$ for $j \neq i$, i.e. only h_i is zero.

Finally, recall the definition of the reset kernels on $\mathcal{H}^{(N)}$ and $\mathcal{H}$:

$$R^{(N)}(\hat{\mathbf{x}}, \cdot) = \sum_{i=1}^m \mathbf{1}\{h_i^{(N)}(\hat{\mathbf{x}}) \geq 0\}(\hat{\mathbf{p}}_i^{(N)}(\hat{\mathbf{x}})/\hat{\mathbf{p}}^{(N)}(\hat{\mathbf{x}}))R_i^{(N)}(\hat{\mathbf{x}}, \cdot),$$

and

$$R(\hat{\mathbf{x}}, \cdot) = \sum_{i=1}^m \mathbf{1}\{h_i(\hat{\mathbf{x}}) \geq 0\}(\hat{\mathbf{p}}_i(\hat{\mathbf{x}})/\hat{\mathbf{p}}(\hat{\mathbf{x}}))R_i(\hat{\mathbf{x}}, \cdot).$$

Under the previous hypothesis, we can prove the following lemma.

Lemma D.3. *Let $R^{(N)}(\hat{\mathbf{y}})$ and $R(\hat{\mathbf{y}})$ defined as before and let $\hat{\mathbf{Y}}^{(N)} \Rightarrow \hat{\mathbf{Y}}$, where $\hat{\mathbf{Y}}^{(N)}$, $\hat{\mathbf{Y}}$ are random elements with support in $\mathcal{H}^{(N)}$ and $\mathcal{H}$, respectively, such that $\mathbb{P}\{\hat{\mathbf{Y}} \in D\} = 0$. Then $R^{(N)}(\hat{\mathbf{Y}}^{(N)}) \Rightarrow R(\hat{\mathbf{Y}})$ and $(\hat{\mathbf{Y}}^{(N)}, R^{(N)}(\hat{\mathbf{Y}}^{(N)})) \Rightarrow (\hat{\mathbf{Y}}, R(\hat{\mathbf{Y}}))$.*

Proof. Fix a bounded and uniformly continuous function $g : E \to \mathbb{R}$. We need to prove that $|\mathbb{E}[g(R^{(N)}(\hat{\mathbf{Y}}^{(N)}))] - \mathbb{E}[g(R(\hat{\mathbf{Y}}))]| \to 0$ as $N \to \infty$. We use the same notation as in Lemma C.1. The idea is to split the integrals $\int_E R^{(N)} g(\hat{\mathbf{y}}) P^{(N)}(\hat{\mathbf{y}})\, \mathrm{d}\hat{\mathbf{y}}$ and $\int_E Rg(\hat{\mathbf{y}})P(\hat{\mathbf{y}})\, \mathrm{d}\hat{\mathbf{y}}$ into several regions, surrounding the discontinuity points by a small probability region in such a way that the probability mass is concentrated on a continuity region, in which we can apply Lemma C.1. There are some technical details that we have to work out, as $\hat{\mathbf{Y}}^{(N)}$ and $\hat{\mathbf{Y}}$ are concentrated on a manifold of E.

Let $\delta > 0$ (to be fixed afterwards) and $K \subseteq E$ be a compact set. Call $D_\delta = \bigcup_{\hat{\mathbf{x}} \in D} B_\delta(\hat{\mathbf{x}})$ the δ-neighbourhood of D. Clearly, $D_\delta \downarrow D$ as $\delta \downarrow 0$, and therefore $P(D_\delta) \downarrow 0$. The same holds for the closure $\overline{D_\delta}$: $P(\overline{D_\delta}) \downarrow 0$

Consider now a size-dependent activation function, $i \in I_{dep}$, and let $\mathcal{H}_{i,\delta} = \mathcal{H}_i \cap D_\delta^c$. By the size-compatibility condition, it follows that $|h_j(\hat{\mathbf{x}})| > 0$ for each $\hat{\mathbf{x}} \in \mathcal{H}_{i,\delta}$ and each $j \neq i$. In particular, $d(\hat{\mathbf{x}}, \mathcal{H}_j) > 0$ for each $\hat{\mathbf{x}} \in \mathcal{H}_{i,\delta}$, where the distance between a point and a set is defined in the usual way as $d(\hat{\mathbf{x}}, A) = \inf_{\hat{\mathbf{y}} \in A} d(\hat{\mathbf{x}}, \hat{\mathbf{y}})$. Now, notice that $\mathcal{H}_{i,\delta}$ is closed and so $\mathcal{H}_{i\,\delta} \cap K$ is compact. Therefore, by

