Programming Discrete Distributions with Chemical Reaction Networks*

Luca Cardelli[1,2], Marta Kwiatkowska[2], and Luca Laurenti[2]

[1] Microsoft Research
[2] Department of Computer Science, University of Oxford

Abstract. We explore the range of probabilistic behaviours that can be engineered with Chemical Reaction Networks (CRNs). We show that at steady state CRNs are able to "program" any distribution with finite support in $\mathbb{N}^m$, with $m \geq 1$. Moreover, any distribution with countable infinite support can be approximated with arbitrarily small error under the L^1 norm. We also give optimized schemes for special distributions, including the uniform distribution. Finally, we formulate a calculus to compute on distributions that is complete for finite support distributions, and can be compiled to a restricted class of CRNs that at steady state realize those distributions.

1 Introduction

Individual cells and viruses operate in a noisy environment and molecular interactions are inherently stochastic. How cells can tolerate and take advantage of noise (stochastic fluctuations) is a question of primary importance. It has been shown that noise has a functional role in cells [11]; indeed, some critical functions depend on the stochastic fluctuations of molecular populations and would be impossible in a deterministic setting. For instance, noise is fundamental for probabilistic differentiation of strategies in organisms, and is a key factor for evolution and adaptation [5]. In Escherichia coli, randomly and independently of external inputs, a small sub-population of cells enters a non-growing state in which they can elude the action of antibiotics that can only kill actively growing bacterial cells. Thus, when a population of E. coli cells is treated with antibiotics, the persisted cells survive by virtue of their quiescence before resuming growth [14]. This is an example in which molecular systems compute by producing a distribution. In other cases cells need to shape noise and compute on distributions instead of simply mean values. For example, in [16] the authors show, both mathematically and experimentally, that microRNA confers precision on the protein expression: it shapes the noise of genes in a way that decreases the intrinsic noise in protein expression, maintaining its expected value almost constant. Thus, although fundamentally important, the mechanisms used by cells to compute in a stochastic environment are not well understood.

Chemical Reaction Networks (CRNs) with mass action kinetics are a well studied formalism for modelling biochemical systems, more recently also used as a formal programming language [10]. It has been shown that any CRN can be

*This research is supported by a Royal Society Research Professorship and ERC AdG VERIWARE.

physically implemented by a corresponding DNA strand displacement circuit in a well-mixed solution [18]. DNA-based circuits thus have the potential to operate inside cells and control their activity. Winfree and Qian have also shown that CRNs can be implemented on the surface of a DNA nanostructure [15], enabling localized computation and engineering biochemical systems where the molecular interactions occur between few components. When the number of interacting entities is small, the stochastic fluctuations intrinsic in molecular interactions play a predominant role in the time evolution of the system. As a consequence, "programming" a CRN to provide a particular probabilistic response for a subset of species, for example in response to environmental conditions, is important for engineering complex biochemical nano-devices and randomized algorithms. In this paper, we explore the capacity of CRNs to "program" discrete probability distributions. We aim to characterize the probabilistic behaviour that can be obtained, exploring both the capabilities of CRNs for producing distributions and for computing on distributions by composing them.

Contributions. We show that at steady state CRNs are able to compute any distribution with support in $\mathbb{N}^m$, with $m \geq 1$. We propose an algorithm to systematically "program" a CRN so that its stochastic semantics at steady state approximates a given distribution with arbitrarily small error under the L^1 norm. The approximation is exact if the support of the distribution is finite. The resulting network has a number of reactions linear in the dimension of the support of the distribution and the output is produced monotonically allowing composition. Since distributions with large support can result in unwieldy networks, we also give optimised networks for special distributions, including a novel scheme for the uniform distribution. We formulate a calculus that is complete for finite support distributions, which can be compiled to a restricted class of CRNs that at steady state compute those distributions. The calculus allows for modelling of external influences on the species. Our results are of interest for a variety of scenarios in systems and synthetic biology. For example, they can be used to program a biased stochastic coin or a uniform distribution, thus enabling implementation of randomized algorithms and protocols in CRNs.

Related work. It has been shown that CRNs with stochastic semantics are Turing complete, up to an arbitrarily small error [17]. If we assume error-free computation, their computational power decreases: they can decide the class of the semi-linear predicates [4] and compute semi-linear functions [9]. A first attempt to model distributions with CRNs can be found in [12], where the problem of producing a single distribution is studied. However, their circuits are approximated and cannot be composed to compute operations on distributions.

2 Chemical Reaction Networks

A *chemical reaction network (CRN)* (Λ, R) is a pair of finite sets, where Λ is the set of *chemical species*, $|\Lambda|$ denotes its size, and R is a set of reactions. A *reaction* $\tau \in R$ is a triple $\tau = (r_\tau, p_\tau, k_\tau)$, where $r_\tau \in \mathbb{N}^{|\Lambda|}$ is the *source complex*, $p_\tau \in \mathbb{N}^{|\Lambda|}$ is the *product complex* and $k_\tau \in \mathbb{R}_{>0}$ is the coefficient associated to the rate of the reaction, where we assume $k_\tau = 1$ if not specified; r_τ and p_τ

represent the stoichiometry of reactants and products. Given a reaction $\tau_1 = ([1,0,1],[0,2,0],k_1)$ we often refer to it as $\tau_1 : \lambda_1 + \lambda_3 \rightarrow^{k_1} 2\lambda_2$. The *net change* associated to τ is defined by $\upsilon_\tau = p_\tau - r_\tau$.

We assume that the system is well stirred, that is, the probability of the next reaction occurring between two molecules is independent of the location of those molecules, at fixed volume V and temperature. Under these assumptions a *configuration* or *state* of the system $x \in \mathbb{N}^{|\Lambda|}$ is given by the number of molecules of each species. A *chemical reaction system* (CRS) $C = (\Lambda, R, x_0)$ is a tuple where (Λ, R) is a CRN and $x_0 \in \mathbb{N}^{|\Lambda|}$ represents its initial condition.

Stochastic semantics. In this paper we consider CRNs with stochastic semantics. The propensity rate α_τ of a reaction τ is a function of the current configuration of the system x such that $\alpha_\tau(x)dt$ is the probability that a reaction event occurs in the next infinitesimal interval dt. We assume mass action kinetics [2]. That is, if $\tau : \lambda_1 + ... + \lambda_k \rightarrow^k \cdot$, then $\alpha_\tau(x) = k \cdot \prod_{i=1}^{|\Lambda|} \frac{x(\lambda_i)!}{(x(\lambda_i)-r_{\tau,i})!}$, where $r_{\tau,i}$ is the $i-$th component of vector r.[3] The time evolution of a CRS $C = (\Lambda, R, x_0)$ can be modelled as a time-homogeneous *Continuous Time Markov Chain* (CTMC) $(X^C(t), t \in \mathbb{R}_{\geq 0})$, with state space S [2]. When clear from the context we write $X(t)$ instead of $X^C(t)$. $Q : S \times S \rightarrow \mathbb{R}$ is the generator matrix of X, given by $Q(x_i, x_j) = \sum_{\{\tau \in R | x_j = x_i + \upsilon_\tau\}} \alpha_\tau(x_i)$ for $i \neq j$ and $Q(x_i, x_i) = -\sum_{j=1 \wedge j \neq i}^{|S|} Q(x_i, x_j)$. We denote $P^C(t)(x) = Prob(X^C(t) = x | X^C(0) = x_0)$, where x_0 is the initial configuration. $P^C(t)(x)$ represents the transient evolution of X, and can be calculated exactly by solving the *Chemical Master Equation (CME)* or by approximation of the CME [7].