continuity of the distance d, there is a $\rho_{i,j} > 0$ such that $d(\hat{\mathbf{x}}, \mathcal{H}_j) > \rho_{i,j}$ for each $\hat{\mathbf{x}} \in \mathcal{H}_{i\,\delta} \cap K$.[22] Let now $\delta_i = \min_{j \neq i} \rho_{i,j}/2$, and notice that, for each $\hat{\mathbf{x}} \in \mathcal{H}_{i,\delta} \cap K$ and $\hat{\mathbf{y}} \in B_{\delta_i}(\hat{\mathbf{x}})$, we have $d(\hat{\mathbf{y}}, \mathcal{H}_j) > \delta_i > 0$. Let $A_{i,\delta} = \left[\bigcup_{\hat{\mathbf{x}} \in \mathcal{H}_{i,\delta} \cap K} B_{\delta_i}(\hat{\mathbf{x}})\right] \cap D_\delta^c$, then $d(\hat{\mathbf{y}}, \mathcal{H}_j) \geq \delta_i$ for each $\hat{\mathbf{y}} \in \overline{A_{i,\delta}}$. It follows that $h_j(\hat{\mathbf{y}}) > 0$ in $\overline{A_{i,\delta}}$, which is compact, so that we find a $\rho_i > 0$ such that $\|h_j(\hat{\mathbf{y}})\| \geq \rho_i$ for $\hat{\mathbf{y}} \in A_{i\,\delta}$ and each $j \neq i$.

By possibly invoking uniform convergence of $h_j^{(N)}$ to h_j, for $j \in I_{dep}$, the property of $A_{i,\delta}$ allows us to conclude that, for N large enough, $h_j^{(N)}(\hat{\mathbf{y}}) \neq 0$ in $A_{i,\delta}$. Furthermore, by the robust activation of h_i, $h_i^{(N)}$ ultimately changes sign within $A_{i,\delta}$. In particular, combining this with the fact that $\hat{\mathbf{Y}}^{(N)}$ is supported in $\mathcal{H}^{(N)}$ and $\hat{\mathbf{Y}}$ is supported in $\mathcal{H}$, we get that in $A_{i,\delta}$, $R^{(N)}$ coincides with $R_i^{(N)}$ and R with R_i, hence they satisfy the continuity property $R^{(N)}(\hat{\mathbf{x}}^{(N)}) \to R(\hat{\mathbf{x}})$ as $\hat{\mathbf{x}}^{(N)} \to \hat{\mathbf{x}}$.

We can deal similarly with size-independent activation functions h_j, $j \in I_{ind}$. In this case, however, we may have more than one guard robustly active in $\mathcal{H}$, so we really need to consider each possible activation profile. Let α be a boolean vector, $\alpha \in \{0,1\}^m$, such that $\alpha_i = 0$ for $i \in I_{dep}$. Call J_{ind} the set of such vectors. Then we can define $\mathcal{H}_\alpha = \mathcal{H} \cap \bigcap_{j:\alpha_j=1} \mathcal{H}_j$. In $int_{\mathcal{H}}(\mathcal{H}_\alpha)$, $h_i(\hat{\mathbf{x}}) \neq 0$ if and only if $\alpha_i = 0$, hence we can reason as for the size-dependent case to construct an open neighbourhood $A_{\alpha,\delta}$ of $\mathcal{H}_\alpha \cap D_\delta^c$ in which $h_i^{(N)}(\hat{\mathbf{x}}) \neq 0$ for N large enough and all $\hat{\mathbf{x}} \in A_{\alpha,\delta}$. Since $h_j^{(N)} = h_j$ for $\alpha_j = 1$, and since $P^{(N)}$ and P are supported in $\mathcal{H}_\alpha$, when restricted to $A_{\alpha,\delta}$, we can conclude that $R^{(N)}(\hat{\mathbf{x}}^{(N)}) \to R(\hat{\mathbf{x}})$ as $\hat{\mathbf{x}}^{(N)} \to \hat{\mathbf{x}}$ in $A_{\alpha,\delta}$.

Recall the definition of J_{ind}, and let J_{dep} be the set of boolean vectors $\alpha \in \{0,1\}^m$ equal to one only for a single $i \in I_{dep}$, and zero elsewhere, and $J = J_{ind} \cup J_{dep}$. For each compact K and δ, we have constructed an open set $A_\delta = \bigcup_{\alpha \in J} A_{\alpha,\delta}$ such that $R^{(N)}$ and R behave nicely in it.