Definition 1 *The steady state distribution (or limit distribution) of X^C is defined as $\pi^C = \lim_{t \rightarrow \infty} P^C(t)$.*

Again, when clear from the context, instead of π^C we simply write π. π calculates the percentage of time, in the long-run, that X spends in each state $x \in S$. If S is finite, then the above limit distribution always exists and is unique [13]. In this paper we focus on discrete distributions, and will sometimes conflate the term distribution with probability mass function, defined next.

Definition 2 *Suppose that $M : S \rightarrow \mathbb{R}^m$ with $m > 0$ is a discrete random variable defined on a countable sample space S. Then the probability mass function (pmf) $f : \mathbb{R}^m \rightarrow [0,1]$ for M is defined as $f(x) = Prob(s \in S \mid M(s) = x)$.*

For a pmf $\pi : \mathbb{N}^m \rightarrow [0,1]$ we call $J = \{y \in \mathbb{N}^m | \pi(y) \neq 0\}$ the support of π. A pmf is always associated to a discrete random variable whose distribution is described by the pmf. However, sometimes, when we refer to a pmf, we imply the associated random variable. Given two pmfs f_1 and f_2 with values in $\mathbb{N}^m$, $m > 0$, we define the L^1 norm (or distance) between them as $d_1(f_1, f_2) = \sum_{n \in \mathbb{N}^m} (|f_1(n) - f_2(n)|)$. Note that, as f_1, f_2 are pmfs, then $d_1(f_1, f_2) \leq 2$. It is worth stressing that, given the CTMC X, for each $t \in \mathbb{R}_{\geq 0}$, $X(t)$ is a random variable defined on a countable

[3] The reaction rate k depends on the volume V. However, as the volume is fixed, in our notation V is embedded inside k.

state space. As a consequence, its distribution is given by a pmf. Likewise, the limit distribution of a CTMC, if it exists, is a pmf.

Definition 3 *Given $C = (\Lambda, R)$ and $\lambda \in \Lambda$, we define $\pi_\lambda(k) = \sum_{\{x \in S | x(\lambda)=k\}} \pi(x)$ as the probability that for $t \to \infty$, in X^C, there are k molecules of λ.*

π_λ is a pmf representing the steady state distribution of species λ.

3 On approximating discrete distributions with CRNs

We now show that, ffor a pmf with support in $\mathbb{N}$, we can always build a CRS such that, at steady state (i.e. for $t \to \infty$) the random variable representing the molecular population of a given species in the CRN approximates that distribution with arbitrarily small error under the L^1 norm. The result is then generalised to distributions with domain in $\mathbb{N}^m$, with $m \geq 1$. The approximation is exact in case of finite support.

3.1 Programming pmfs

Definition 4 *Given $f : \mathbb{N} \to [0, 1]$ with finite support $J = (z_1, ..., z_{|J|})$ such that $\sum_{i=1}^{|J|} f(z_i) = 1$, we define the CRS $C_f = (\Lambda, R, x_0)$ as follows. C_f is composed of $2|J|$ reactions and $2|J|+2$ species. For any $z_i \in J$ we have two species $\lambda_i, \lambda_{i,i} \in \Lambda$ such that $x_0(\lambda_i) = z_i$ and $x_0(\lambda_{i,i}) = 0$. Then, we consider a species $\lambda_z \in \lambda$ such that $x_0(\lambda_z) = 1$, and the species $\lambda_{out} \in \Lambda$, which represents the output of the network and such that $x_0(\lambda_{out}) = 0$. For every $z_i \in J$, R has the following two reactions: $\tau_{i,1} : \lambda_z \to^{f(z_i)} \lambda_{i,i}$ and $\tau_{i,2} : \lambda_i + \lambda_{i,i} \to \lambda_{out} + \lambda_{i,i}$.*

Example 1. Consider the probability mass function $f : \mathbb{N} \to [0, 1]$ defined as $f(y) = \{\frac{1}{6}$ if $y = 2; \frac{1}{3}$ if $y = 5; \frac{1}{2}$ if $y = 10; 0$ otherwise$\}$. Let $\Lambda = \{\lambda_1, \lambda_2, \lambda_3, \lambda_z, \lambda_{1,1}, \lambda_{2,2}, \lambda_{3,3}, \lambda_{out}\}$, then we build the CRS $C = (\Lambda, R, x_0)$ following Definition 4, where R is given by the following set of reactions:

$$\lambda_z \to^{\frac{1}{6}} \lambda_{1,1}; \quad \lambda_z \to^{\frac{1}{3}} \lambda_{2,2}; \quad \lambda_z \to^{\frac{1}{2}} \lambda_{3,3};$$

$$\lambda_1 + \lambda_{1,1} \to^1 \lambda_{1,1} + \lambda_{out}; \quad \lambda_2 + \lambda_{2,2} \to^1 \lambda_{2,2} + \lambda_{out}; \quad \lambda_3 + \lambda_{3,3} \to^1 \lambda_{3,3} + \lambda_{out}.$$

The initial condition x_0 is $x_0(\lambda_{out}) = x_0(\lambda_{1,1}) = x_0(\lambda_{2,2}) = x_0(\lambda_{3,3}) = 0$; $x_0(\lambda_1) = 2$; $x_0(\lambda_2) = 5$; $x_0(\lambda_3) = 10$; $x_0(\lambda_z) = 1$. Theorem 1 ensures $\pi_{\lambda_{out}} = f$.

Theorem 1. *Given a pmf $f : \mathbb{N} \to [0, 1]$ with finite support J, the CRS C_f as defined in Definition 4 is such that $\pi^{C_f}_{\lambda_{out}} = f$.*

Proof. Let $J = (z_1, .., z_{|J|})$ be the support of f, and $|J|$ its size. Suppose $|J|$ is finite, then the set of reachable states from x_0 is finite by construction and the limit distribution of X^{C_f}, the induced CTMC, exists. By construction, in the initial state x_0 only reactions of type $\tau_{i,1}$ can fire, and the probability that a specific $\tau_{i,1}$ fires first is exactly:

$$\frac{\alpha_{\tau_{i,1}}(x_0)}{\sum_{j=1}^{|J|} \alpha_{\tau_{j,1}}(x_0)} = \frac{f(z_i) \cdot 1}{\sum_{j=1}^{|J|} f(z_j) \cdot 1} = \frac{f(z_i)}{\sum_{j=1}^{|J|} f(z_j)} = \frac{f(z_i)}{1} = f(z_i)$$

Observe that the firing of the first reaction uniquely defines the limit distribution of X^{C_f}, because λ_z is consumed immediately and only reaction $\tau_{i,2}$ can fire, with no race condition, until λ_i are consumed. This implies that at steady state λ_{out} will be equal to $x_0(\lambda_i)$, and this happens with probability $f(x_0(\lambda_i))$. Since $x_0(\lambda_i) = z_i$ for $i \in [1, |J|]$, we have $\pi^{C_f}_{\lambda_{out}} = f$. □

Then, we can state the following corollary of Theorem 1.

Corollary 1. *Given a pmf $f : \mathbb{N} \to [0, 1]$ with countable support J, we can always find a finite CRS C_f such that $\pi^{C_f}_{\lambda_{out}} = f$ with arbitrarily small error under the L^1 norm.*

Proof. Let $J = \{z_1, ..., z_{|J|}\}$. Suppose J is (countably) infinite, that is, $|J| \to \infty$. Then, we can always consider an arbitrarily large but finite number of points in the support, such that the probability mass lost is arbitrarily small, and applying Definition 4 on this finite subset of the support we have the result.

In order to prove the result consider the function f' with support $J' = \{z_1, ..., z_k\}$, $k \in \mathbb{N}$, such that $f(z_i) = f'(z_i)$, for all $i \in \mathbb{N}_{\leq k}$. Consider the series $\sum_{i=1}^{\infty} f(n)$. This is an absolute convergent series by definition of pmf. Then, we have that $\lim_{i\to\infty} f(i) = 0$ and, for any $\epsilon > 0$, we can choose some $\kappa_\varepsilon \in \mathbb{N}$, such that:

$$\forall k > \kappa_\varepsilon \quad |\sum_{i=1}^{k} f'(i) - \sum_{i=1}^{\infty} f(i)| < \frac{\epsilon}{2}.$$

This implies that for $k > \kappa_\varepsilon$ given $f'_k = \sum_{i=1}^{k} f'(i)$ we have, $d_1(f'_k, f) < \epsilon$. □

The following remark shows that the need for precisely tuning the value of reaction rates in Theorem 1 can be dropped by introducing some auxiliary species.

Remark 1. In practice, tuning the rates of a reaction can be difficult or impossible. However, it is possible to modify the CRS derived using Definition 4 in such a way the probability value is not encoded in the rates, and we just require that all reactions have the same rates. We can do that by using some auxiliary species $\Lambda_c = \{\lambda_{c_1}, \lambda_{c_2}, ..., \lambda_{c_{|\Lambda_c|}}\}$. Then, the reactions $\tau_{i,1}$ for $i \in [1, J]$ become $\tau_{i,1} : \lambda_z + \lambda_{c_i} \to^k \lambda_{i,i}$, for $k \geq 0$, instead of $\tau_{i,1} : \lambda_z \to^{f(y_i)} \lambda_{i,i}$, as in the original definition. The initial condition of λ_{c_i} is $x_0(\lambda_{c_i}) = f(y_i) \cdot L$, where $L \in \mathbb{N}$ is such that for $j \in [1, |J|]$ and $J = \{z_1, ..., z_{|J|}\}$ we have that $f(z_j) \cdot L$ is a natural number, assuming all the $f(z_j)$ are rationals.

Remark 2. In biological circuits the probability distribution of a species may depend on some external conditions. For example, the *lambda Bacteriofage* decides to lyse or not to lyse with a probabilistic distribution based also on environmental conditions [5]. Programming similar behaviour is possible by extension of Theorem 1. For instance, suppose, we want to program a switch that with rate $50 + Com$ goes in a state O_1, and with rate 5000 goes in a different state O_2, where Com is an external input. To program this logic we can use the following

reactions: $\tau_{1,1} : \lambda_z + \lambda_{c_1} \rightarrow^{k_1} \lambda_{O_1}$ and $\tau_{1,2} : \lambda_z + \lambda_{c_2} \rightarrow^{k_1} \lambda_{O_2}$, where λ_{O_1} and λ_{O_2} model the two logic states, initialized at 0. The initial condition x_0 is such that $x_0(\lambda_z) = 1$, $x_0(\lambda_{c_1}) = 50$ and $x_0(\lambda_{c_2}) = 5000$. Then, we add the following reaction $Com \rightarrow^{k_2} \lambda_{c_1}$. It is easy to show that if $k_2 >>> k_1$ then we have the desired probabilistic behaviour for any initial value of $Com \in \mathbb{N}$. This may be of interest also for practical scenarios in synthetic biology, where for instance the behaviour of synthetic bacteria needs to be externally controlled [3]; and, if each bacteria is endowed with a similar logic, then, by tuning Com, at the population level, it is possible to control the fraction of bacteria that perform this task.

In the next theorem we generalize to the multi-dimensional case.

Theorem 2. *Given $f : \mathbb{N}^m \rightarrow [0, 1]$ with $m \geq 1$ such that $\sum_{i \in \mathbb{N}} f(i) = 1$, then there exists a CRS $C = (\Lambda, R, x_0)$ such that the joint limit distribution of $(\lambda_{out_1}, \lambda_{out_2}, ..., \lambda_{out_m}) \in \Lambda$ approximates f with arbitrarily small error under the L^1 distance. The approximation is exact if the support of f is finite.*

To prove this theorem we can derive a CRS similar to that in the uni-dimensional case. The firing of the first reaction can be used to probabilistically determine the value at steady state of the m output species, using some auxiliary species.

3.2 Special distributions

For a given pmf the number of reactions of the CRS derived from Definition 4 is linear in the dimension of its support. As a consequence, if the support is large then the CRSs derived using Theorems 1 and 2 can be unwieldy. In the following we show three optimised CRSs to calculate the Poisson, binomial and uniform distributions. These CRNs are compact and applicable in many practical scenarios. However, using Definition 4 the output is always produced monotonically. In the circuits below this does not happen, but, on the other hand, the gain in compactness is substantial. The first two circuits have been derived from the literature, while the CRN for the uniform distribution is new.

Poisson distribution. The main result of [1] guarantees that all the CRNs that respect some conditions (weakly reversible, deficiency zero and irreducible state space, see [1]) have a distribution given by the product of Poisson distributions. As a particular case, we consider the following CRS composed of only one species λ and the following two reactions $\tau_1 : \emptyset \rightarrow^{k_1} \lambda$; $\tau_2 : \lambda \rightarrow^{k_2} \emptyset$. Then, at steady state, λ has a Poisson distribution with expected value $\frac{k_1}{k_2}$.

Binomial distribution. We consider the network introduced in [1]. The CRS is composed of two species, λ_1 and λ_2, with initial condition x_0 such that $x_0(\lambda_1) + x_0(\lambda_2) = K$ and the following set of reactions: $\tau_1 : \lambda_1 \rightarrow^{k_1} \lambda_2; \tau_2 : \lambda_2 \rightarrow^{k_2} \lambda_1$. As shown in [1], λ_1 and λ_2 at steady state have a binomial distribution such that: $\pi_{\lambda_1}(y) = (\frac{K}{y}){c_1}^y(1 - c_1)^{K-y}$ and $\pi_{\lambda_2}(y) = (\frac{K}{y}){c_2}^y(1 - c_2)^{K-y}$.

Uniform distribution. The following CRS computes the uniform distribution over the sum of the initial number of molecules in the system, independently of the initial value of each species. It has species λ_1 and λ_2 and reactions:

$\tau_1 : \lambda_1 \rightarrow^k \lambda_2; \quad \tau_2 : \lambda_2 \rightarrow^k \lambda_1; \quad \tau_3 : \lambda_1 + \lambda_2 \rightarrow^k \lambda_1 + \lambda_1; \quad \tau_4 : \lambda_1 + \lambda_2 \rightarrow^k \lambda_2 + \lambda_2$

For $k > 0$, τ_1 and τ_2 implement the binomial distribution. These are combined with τ_3 and τ_4, which implement a Direct Competition system [6], which has a bimodal limit distribution in 0 and in K, where $x_0(\lambda_1) + x_0(\lambda_2) = K$, with x_0 initial condition. This network, surprisingly, according to the next theorem, at steady state produces a distribution which varies uniformly between 0 and K.