Now, fix $\varepsilon > 0$, and, invoking the uniform tightness of $P^{(N)}$ and P, choose K_ε compact such that $P^{(N)}(K_\varepsilon) \geq 1 - \varepsilon/4\|g\|_\infty$. Furthermore, pick $\delta > 0$ such that $P(\overline{D_\delta}) \leq \varepsilon/4\|g\|_\infty$. Then we have

$$
\begin{aligned}
|\mathbb{E}[g(R^{(N)}(\hat{\mathbf{Y}}^{(N)}))] \quad - \quad \mathbb{E}[g(R(\hat{\mathbf{Y}}))]| \leq & \underbrace{\left| \int_{A_\delta} R^{(N)} g(\hat{\mathbf{y}}) P^{(N)}(\hat{\mathbf{y}}) \, \mathrm{d}\hat{\mathbf{y}} - \int_{A_\delta} R g(\hat{\mathbf{y}}) P(\hat{\mathbf{y}}) \, \mathrm{d}\hat{\mathbf{y}} \right|}_{(a)} \\
+ & \underbrace{\left| \int_{K_\varepsilon \setminus (A_\delta \cup D_\delta)} R^{(N)} g(\hat{\mathbf{y}}) P^{(N)}(\hat{\mathbf{y}}) \, \mathrm{d}\hat{\mathbf{y}} - \int_{K_\varepsilon \setminus (A_\delta \cup D_\delta)} R g(\hat{\mathbf{y}}) P(\hat{\mathbf{y}}) \, \mathrm{d}\hat{\mathbf{y}} \right|}_{(b)} \\
+ & \underbrace{\left| \int_{D_\delta} R^{(N)} g(\hat{\mathbf{y}}) P^{(N)}(\hat{\mathbf{y}}) \, \mathrm{d}\hat{\mathbf{y}} - \int_{D_\delta} R g(\hat{\mathbf{y}}) P(\hat{\mathbf{y}}) \, \mathrm{d}\hat{\mathbf{y}} \right|}_{(c)} \\
+ & \underbrace{\left| \int_{K_\varepsilon^c} R^{(N)} g(\hat{\mathbf{y}}) P^{(N)}(\hat{\mathbf{y}}) \, \mathrm{d}\hat{\mathbf{y}} - \int_{K_\varepsilon^c} R g(\hat{\mathbf{y}}) P(\hat{\mathbf{y}}) \, \mathrm{d}\hat{\mathbf{y}} \right|}_{(d)}
\end{aligned}
$$

Now, term (a) goes to zero invoking Lemma C.1, given the continuity of resets in A_δ. To deal with term (b), notice that $K_\varepsilon \setminus (A_\delta \cup D_\delta)$ is closed and $P(K_\varepsilon \setminus (A_\delta \cup D_\delta)) = 0$, so that $\limsup_N P^{(N)}(K_\varepsilon \setminus (A_\delta \cup D_\delta)) = 0$. Term (c) is dealt with by observing that $\limsup_N P^{(N)}(\overline{D_\delta}) \leq P(\overline{D_\delta}) \leq \varepsilon/4\|g\|_\infty$, and so $\limsup_N P^{(N)}(D_\delta) \leq \varepsilon/4\|g\|_\infty$. Therefore (c) is less than $\varepsilon/2$. Finally, (d) is less than $\varepsilon/2$ by the choice of K_ε. It follows that

$$
\limsup_{N \to \infty} |\mathbb{E}[g(R^{(N)}(\hat{\mathbf{Y}}^{(N)}))] - \mathbb{E}[g(R(\hat{\mathbf{Y}}))]| \leq \varepsilon,
$$

which implies $|\mathbb{E}[g(R^{(N)}(\hat{\mathbf{Y}}^{(N)}))] - \mathbb{E}[g(R(\hat{\mathbf{Y}}))]| \to 0$. Proof of the second statement of the theorem can be copied verbatim from Lemma C.1. □

We now give the proof of Theorem 6.1.

[22] Let $f : K \to \mathbb{R}$ be a continuous function on a compact set K such that $|f(\hat{\mathbf{x}})| > 0$ for each $\hat{\mathbf{x}} \in K$. Then there is $\varepsilon > 0$ such that $|f(\hat{\mathbf{x}})| \geq \varepsilon$ for each $\hat{\mathbf{x}} \in K$. Suppose not, and choosing $\varepsilon = 1/n$, construct a sequence $\hat{\mathbf{x}}_n$ such that $f(\hat{\mathbf{x}}_n) \leq 1/n$ and so $f(\hat{\mathbf{x}}_n) \to 0$. By compactness of K, extract a convergent subsequence $\hat{\mathbf{x}}_{n_k} \to \hat{\mathbf{y}} \in K$. Then $0 = \lim_k f(\hat{\mathbf{x}}_{n_k}) = f(\hat{\mathbf{y}})$, a contradiction.