Theorem 3. *Let $x_0(\lambda_1) + x_0(\lambda_2) = K \in \mathbb{N}$. Then, the CRS described above has the following steady state distribution for λ_1 and λ_2:*

$$\pi_{\lambda_1}(y) = \pi_{\lambda_2}(y) = \begin{cases} \frac{1}{K+1}, & \text{if } y \in [0, K] \\ 0, & \text{otherwise} \end{cases}$$

Proof. We consider a general initial condition x_0 such that $x_0(\lambda_1) = K - M$ and $x_0(\lambda_2) = M$ for $0 \leq M \leq K$ and $K, M \in \mathbb{N}$. Because any reaction has exactly 2 reagents and 2 products, we have the invariant that for any configuration x reachable from x_0 it holds that $x(\lambda_1) + x(\lambda_2) = K$. Figure 1 plots the CTMC semantics of the system. For any fixed K the set of reachable states

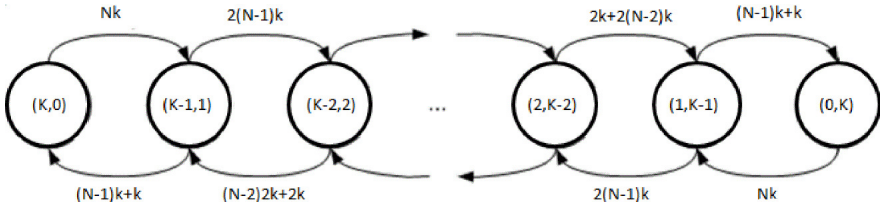

Fig. 1: The figure shows the CTMC induced by the CRS implementing the uniform distribution for initial condition x_0 such that $x_0(\lambda_1) + x_0(\lambda_2) = K$.

from any initial condition in the induced CTMC is finite (exactly K states are reachable from any initial condition) and irreducible. Therefore, the steady state solution exists, is unique and independent of the initial conditions. To find this limit distribution we can calculate Q, the infinitesimal generator of the CTMC, and then solve the linear equations system $\pi Q = 0$, with the constraint that $\sum_{i \in [0,K]} \pi_i = 1$, where π_i is the ith component of the vector π, as shown in [13]. Because the CTMC we are considering is irreducible, this is equivalent to solve the balance equations with the same constraint. The resulting π is the steady state distribution of the system.

We consider 3 cases, where $(K - j, j)$ for $j \in [0, K]$ represents the state of the system in terms of molecules of λ_1 and λ_2.

- Case $j = 0$. For the state $(K, 0)$, whose limit distribution is defined as $\pi(K, 0)$, we have the following balance equation:

$$-\pi(K,0)Kk + \pi(K-1,1)[(K-1)k + k] = 0 \implies$$

$$\pi(K,0) = \pi(K-1,1).$$

– Case $j \in [1, K-1]$. Observing Figure 1 we see that the states and the rates follow a precise pattern: every state is directly connected with only two states and for any transition the rates depend on two reactions, therefore we can consider the balance equations for a general state $(K-j, j)$ for $j \in [1, K-1]$ (for the sake of a lighter notation instead of $\pi(K-j, j)$ we write π^j):

$$\begin{aligned}
&\pi^{j-1}[K+1-j+(K+1-j)(j-1)] \\
&\qquad -\pi^j[2(K-j)j+j+K-j]+\pi^{j+1}[j+1+(K-j-1)(j+1)]=0 \\
&\qquad\qquad \Longrightarrow \\
&\pi^{j-1}[Kj-j^2+j]-\pi^j[2Kj-2j^2+K]+\pi^{j+1}[Kj+K-j^2-j]=0
\end{aligned}$$

  It is easy to verify that if $\pi^{j-1} = \pi^j = \pi^{j+1}$ then the equation is verified.
– Case $j = K$. The case for the state $(0, K)$ is similar to the case $(K, 0)$.

We have shown that each reachable state has equal probability at steady state for any possible initial condition. Therefore, because $\sum_{i=0}^{K} \pi^i = 1$ and $\pi_{\lambda_i}(y) = \sum_{x_j \in S | x_j(\lambda_i)=y} \pi^j$ for $y \geq 0$, we have that for both λ_1 and λ_2

$$\pi_{\lambda_1}(y) = \pi_{\lambda_2}(y) = \begin{cases} \frac{1}{K+1}, & \text{if } y \in [0, K] \\ 0, & \text{otherwise} \end{cases}$$

□

4 Calculus of limit distributions of CRNs

In the previous section we have shown that CRNs are able to program any pmf on $\mathbb{N}$. We now define a calculus to compose and compute on pmfs. We show it is complete with respect to finite support pmfs on $\mathbb{N}$. Then, we define a translation of this calculus into a restricted class of CRNs. We prove the soundness of such a translation, which thus yields an abstract calculus of limit distributions of CRNs. For simplicity, in what follows we consider only pmfs with support in $\mathbb{N}$, but the results can be generalised to the multi-dimensional case in a straightforward way.

Definition 5 *(Syntax). The syntax of formulae of our calculus is given by*

$$P := (P+P) \,|\, min(P,P) \,|\, k \cdot P \,|\, (P)_D : P \,|\, one \,|\, zero$$

$$D := p \,|\, p \cdot c_i + D$$

where $k \in \mathbb{Q}_{\geq 0}$, $p \in \mathbb{Q}_{[0,1]}$ are rational and $V = \{c_1, ..., c_n\}$ is a set of variables with values in $\mathbb{N}$.

A formula P denotes a pmf that can be obtained as a sum, minimum, multiplication by a rational, or convex combination of pmfs *one* and *zero*. Given a formula P, variables $V = \{c_1, ..., c_n\}$, called *environmental inputs*, model the influence of external factors on the probability distributions of the system. $V(P)$ represents the variables in P. An *environment* $E : V \to \mathbb{Q}_{[0,1]}$ is a partial function which maps each input c_i to its valuation normalized to $[0, 1]$. Given a formula

P and an environment E, where $V(P) \subseteq dom(E)$, with $dom(E)$ domain of E, we define its semantics, $[\![P]\!]_E$, as a pmf (the empty environment is denoted as $\emptyset$). D expresses a summation of valuations of inputs c_i weighted by rational probabilities p, which evaluates to a rational $[\![D]\!]_E$ for a given environment. We require that, for any D, the sum of p coefficients in D is in $[0, 1]$. This ensures that $0 \leq [\![D]\!]_E \leq 1$. The semantics is defined inductively as follows, where the operations on pmfs are defined in Section 4.1.

Definition 6 *(Semantics). Given formulae P, P_1, P_2 and an environment E, such that $V(P) \cup V(P_1) \cup V(P_2) \subseteq dom(E)$, we define*

$$[\![one]\!]_E = \pi_{one} \qquad [\![zero]\!]_E = \pi_{zero} \qquad [\![P_1 + P_2]\!]_E = [\![P_1]\!]_E + [\![P_2]\!]_E$$

$$[\![min(P_1, P_2)]\!]_E = min([\![P_1]\!]_E, [\![P_2]\!]_E)$$

$$[\![k \cdot P]\!]_E = \frac{k_1 \cdot ([\![P]\!]_E)}{k_2} \text{ for } k = \frac{k_1}{k_2} \text{ and } k_1, k_2 \in \mathbb{N}$$

$$[\![(P_1)_D : (P_2)]\!]_E = ([\![P_1]\!]_E)_{[\![D]\!]_E} : ([\![P_2]\!]_E)$$

$$[\![p]\!]_E = p \qquad [\![p \cdot c_i + D]\!]_E = p \cdot E(c_i) + ([\![D]\!]_E)$$

where $\pi_{one}(y) = \begin{cases} 1, & \text{if } y = 1 \\ 0, & \text{otherwise} \end{cases}$ *and* $\pi_{zero}(y) = \begin{cases} 1, & \text{if } y = 0 \\ 0, & \text{otherwise} \end{cases}$.

To illustrate the calculus, consider the Bernoulli distribution with parameter $p \in \mathbb{Q}_{[0,1]}$. We have $bern^p = (one)_p : zero$, where $[\![bern^p]\!]_\emptyset(y) = \{p \text{ if } y = 1; 1 - p \text{ if } y = 0; 0 \text{ otherwise}\}$. The binomial distribution can be obtained as a sum of n independent Bernoulli distributions of the same parameter. Given a random variable with a binomial distribution with parameters (n, p), if n is sufficiently large and p sufficiently small then this approximates a Poisson distribution with parameter $n \cdot p$.

4.1 Operations on distributions

In this section, we define a set of operations on pmfs needed to define the semantics of the calculus. We conclude the section by showing that these operations are sufficient to represent pmfs with finite support in $\mathbb{N}$.

Definition 7 *Let $\pi_1 : \mathbb{N} \to [0, 1]$, $\pi_2 : \mathbb{N} \to [0, 1]$ be two pmfs. Assume $p \in \mathbb{Q}_{[0,1]}$, $y \in \mathbb{N}$, $k_1 \in \mathbb{N}$ and $k_2 \in \mathbb{N}_{>0}$, then we define the following operations on pmfs:*

- *The sum or convolution of π_1 and π_2 is defined as $(\pi_1 + \pi_2)(y) = \sum_{(y_i, y_j) \in \mathbb{N} \times \mathbb{N} \, s.t. \, y_i + y_j = y} \pi_1(y_i)\pi_2(y_j)$.*
- *The minimum of π_1 and π_2 is defined as $min(\pi_1, \pi_2)(y) = \sum_{(y_i, y_j) \in \mathbb{N} \times \mathbb{N} \, s.t. \, min(y_i, y_j) = y} \pi_1(y_i)\pi_2(y_j)$.*
- *The multiplication of π_1 by the constant k_1 is defined as $(k_1\pi_1)(y) = \begin{cases} \pi_1(\frac{y}{k_1}), & \text{if } \frac{y}{k_1} \in \mathbb{N} \\ 0, & \text{otherwise} \end{cases}$*
- *The division of π_1 by the constant k_2 is defined as $\frac{\pi}{k_2}(y) = \sum_{y_i \in \mathbb{N} \, s.t. \, y = \lfloor y_i / k_2 \rfloor} \pi(y_i)$.*

– *The convex combination of* π_1 *and* π_2*, for* $y \in \mathbb{N}$*, is defined* $((\pi_1)_p : (\pi_2))(y) = p\pi_1(y) + (1-p)\pi_2(y)$.

The convex combination operator is the only one that is not closed with respect to pmfs whose support is a single point. Lemma 1 shows that this operator is not associative with respect to minimum and sum of pmfs.

Lemma 1. *Given probability mass functions* π_1, $\pi_2 : \mathbb{N} \to [0,1]$, $p_1, p_2, p_3, p_4 \in [0,1]$ *and* $k \in \mathbb{Q}_{\geq 0}$*, then the following equations hold:*

– $k((\pi_1)_p : \pi_2) = (k\pi_1)_p : (k\pi_2)$
– $((\pi_1)_{p_1} : \pi_2)_{p_2} : \pi_3 = (\pi_1)_{p_3} : ((\pi_2)_{p_4} : \pi_3)$ *iff* $p_3 = p_1 p_2$ *and* $p_4 = \frac{(1-p_1)p_2}{1-p_1 p_2}$
– $(\pi_1)_p : \pi_2 = (\pi_2)_{1-p} : \pi_1$
– $(\pi_1)_p : \pi_1 = \pi_1$.

Example 2. Consider the following formula

$$P_1 = (one)_{0.001 \cdot c + 0.2}(4 \cdot one) + (2 \cdot one)_{0.4}(3 \cdot one),$$

with set of environmental variables $V = \{c\}$ and an enviroment E such that $V(P_1) \subseteq dom(E)$. Then, according to Definition 7 we have that

$$[\![P_1]\!]_E(y) = \begin{cases} (0.001 \cdot [\![c]\!]_E + 0.2) \cdot 0.4, & \text{if } y = 3 \\ (0.001 \cdot [\![c]\!]_E + 0.2) \cdot 0.6, & \text{if } y = 4 \\ (1 - (0.001 \cdot [\![c]\!]_E + 0.2)) \cdot 0.4, & \text{if } y = 6 \\ (1 - (0.001 \cdot [\![c]\!]_E + 0.2)) \cdot 0.6, & \text{if } y = 7 \\ 0, & \text{otherwise} \end{cases}$$

Having formally defined all the operations on pmfs, we can finally state the following proposition guaranteeing that the semantics of any formula of the calculus is a pmf.

Proposition 1. *Given* P*, a formula of the calculus defined in Definition 5, and an environment* E *such that* $V(P) \subseteq dom(E)$*, then* $[\![P]\!]_E$ *is a pmf.*

The following theorem shows that our calculus is complete with respect to finite support distributions.

Theorem 4. *For any pmf* $f : \mathbb{N} \to [0,1]$ *with finite support there exists a formula* P *such that* $[\![P]\!]_\emptyset = f$.

Proof. Given a pmf $f : \mathbb{N} \to [0,1]$ with finite support $J = (z_1, ..., z_{|J|})$ we can define $P = (z_1 \cdot one)_{f(z_1)} : ((z_2 \cdot one)_{\frac{f(z_2)}{1-f(z_1)}} : (... : ((z_i \cdot one)_{\frac{f(z_i)}{\prod_{j=1}^{i-1}(1-f(z_j))}} : ... : ((z_n \cdot one))))$. Then, $[\![P]\!]_\emptyset = f$. □

Proof of Theorem 4 relies only on a subset of the operators, but the other operators are useful for composing previously defined pmfs.

5 CRN implementation

We now show how the operators of the calculus can be realized by operators on CRSs. The resulting CRSs produce the required distributions at steady state, that is, in terms of the steady state distribution of the induced CTMC. Thus, we need to consider a restricted class of CRNs that always stabilize and that can be incrementally composed. The key idea is that each such CRN has output species that cannot act as a reactant in any reaction, and hence the counts of those species increase monotonically.[4] This implies that the optimized CRSs shown in Section 3.2 cannot be used compositionally.

5.1 Non-reacting output CRSs (NRO-CRSs)

Since in the calculus presented in Definition 5 we consider only finite support pmfs, in this section we are limited to finite state CTMCs. This is important because some results valid for finite state CTMCs are not valid in infinite state spaces. Moreover, any pmf with infinite support on natural numbers can always be approximated under the L^1 norm (see Corollary 1).

Given a CRS $C = (\Lambda, R, x_0)$, we call the *non-reacting species* of C the subset of species $\Lambda_r \subseteq \Lambda$ such that given $\lambda_r \in \Lambda_r$ there does not exist $\tau \in R$ such that $r_\tau^{\lambda_r} > 0$, where $r_\tau^{\lambda_r}$ is the component of the source complex of the reaction τ relative to λ_r, that is, λ_r is not a reactant in any reaction. Given C we also define a subset of species, $\Lambda_o \subseteq \Lambda$, as the *output species* of C. Output species are those whose limit distribution is of interest. In general, they may or may not be *non-reacting species*; they depend on the observer and on what he/she is interested in observing.

Definition 8 *A non-reacting output CRS (NRO-CRS) is a tuple $C = (\Lambda, \Lambda_o, R, x_0)$, where $\Lambda_o \subseteq \Lambda$ are the output species of C such that $\Lambda_o \subseteq \Lambda_r$, where Λ_r are the non-reacting species of C.*

NRO-CRNs are CRSs in which the output species are produced monotonically and cannot act as a reactant in any reaction. A consequence of Theorem 1 is the following lemma, which shows that this class of CRNs can approximate any pmf with support on natural numbers, up to an arbitrarily small error.

Lemma 2. *For any probability mass function $f : \mathbb{N}^m \to [0, 1]$ there exists a* NRO-CRS *such that the joint limit distribution of its output species approximates f with arbitrarily small error under the L^1 norm. The approximation is exact if the support of f is finite.*

In Table 1, we define a set of operators on NRO-CRSs. Let $C_1 = (\Lambda_1, \Lambda_{o_1}, R_1, x_{0_1})$ and $C_2 = (\Lambda_2, \Lambda_{o_2}, R_2, x_{0_2})$ be NRO-CRSs such that $\Lambda_1 \cap \Lambda_2 =$

[4] Note that this is a stricter requirement than those in [9], where output species are produced monotonically, but they are allowed to act as catalysts in some reactions. We cannot allow that because catalyst species influence the value of the propensity rate of a reaction and so the probability that it fires.