Theorem (6.1)**.** *Let $(\mathcal{A}, \gamma_N)$ be a sequence of population-**sCCP** models for increasing system size $\gamma_N \to \infty$, as $N \to \infty$, with variables partitioned into $\mathbf{X} = (\mathbf{X}_d, \mathbf{X}_c, \mathbf{X}_e)$, with discrete stochastic actions satisfying scaling 3, instantaneous actions satisfying scaling 4, and continuous actions satisfying scaling 2. Let $\hat{\mathbf{X}}^{(N)}(t)$ be the associated sequence of normalized CTMC and $\hat{\mathbf{x}}(t)$ be the limit PDMP associated with the normalized limit TDSHA $\hat{\mathcal{T}}(\mathcal{A})$.*

If $\hat{\mathbf{x}}_0^{(N)} \Rightarrow \hat{\mathbf{x}}_0$ (weakly) and the PDMP is non-Zeno*,* robustly transversal*, has the* robust activation property *and it is* size-compatible*, then $\hat{\mathbf{X}}^{(N)}(t)$ converges weakly to $\hat{\mathbf{x}}(t)$, $\hat{\mathbf{X}}^{(N)} \Rightarrow \hat{\mathbf{x}}$, as random elements in the space of cadlag function with the Skorohod metric.*

Proof. The argument closely follows the proof of Theorem 5.1. The only difference is the definition of jump times $T_i^{(N)}$ and T_i. In this case, in fact, these are defined as the minimum of stochastic jump times $\tau_i^{(N)}$ and τ_i (conditional on having observed $i-1$ jumps) and instantaneous jump times $\zeta_i^{(N)}$ and ζ_i (conditional on having observed $i-1$ jumps). Now, as $\tau_i^{(N)} \Rightarrow \tau_i$ by Lemma D.2 and $\zeta_i^{(N)} \Rightarrow \zeta_i$ by Lemma 6.1, by the continuous mapping theorem it follows that $T_i^{(N)} = \min\{\tau_i^{(N)}, \zeta_i^{(N)}\} \Rightarrow \min\{\tau_i, \zeta_i\} = T_i$.

This allows us to extend Corollary D.1, by replacing $\tau^{(N)} \Rightarrow \tau$ with $T^{(N)} \Rightarrow T$ in point 2, and then showing $\hat{\mathbf{X}}^{(N)}(T^{(N)}) \Rightarrow \hat{\mathbf{x}}(T)$ (use Proposition D.2 conditional on $T < \infty$). As for convergence of the state after the reset, notice that $\hat{\mathbf{Y}}^{(N)} = \hat{\mathbf{X}}^{(N)}(T^{(N)})$ and $\hat{\mathbf{y}} = \hat{\mathbf{x}}(T)$ satisfy the conditions of Lemma D.3. Moreover, we have $R^{(N)}(\hat{\mathbf{Y}}^{(N)}) = R_s^{(N)}(\hat{\mathbf{Y}}^{(N)})\mathbf{I}\{\tau_i^{(N)} < \zeta_i^{(N)}\} + R_i^{(N)}(\hat{\mathbf{Y}}^{(N)})\mathbf{I}\{\tau_i^{(N)} > \zeta_i^{(N)}\}$, and $R(\hat{\mathbf{y}}) = R_s(\hat{\mathbf{y}})\mathbf{I}\{\tau_i < \zeta_i\} + R_i(\hat{\mathbf{y}})\mathbf{I}\{\tau_i < \zeta_i\}$, where $R_s^{(N)}(\hat{\mathbf{Y}}^{(N)})$ and $R_s(\hat{\mathbf{y}})$ are the resets kernels for stochastic jumps (constructed from instantaneous transitions) and $R_i^{(N)}(\hat{\mathbf{Y}}^{(N)})$ and $R_i(\hat{\mathbf{y}})$ are the resets kernels for the instantaneous jumps.
Both satisfy $R_s^{(N)}(\hat{\mathbf{Y}}^{(N)}) \Rightarrow R_s(\hat{\mathbf{y}})$ and $R_i^{(N)}(\hat{\mathbf{Y}}^{(N)}) \Rightarrow R_i(\hat{\mathbf{y}})$, as $\hat{\mathbf{Y}}^{(N)} \to \hat{\mathbf{y}}$. Now, as $\mathbb{P}\{\zeta_i = \tau_i\} = 0$, we can apply the continuous mapping theorem first to the indicator functions $\mathbf{I}\{\tau < \zeta\}$ and $\mathbf{I}\{\tau > \zeta\}$, to show that $\mathbf{I}\{\tau_i^{(N)} < \zeta_i^{(N)}\} \Rightarrow \mathbf{I}\{\tau_i < \zeta_i\}$ and $\mathbf{I}\{\tau_i^{(N)} > \zeta_i^{(N)}\} \Rightarrow \mathbf{I}\{\tau_i > \zeta_i\}$, and then to the definition of $R^{(N)}$ and R, to show that $R^{(N)}(\hat{\mathbf{Y}}^{(N)}) \Rightarrow R(\hat{\mathbf{y}})$.
Reasoning similarly to Lemma C.1, we have then proved the equivalent of point 4 and 5 of Corollary D.1. Then, the proof of the theorem works as in Theorem 5.1. □

D.2 Guards in discrete stochastic transitions

We turn now to prove convergence in the presence of guarded discrete stochastic transitions. As discussed in the paper, there are two main issues to deal in this case, caused by the introduction of discontinuities in the rate functions. The first is the convergence of jump times, the second is the convergence of states after the resets. Both points require an additional regularity property of the PDMP, namely the *robust compatibility* with respect to guards of discrete stochastic transitions.

We start by showing convergence of exit times. Recall that we have m, say, discrete stochastic transitions, with rate functions $\hat{\lambda}_i^{(N)}$, $\hat{\lambda}_i$, and activation functions $h_i^{(N)}$ and h_i associated with guards, such that $\hat{\lambda}_i$ and h_i are continuous, and $\hat{\lambda}_i^{(N)}$, $h_i^{(N)}$ converge uniformly on compact sets to $\hat{\lambda}_i$ and h_i, respectively. Let $\hat{\lambda}^{(N)}(\hat{\mathbf{x}}) = \sum_{i=1}^m \mathbf{I}\{h_i^{(N)}(\hat{\mathbf{x}}) \geq 0\}\hat{\lambda}_i^{(N)}(\hat{\mathbf{x}})$ and $\hat{\lambda}(\hat{\mathbf{x}}) = \sum_{i=1}^m \mathbf{I}\{h_i(\hat{\mathbf{x}}) \geq 0\}\hat{\lambda}_i(\hat{\mathbf{x}})$. Furthermore, we consider the following discontinuity surfaces: $\mathcal{H}_i^{(N)} = \{h_i^{(N)}(\hat{\mathbf{x}}) = 0\}$ and $\mathcal{H}_i = \{h_i(\hat{\mathbf{x}}) = 0\}$, $\mathcal{H}^{(N)} = \bigcup_{i=1}^m \mathcal{H}_i^{(N)}$, and $\mathcal{H} = \bigcup_{i=1}^m \mathcal{H}_i$.

We then can prove the following

Lemma D.4. *Let $\hat{\mathbf{X}}^{(N)} \Rightarrow \hat{\mathbf{X}}$, as random variables in the space of cadlag functions, with $\hat{\mathbf{X}}$ a.s. continuous and such that, with probability 1, $\{t \in \mathbb{R}^+ \mid \hat{\mathbf{X}}(t) \in \mathcal{H}\}$ has Lebesgue measure 0. Let $\tau^{(N)}$, τ be the jump times associated with rates $\hat{\lambda}^{(N)}$ and $\hat{\lambda}$ defined above. Then $\tau^{(N)} \Rightarrow \tau$, as $N \to \infty$.*

Proof. We first prove that, if $\hat{\mathbf{x}}^{(N)} \to \hat{\mathbf{x}}$ as elements in the space of cadlag functions, $\hat{\mathbf{x}}$ is continuous, and $\hat{\lambda}(\hat{\mathbf{x}}(t))$ is almost everywhere continuous, then $\Lambda^{(N)}(T) = \int_0^T \hat{\lambda}^{(N)}(\hat{\mathbf{x}}^{(N)}(t))\, \mathrm{d}t \to \int_0^T \hat{\lambda}(\hat{\mathbf{x}}(t))\, \mathrm{d}t = \Lambda(T)$ for every $T > 0$. In fact, for each continuity point $\hat{\mathbf{y}}$ of $\hat{\lambda}$ we have that $\hat{\lambda}^{(N)}(\hat{\mathbf{y}}^{(N)}) \to \hat{\lambda}(\hat{\mathbf{y}})$ as $\hat{\mathbf{y}}^{(N)} \to \hat{\mathbf{y}}$. It follows that, for each $t > 0$ such that $\hat{\lambda}(\hat{\mathbf{x}}(t))$ is continuous, then $\hat{\lambda}^{(N)}(\hat{\mathbf{x}}^{(N)}(t)) \to \hat{\lambda}(\hat{\mathbf{x}}(t))$, as $\hat{\mathbf{x}}^{(N)}(t) \to \hat{\mathbf{x}}(t)$ by continuity of $\hat{\mathbf{x}}$. Therefore, $\hat{\lambda}^{(N)}(\hat{\mathbf{x}}^{(N)}) \to \hat{\lambda}(\hat{\mathbf{x}})$ pointwise almost everywhere in $[0, T]$.