$\emptyset$. Let $\lambda_{o_1} \in \Lambda_{o_1}$ and $\lambda_{o_2} \in \Lambda_{o_2}$, then we define the set of reactions which implements the operators of Sum, Minimum, Multiplication by a constant $k_1 \in \mathbb{N}$ and Division by a constant $k_2 \in \mathbb{N}_{\geq 0}$ over the steady state distribution of λ_{o_1} and λ_{o_2}. The output species of each composed NRO-CRS is λ_{out}, and we assume $\{\lambda_{out}\} \cap (\Lambda_1 \cup \Lambda_2) = \emptyset$ and $x_0(\lambda_{out}) = 0$.

Table 1: CRS operators

Operator	Resulting NRO-CRS
Sum	$(\Lambda_1 \cup \Lambda_2 \cup \{\lambda_{out}\}, \{\lambda_{out}\}, R_1 \cup R_2 \cup \{\lambda_{o_1} \to \lambda_{out}, \lambda_{o_2} \to \lambda_{out}\}, x_0)$
Min	$(\Lambda_1 \cup \Lambda_2 \cup \{\lambda_{out}\}, \{\lambda_{out}\},\ R_1 \cup R_2 \cup \{\lambda_{o_1} + \lambda_{o_2} \to \lambda_{out}\}, x_0)$
Mul by k_1	$(\Lambda_1 \cup \{\lambda_{out}\}, \{\lambda_{out}\}, R_1 \cup \{\lambda_{o1} \to \underbrace{\lambda_{out} + ... + \lambda_{out}}_{k_1\ times}\}, x_0)$
Div by k_2	$(\Lambda_1 \cup \{\lambda_{out}\}, \{\lambda_{out}\}, R_1 \cup \{\underbrace{\lambda_{o_1} + ... + \lambda_{o_1}}_{k_2\ times} \to \lambda_{out}\}, x_0)$

We emphasize that proving that CRS operators of Table 1 implement the operations in Definition 7 is not trivial. In fact, we need to compose stochastic processes and show that the resulting process has the required properties. Fundamental to that end is a convenient representation of X in terms of a summation of time-inhomogeneous Poisson processes, one for each reaction [2]. In what follows we present in slightly extended form the operators for convex combination, with or without external inputs (respectively $Con(\cdot)$ and $ConE(\cdot)$). Formal definitions and proofs of correctness of all the circuits are presented in [8].
Considering C_1 and C_2, as previously, then we need to derive a CRS operator $Con(C_1, \lambda_{o_1}, C_2, \lambda_{o_2}, p, \lambda_{out})$ such that $\pi_{\lambda_{out}} = (\pi^{C_1}_{\lambda_{o_1}})_p : (\pi^{C_2}_{\lambda_{o_2}})$. That is, at steady stade, λ_{out} equals $\pi^{C_1}_{\lambda_{o_1}}$ with probability p and $\pi^{C_2}_{\lambda_{o_2}}$ with probability $1 - p$. This can be done by using Theorem 2 to generate a bi-dimensional synthetic coin with output species $\lambda_{r_1}, \lambda_{r_2}$ such that their joint limit distribution is $\pi_{\lambda_{r_1},\lambda_{r_2}}(y_1, y_2) = \{p \quad \text{if } y_1 = 1 \text{ and } y_2 = 0\ ; 1 - p \quad \text{if } y_1 = 0 \text{ and } y_2 = 1; 0 \quad \text{otherwise}\}$. That is, λ_{r_1} and λ_{r_2} are mutually exclusive at steady state. Using these species as catalysts in $\tau_3 : \lambda_{o_1} + \lambda_{r_1} \to \lambda_{r_1} + \lambda_{out}$ and $\tau_4 : \lambda_{o_2} + \lambda_{r_2} \to \lambda_{r_2} + \lambda_{out}$ we have exactly the desired result at steady state.

Example 3. Consider the following NRO-CRSs $C_1 = (\{\lambda_{o_1}\}, \{\lambda_{o_1}\}, \{\}, x_{0_1})$ and $C_2 = (\{\lambda_{o_2}\}, \{\lambda_{o_2}\}, \{\}, x_{0_2})$, with initial condition $x_{0_1}(\lambda_{o_1}) = 10$ and $x_{0_2}(\lambda_{o_2}) = 20$. Then, the operator $ConE(C_1, \lambda_{o_1}, C_2, \lambda_{o_2}, 0.3, \lambda_{out})$ implements the operation $\pi_{\lambda_{out}} = (\pi^{C_1}_{\lambda_{o_1}})_{0.3}(\pi^{C_2}_{\lambda_{o_2}})$ and it is given by the following reactions:

$$\lambda_z \to^{0.3} \lambda_{r_1}; \quad \lambda_z \to^{0.7} \lambda_{r_2}; \quad \lambda_{r_1} + \lambda_{o_1} \to \lambda_{r_1} + \lambda_{out}; \quad \lambda_{r_2} + \lambda_{o_2} \to \lambda_{r_2} + \lambda_{out}$$

with initial condition x_0 such that $x_0(\lambda_z) = 1$, $x_0(\lambda_{r_1}) = x_0(\lambda_{r_2}) = x_0(\lambda_{out}) = 0$.

Let C_1, C_2 be as above and $f = p_0 + p_1 \cdot c_1 + ... + p_n \cdot c_n$ with $p_1, ..., p_n \in \mathbb{Q}_{[0,1]}$, $V = \{c_1, ..., c_n\}$ a set of environmental variables, and E, an environment such that $V \subseteq dom(E)$. Then, computing a CRS operator $ConE(C_1, \lambda_{o_1}, C_2, \lambda_{o_2}, f(E(V)), \lambda_{out})$ such that $\pi_{\lambda_{out}} = (\pi^{C_1}_{\lambda_{o_1}})_{f(E(V))} : (\pi^{C_2}_{\lambda_{o_2}})$ is a matter of extending the previous circuit. First of all, we can derive the CRS

to compute $f(E(V))$ and $1-f(E(V))$ and memorize them in some species. This can be done as $f(E(V))$ is semi-linear [9]. Then, as $f(E(V)) \leq 1$ by assumption, we can use these species as catalysts to determine the output value of λ_{out}, as in the previous case. As shown in [8], this circuit, in the case of external inputs, introduces an arbitrarily small, but non-zero, error, due to the fact that there is no way to know when the computation of $f(E(V))$ terminates.