Furthermore, by continuity of $\hat{\lambda}_i$, we have $\hat{\lambda}_i^{(N)}(\hat{\mathbf{x}}^{(N)}) \to \hat{\lambda}_i(\hat{\mathbf{x}})$, hence $\{\hat{\lambda}_i^{(N)}(\hat{\mathbf{x}}^{(N)}), \hat{\lambda}_i(\hat{\mathbf{x}})\}$ is relatively compact in the space of cadlag functions, and so bounded uniformly. It means that there is $M_i > 0$ such that $\|\hat{\lambda}_i^{(N)}(\hat{\mathbf{x}}^{(N)}(t))\| \leq M_i$ and $\|\hat{\lambda}_i(\hat{\mathbf{x}}(t))\| \leq M_i$. But as $\hat{\lambda}^{(N)}(\hat{\mathbf{x}}^{(N)}(t)) \leq \sum_i \hat{\lambda}_i^{(N)}(\hat{\mathbf{x}}^{(N)}(t))$ and $\hat{\lambda}(\hat{\mathbf{x}}(t)) \leq \sum_i \hat{\lambda}_i(\hat{\mathbf{x}}(t))$, it follows that $\hat{\lambda}^{(N)}(\hat{\mathbf{x}}^{(N)}), \hat{\lambda}(\hat{\mathbf{x}})$ are bounded by $\sum_i M_i$. Hence, by the bounded convergence theorem, $\int_0^T \hat{\lambda}^{(N)}(\hat{\mathbf{x}}^{(N)}(t))\ \mathrm{d}t \to \int_0^T \hat{\lambda}(\hat{\mathbf{x}}(t))\ \mathrm{d}t$ for every $T > 0$.

Now, the statement follows by applying the Skorohod representation theorem, as in Lemma D.2. Let $\tilde{\mathbf{X}}^{(N)}$, $\tilde{\mathbf{X}}$ be representations of $\hat{\mathbf{X}}^{(N)}$, $\hat{\mathbf{X}}$ on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ such that $\tilde{\mathbf{X}}^{(N)} \to \tilde{\mathbf{X}}$ almost surely. Hence, due to the hypothesis, for ω in a subset of probability 1 of Ω, we have that $\tilde{\mathbf{X}}^{(N)}(\omega) \to \tilde{\mathbf{X}}(\omega)$, $\tilde{\mathbf{X}}(\omega)$ is continuous, and $\hat{\lambda}(\tilde{\mathbf{X}}(\omega))$ is almost everywhere continuous. Then we can apply the previous argument to $\tilde{\mathbf{X}}^{(N)}(\omega)$, $\tilde{\mathbf{X}}(\omega)$, and conclude that $\Lambda^{(N)}(\omega, T)$ converges pointwise to $\Lambda(\omega, T)$ for each T, from which we get a.s. convergence of the representation of jump times $\tau^{(\tilde{N})}$ and $\tilde{\tau}$. Hence $\tau^{(N)} \Rightarrow \tau$, as desired. □

We turn now our attention to reset kernels. We will extend Lemma C.1 to deal with the discontinuities in the limit kernel using the hypothesis that there is zero probability of being in a discontinuous state when we jump. Recall the definition of $\hat{\lambda}^{(N)}$, $\hat{\lambda}$, $h_i^{(N)}$ and h_i, and further let $R_i^{(N)}$, R_i be the reset kernels, satisfying $R_i^{(N)}(\hat{\mathbf{x}}^{(N)}) \Rightarrow R_i(\hat{\mathbf{x}})$, for each $\hat{\mathbf{x}}^{(N)} \to \hat{\mathbf{x}}$. Then the full reset kernels are $R^{(N)}(\hat{\mathbf{x}}, \cdot) = \sum_{i=1}^m \mathbf{I}\{h_i^{(N)}(\hat{\mathbf{x}}) \geq 0\}(\hat{\lambda}_i^{(N)}(\hat{\mathbf{x}})/\hat{\lambda}^{(N)}(\hat{\mathbf{x}}))R_i^{(N)}(\hat{\mathbf{x}}, \cdot)$, and $R(\hat{\mathbf{x}}, \cdot) = \sum_{i=1}^m \mathbf{I}\{h_i(\hat{\mathbf{x}}) \geq 0\}(\hat{\lambda}_i(\hat{\mathbf{x}})/\hat{\lambda}(\hat{\mathbf{x}}))R_i(\hat{\mathbf{x}}, \cdot)$.