Example 4. Consider the following NRO-CRSs $C_1 = (\{\lambda_{o_1}\}, \{\lambda_{o_1}\}, \{\}, x_{0_1})$ and $C_2 = (\{\lambda_{o_2}\}, \{\lambda_{o_2}\}, \{\}, x_{0_2})$, with initial condition $x_{0_1}(\lambda_{o_1}) = 10$ and $x_{0_2}(\lambda_{o_2}) = 20$. Then, consider the following functions $f(E(c)) = E(c)$, where E is a partial function assigning values to c, and it is assumed $0.001 \leq E(c) \leq 1$ and that $E(c) \cdot 1000 \in \mathbb{N}$. Then, the operator $ConE(C_1, \lambda_{o_1}, C_2, \lambda_{o_2}, f, \lambda_{out})$, implements the operation $\pi_{\lambda_{out}} = (\pi^{C_1}_{\lambda_{o_1}})_{E(c)}(\pi^{C_2}_{\lambda_{o_2}})$ and it is given by the following reactions:

$$\tau_1 : \lambda_c \to^{k_1} \lambda_{\mathrm{Cat}_1} + \lambda_{\mathrm{Cat}_2}; \quad \tau_2 : \lambda_{Tot} + \lambda_{\mathrm{Cat}_2} \to^{k_1} \emptyset$$
$$\tau_3 : \lambda_z + \lambda_{\mathrm{Cat}_1} \to^{k_2} \lambda_1; \quad \tau_4 : \lambda_z + \lambda_{Tot} \to^{k_2} \lambda_2$$
$$\tau_5 : \lambda_{o_1} + \lambda_1 \to^{k_2} \lambda_1 + \lambda_{out}; \quad \tau_6 : \lambda_{o_2} + \lambda_2 \to^{k_2} \lambda_2 + \lambda_{out}$$

where $\lambda_c, \lambda_{\mathrm{Cat}_1}, \lambda_{\mathrm{Cat}_2}, \lambda_z, \lambda_1, \lambda_2$ are auxiliary species with initial condition x_0 such that $x_0(\lambda_{\mathrm{Cat}_1}) = x_0(\lambda_{\mathrm{Cat}_2}) = x_0(\lambda_1) = x_0(\lambda_2) = 0$, $x_0(\lambda_{Tot}) = 1000, x_0(\lambda_z) = 1$, $x_0(\lambda_c) = E(c) \cdot 1000$ and $k_1 \gg k_2$. Reactions τ_1, τ_2 implement $f(E(c))$ and $1 - f(E(c))$ and store these values in λ_{Cat_1} and λ_{Tot}. These are used in reactions τ_3 and τ_4 to determine the probability that the steady state value of λ_{out} is going to be determined by reaction τ_5 or τ_6.

5.2 Compiling into the class of NRO-CRSs

Given a formula P as defined in Definition 5, then $[\![P]\!]_E$ associates to P and an environment E a pmf. We now define a translation of P, $T(P)$, into the class of NRO-CRSs that guarantees that the unique output species of $T(P)$, at steady state, approximates $[\![P]\!]_E$ with arbitrarily small error for any environment E such that $V(P) \subseteq dom(E)$. In order to define such a translation we need the following renaming operator.

Definition 9 *Given a CRS $C = (\Lambda, R, x_0)$, for $\lambda_t \in \Lambda$ and $\lambda_1 \notin \Lambda$ we define the renaming operator $C\{\lambda_1 \leftarrow \lambda_t\} = C_c$ such that $C_c = ((\Lambda - \{\lambda_t\}) \cup \{\lambda_1\}, R\{\lambda_1 \leftarrow \lambda_t\}, x_0')$, where $R\{\lambda_1 \leftarrow \lambda_t\}$ substitutes any occurrence of λ_t with an occurrence of λ_1 for any $\tau \in R$ and $x_0'(\lambda) = \{x_0(\lambda)$ if $\lambda \neq \lambda_t$; $x_0(\lambda_t)$ if $\lambda = \lambda_1\}$.*

This operator produces a new CRS where any occurrence of a species is substituted with an occurrence of another species previously not present.

Definition 10 *(Translation into NRO-CRSs) Define the mapping T by induction on syntax of formulae P:*

$T(one) = (\{\lambda_{out}\}, \{\lambda_{out}\}, \emptyset, x_0)$ *with* $x_0(\lambda_{out}) = 1$;
$T(zero) = (\{\lambda_{out}\}, \{\lambda_{out}\}, \emptyset, x_0)$ *with* $x_0(\lambda_{out}) = 0$;
$T(P_1 + P_2) = Sum(T(P_1)\{\lambda_{o_1} \leftarrow \lambda_{out}\}, \lambda_{o_1}, T(P_2)\{\lambda_{o_2} \leftarrow \lambda_{out}\}, \lambda_{o_2}, \lambda_{out})$;
$T(k \cdot P) = Div(Mul(T(P)\{\lambda_o \leftarrow \lambda_{out}\}, \lambda_o, k_1, \lambda_{out})\{\lambda_{o'} \leftarrow \lambda_{out}\}), \lambda_{o'}, k_2, \lambda_{out})$;
$T(min(P_1, P_2) = Min(T(P_1)\{\lambda_{o_1} \leftarrow \lambda_{out}\}, \lambda_{o_1}, T(P_2)\{\lambda_{o_2} \leftarrow \lambda_{out}\}, \lambda_{o_2}, \lambda_{out})$;
$T((P_1)_D : P_2) =$
$$\begin{cases} Con(T(P_1)\{\lambda_{o_1} \leftarrow \lambda_{out}\}, \lambda_{o_1}, T(P_2)\{\lambda_{o_2} \leftarrow \lambda_{out}\}, \lambda_{o_2}, D, \lambda_{out}), & \text{if } D = p \\ ConE(T(P_1)\{\lambda_{o_1} \leftarrow \lambda_{out}\}, \lambda_{o_1}, T(P_2)\{\lambda_{o_2} \leftarrow \lambda_{out}\}, \lambda_{o_2}, D, \lambda_{out}), & \\ \qquad\qquad \text{if } D = p + \sum_{i=1}^{m} p_i \cdot c_i \end{cases}$$
for $m > 1$, $k \in \mathbb{Q}_{>0}$, $k_1, k_2 \in \mathbb{N}$ *such that* $k = \frac{k_1}{k_2}$ *and formulae* P_1, P_2, *which are assumed to not contain species* $\lambda_{o_1}, \lambda_{o_2}$.

Example 5. Consider the formula $P_1 = (one)_{0.001 \cdot c + 0.2}(4 \cdot one) + (2 \cdot one)_{0.4}(3 \cdot one)$ of Example 2, and an environment E such that $0.000125 \leq E(c) \leq 1$ and suppose $E(c) \cdot 800 \in \mathbb{N}$. We show how the translation defined in Definition 10 produces a NRO-CRS C with output species λ_{out} such that $\pi_{\lambda_{out}} = [\![P_1]\!]_E$. Consider the following NRO-CRSs C_1, C_2, C_3, C_4 defined as $C_1 = (\{\lambda_{c_1}\}, \{\lambda_{c_1}\}, \{\}, x_0')$ with $x_0(\lambda_{c_1}) = 1$, $C_2 = (\{\lambda_{c_2}\}, \{\lambda_{c_2}\}, \{\}, x_0)$ with $x_0(\lambda_{c_2}) = 1$, , $C_3 = (\{\lambda_{c_3}\}, \{\lambda_{c_3}\}, \{\}, x_0)$ with $x_0(\lambda_{c_3}) = 1$, and $C_4 = (\{\lambda_{c_4}\}, \{\lambda_{c_4}\}, \{\}, x_0)$ with $x_0(\lambda_{c_2}) = 1$. Then, we have that :

$$C_1^c = ConE(C_1, \lambda_{c_1}, Mul(C_2, \lambda_{c_2}, 4, \lambda_{out})\{\lambda_{o_2} \leftarrow \lambda_{out}\}, \lambda_{o_2}, 0.001 \cdot c + 0.2, \lambda_{out_1})$$
$$C_2^c = Con(Mul(C_3, \lambda_{c_3}, 2, \lambda_{out})\{\lambda_{o_3} \leftarrow \lambda_{out}\}, \lambda_{o_3}, Mul(C_4, \lambda_{c_4}, 3, \lambda_{out})\{\lambda_{o_4} \leftarrow \lambda_{out}\}, \lambda_{o_4}, 0.4, \lambda_{out_2})$$