Equipped with these definitions, we can prove the following lemma.

Lemma D.5. *Let $R^{(N)}(\hat{\mathbf{y}})$ and $R(\hat{\mathbf{y}})$ defined as before and let $\hat{\mathbf{Y}}^{(N)} \Rightarrow \hat{\mathbf{Y}}$, where $\hat{\mathbf{Y}}^{(N)}$, $\hat{\mathbf{Y}}$ are random elements in E such that $\mathbb{P}\{\hat{\mathbf{Y}} \in \mathcal{H}\} = 0$. Then $R^{(N)}(\hat{\mathbf{Y}}^{(N)}) \Rightarrow R(\hat{\mathbf{Y}})$ and $(\hat{\mathbf{Y}}^{(N)}, R^{(N)}(\hat{\mathbf{Y}}^{(N)})) \Rightarrow (\hat{\mathbf{Y}}, R(\hat{\mathbf{Y}}))$.*

Proof. The proof is based Lemma C.1, with additional arguments taking care of the discontinuities in $R^{(N)}$ and R. Fix $\varepsilon > 0$ and a bounded and uniformly continuous function $g : E \to \mathbb{R}$. By the same argument of Lemma C.1, $\{\hat{\mathbf{Y}}^{(N)}, \hat{\mathbf{Y}}\}$ is uniformly tight, and so there is a compact set K_ε such that $P^{(N)}(K_\varepsilon) \geq 1 - \varepsilon/4\|g\|_\infty$ for each N, and $P(K_\varepsilon) \geq 1 - \varepsilon/4\|g\|_\infty$. Furthermore, for $\delta \geq 0$, let $\mathcal{H}_{i,\delta}$ be the closed δ-neighbourhood of $\mathcal{H}_i$, defined by $\mathcal{H}_{i,\delta} = \overline{\bigcup_{\hat{\mathbf{x}} \in \mathcal{H}_i} B_\delta(\hat{\mathbf{x}})}$, where $B_\delta(\hat{\mathbf{x}})$ is the ball of radius δ centred in $\hat{\mathbf{x}}$. Let also $\mathcal{H}_\delta = \bigcup_i \mathcal{H}_{i,\delta}$. Clearly $\mathcal{H}_\delta \downarrow \mathcal{H}$ for $\delta \downarrow 0$, and so $P(\mathcal{H}_\delta) \downarrow 0$. Choose δ such that $P(\mathcal{H}_\delta) < \varepsilon/4\|g\|_\infty$. We have that

$$
\begin{aligned}
|\mathbb{E}[g(R^{(N)}(\hat{\mathbf{Y}}^{(N)}))] - \mathbb{E}[g(R(\hat{\mathbf{Y}}))]| \quad \leq \quad & \underbrace{\left| \int_{K_\varepsilon \cap \mathcal{H}_\delta^c} R^{(N)} g(\hat{\mathbf{y}}) P^{(N)}(\hat{\mathbf{y}})\ \mathrm{d}\hat{\mathbf{y}} - \int_{K_\varepsilon \cap \mathcal{H}_\delta^c} R g(\hat{\mathbf{y}}) P(\hat{\mathbf{y}})\ \mathrm{d}\hat{\mathbf{y}} \right|}_{(a)} \\
+ \quad & \underbrace{\left| \int_{\mathcal{H}_\delta} R^{(N)} g(\hat{\mathbf{y}}) P^{(N)}(\hat{\mathbf{y}})\ \mathrm{d}\hat{\mathbf{y}} - \int_{\mathcal{H}_\delta} R g(\hat{\mathbf{y}}) P(\hat{\mathbf{y}})\ \mathrm{d}\hat{\mathbf{y}} \right|}_{(b)} \\
+ \quad & \underbrace{\left| \int_{K_\varepsilon^c} R^{(N)} g(\hat{\mathbf{y}}) P^{(N)}(\hat{\mathbf{y}})\ \mathrm{d}\hat{\mathbf{y}} - \int_{K_\varepsilon^c} R g(\hat{\mathbf{y}}) P(\hat{\mathbf{y}})\ \mathrm{d}\hat{\mathbf{y}} \right|}_{(c)}
\end{aligned}
$$

Now,the reset kernels $R^{(N)}$ and R in $K_\varepsilon \cap \mathcal{H}_\delta^c$ satisfy the continuity property $R^{(N)}(\hat{\mathbf{x}}^{(N)}) \Rightarrow R(\hat{\mathbf{x}})$ as $\hat{\mathbf{x}}^{(N)} \to \hat{\mathbf{x}}$, for $\hat{\mathbf{x}}^{(N)}, \hat{\mathbf{x}} \in K_\varepsilon \cap \mathcal{H}_\delta^c$. This follows because, by uniform convergence of $h_i^{(N)}$ to h_i on the compact set K_ε and the fact that $K_\varepsilon \cap \mathcal{H}_\delta^c$ does not contain any discontinuity surface, if $\hat{\mathbf{x}}^{(N)} \to \hat{\mathbf{x}}$, then $\hat{\mathbf{x}}^{(N)}$ will ultimately satisfy the same guards as $\hat{\mathbf{x}}$, i.e. $\mathbf{I}\{h_i^{(N)}(\hat{\mathbf{x}}^{(N)}) \geq 0\} \to \mathbf{I}\{h_i(\hat{\mathbf{x}}) \geq 0\}$. Then convergence of the reset kernels follows as in the unguarded case. This means that we can apply Lemma C.1 and conclude that term (a) goes to zero.

As for term (b), notice that $\mathcal{H}_\delta$ is closed, hence $\limsup_{N\to\infty} P^{(N)}(\mathcal{H}_\delta) \leq P(\mathcal{H}_\delta) \leq \varepsilon/4\|g\|_\infty$ by the Portmanteau theorem. Finally, term (c) is less than $\varepsilon/2$ by the choice of K_ε and the fact that Rg and $R^{(N)}g$ are both bounded by $\|g\|_\infty$. Hence we have that $\limsup_{N\to\infty} |\mathbb{E}[g(R^{(N)}(\hat{\mathbf{Y}}^{(N)}))] - \mathbb{E}[g(R(\hat{\mathbf{Y}}))]| \leq \varepsilon$, which

implies convergence to zero by the arbitrariness of ε. This proves $R^{(N)}(\hat{\mathbf{Y}}^{(N)}) \Rightarrow R(\hat{\mathbf{Y}})$. The second part of the statement, instead, follows as in Lemma C.1. □

We are finally ready to prove proposition 7.2.

Proposition (7.2)**.** *Let $(\mathcal{A}, \gamma_N)$ be a sequence of population-**sCCP** models for increasing systems size $\gamma_N \to \infty$, as $N \to \infty$, with variables partitioned into $\mathbf{X} = (\mathbf{X}_d, \mathbf{X}_c, \mathbf{X}_e)$, with discrete stochastic actions satisfying either scaling 3 or scaling 8, no instantaneous actions, and continuous actions satisfying scaling 2. Let $\hat{\mathbf{X}}^{(N)}(t)$ be the associated sequence of normalized CTMC and $\hat{\mathbf{x}}(t)$ be the limit PDMP associated with the normalized limit TDSHA $\hat{\mathcal{T}}(\mathcal{A})$.*

If $\hat{\mathbf{x}}_0^{(N)} \Rightarrow \hat{\mathbf{x}}_0$ (weakly) and the PDMP is non-Zeno *and* robustly compatible*, then $\hat{\mathbf{X}}^{(N)}(t)$ converges weakly to $\hat{\mathbf{x}}(t)$, $\hat{\mathbf{X}}^{(N)} \Rightarrow \hat{\mathbf{x}}$, as random elements in the space of cadlag function with the Skorohod metric.*

Proof. The proof proceeds essentially as that of Theorem 5.1. The only difference is that we have to replace Lemma D.2 with Lemma D.4 and Lemma C.1 with Lemma D.5 in corollary D.1 and in the proof of Theorem 5.1. To do this, we just need to show that the robust compatibility of the PDMP guarantees the satisfaction of the hypothesis of the two lemmas. This is trivial for Lemma D.4, as robust compatibility is an explicit hypothesis. The condition of Lemma D.5, instead, holds because robust compatibility of the PDMP $\hat{\mathbf{x}}$ and the absolute continuity of the exponential distribution with respect to the Lebesgue measure imply that the event $\{\hat{\mathbf{x}}(\tau) \in \mathcal{H}\}$ has probability zero. □