are such that $\pi_{\lambda_{out_1}} = \begin{cases} (0.001 \cdot [\![c]\!]_E + 0.2), & \text{if } y = 1 \\ 1 - (0.001 \cdot [\![c]\!]_E + 0.2), & \text{if } y = 4 \\ 0, & \text{otherwise} \end{cases}$,
and $\pi_{\lambda_{out_2}} = \begin{cases} 0.4, & \text{if } y = 2 \\ 0.6, & \text{if } y = 3 \\ 0, & \text{otherwise} \end{cases}$. Then, consider the CRS $C = Sum(C_1^c\{\lambda_{t_1 \leftarrow \lambda_{out_1}}\}, \lambda_{t_1}, C_2^c\{\lambda_{t_2 \leftarrow \lambda_{out_2}}\}, \lambda_{t_2}, \lambda_{out})$ and we have $\pi_{\lambda_{out}} = [\![P_1]\!]_E$ with arbitrarily small error. The reactions of C are shown below

$$Mul \text{ on inputs } \{\tau_1 : \lambda_{C_2} \to 4\lambda_{o_1}; \quad \tau_2 : \lambda_{C_3} \to 2\lambda_{o_2}; \quad \tau_3 : \lambda_{C_4} \to 3\lambda_{o_3}$$
$$C_1^c \begin{cases} \tau_4 : \lambda_{env} \to^k \lambda_{cat_1} + \lambda_{cat_2}; \quad \tau_5 : \lambda_{cat_1} + \lambda_z \to \lambda_1 \\ \tau_6 : \lambda_{cat_2} + \lambda_{tot} \to^k \emptyset; \quad \tau_7 : \lambda_{tot} + \lambda_z \to \lambda_2 \\ \tau_8 : \lambda_1 + \lambda_{o_1} \to \lambda_{o_1} + \lambda_{out_1}; \quad \tau_9 : \lambda_2 + \lambda_{o_2} \to \lambda_{o_2} + \lambda_{out_1} \end{cases}$$
$$C_2^c \begin{cases} \tau_9 : \lambda_{z_1} \to^{0.6} \lambda_{r_1}; \quad \tau_{10} : \lambda_{z_1} \to^{0.4} \lambda_{r_2} \\ \tau_{11} : \lambda_{r_1} + \lambda_{o_3}; \to \lambda_{r_1} + \lambda_{out_2}; \quad \tau_7 : \lambda_{r_2} + \lambda_{o_4} \to \lambda_{r_2} + \lambda_{out_2} \end{cases}$$
$$Sum \, \{\tau_{12} : \lambda_{out_1} \to \lambda_{out}; \quad \tau_{13} : \lambda_{out_2} \to \lambda_{out}$$

for $k \gg 1$ and initial condition such that $x_0(\lambda_{env}) = E(c) \cdot 800$, $x_0(\lambda_{tot}) = 800$, $x_0(\lambda_z) = x_0(\lambda_{z_1}) = x_0(\lambda_{z_2}) = 1 = x_0(\lambda_{c_1}) = x_0(\lambda_{c_2}) = x_0(\lambda_{c_3}) = x_0(\lambda_{c_4}) = 1$, and all other species initialized with 0 molecules.

Proposition 2. *For any formula P we have that $T(P)$ is a NRO-CRS.*

The proof follows by structural induction as shown in [8]. Given a formula P and an environment E such that $V(P) \subseteq dom(E)$, the following theorem guarantees the soundness of $T(P)$ with respect to $[\![P]\!]_E$. In order to prove the soundness of our translation we consider the measure of the multiplicative error between two pmfs f_1 and f_2 with values in $\mathbb{N}^m$, $m > 0$ as $e_m(f_1, f_2) = \max_{n \in \mathbb{N}^m} \min(\frac{f_1(n)}{f_2(n)}, \frac{f_2(n)}{f_1(n)})$.

Theorem 5. *(Soundness) Given a formula P and λ_{out}, unique output species of $T(P)$, then, for an environment E such that $V(P) \subseteq dom(E)$, it holds that $\pi^{T(P)}_{\lambda_{out}} = [\![P]\!]_E$ with arbitrarily small error under multiplicative error measure.*

The proof follows by structural induction.

Remark 3. A formula P is finite by definition, so Theorem 5 is valid because the only production rule which can introduce an error is $(P_1)_D : (P_2)$ in the case $D \neq p_0$, and we can always find reaction rates to make the total probability of error arbitrarily small. Note that, by using the results of [17], it would also be possible to show that the total error can be kept arbitrarily small, even if a formula is composed from an unbounded number of production rules. This requires small modifications to the ConE operator following ideas in [17].

Note that compositional translation, as defined in Definition 10, generally produces more compact CRNs respect to the direct translation in Theorem 1, and in both cases the output is non-reacting, so the resulting CRN can be used for composition. For a distribution with support J direct translation yields a CRN with $2|J|$ reactions, whereas, for instance, the support of the sum pmf has the cardinality of the Cartesian product of the supports of the input pmfs.

6 Discussion

Our goal was to explore the capacity of CRNs to compute with distributions. This is an important goal because, when molecular interactions are in low number, as is common in various experimental scenarios [15], deterministic methods are not accurate, and stochasticity is essential for cellular circuits. Moreover, there is a large body of literature in biology where stochasticity has been shown to be essential and not only a nuisance [11]. Our work is a step forward towards better understanding of molecular computation. In this paper we focused on error-free computation for distributions. It would be interesting to understand and characterize what would happen when relaxing this constraint. That is, if we admit a probabilistically (arbitrarily) small error, does the ability of CRNs to compute on distributions increase? Can we relax the constraint that output species need to be produced monotonically? Can we produce more compact networks? Another topic we would like to address is if it is possible to implement the CRNs without leaders (species being present with initial number of molecules equal to 1). This is a crucial aspect in our theorems and having the same results without these constraints would make the implementation easier. However, it is worth noting that, in a practical scenario, such species could be thought of as a single gene or as localized structures [15].

References

1. D. F. Anderson, G. Craciun, and T. G. Kurtz. Product-form stationary distributions for deficiency zero chemical reaction networks. *Bulletin of mathematical biology*, 72(8):1947–1970, 2010.
2. D. F. Anderson and T. G. Kurtz. Continuous time Markov chain models for chemical reaction networks. In *Design and analysis of biomolecular circuits*, pages 3–42. Springer, 2011.
3. J. C. Anderson, E. J. Clarke, A. P. Arkin, and C. A. Voigt. Environmentally controlled invasion of cancer cells by engineered bacteria. *Journal of molecular biology*, 355(4):619–627, 2006.
4. D. Angluin, J. Aspnes, D. Eisenstat, and E. Ruppert. The computational power of population protocols. *Distributed Computing*, 20(4):279–304, 2007.
5. A. Arkin, J. Ross, and H. H. McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected escherichia coli cells. *Genetics*, 149(4):1633–1648, 1998.
6. L. Cardelli and A. Csikász-Nagy. The cell cycle switch computes approximate majority. *Scientific reports*, 2, 2012.
7. L. Cardelli, M. Kwiatkowska, and L. Laurenti. Stochastic analysis of chemical reaction networks using linear noise approximation. In *Computational Methods in Systems Biology*, pages 64–76. Springer, 2015.
8. L. Cardelli, M. Kwiatkowska, and L. Laurenti. Programming discrete distributions with chemical reaction networks. *arXiv preprint arXiv:1601.02578*, 2016.
9. H.-L. Chen, D. Doty, and D. Soloveichik. Deterministic function computation with chemical reaction networks. *Natural computing*, 13(4):517–534, 2014.
10. Y.-J. Chen, N. Dalchau, N. Srinivas, A. Phillips, L. Cardelli, D. Soloveichik, and G. Seelig. Programmable chemical controllers made from DNA. *Nature nanotechnology*, 8(10):755–762, 2013.
11. A. Eldar and M. B. Elowitz. Functional roles for noise in genetic circuits. *Nature*, 467(7312):167–173, 2010.
12. B. Fett, J. Bruck, and M. D. Riedel. Synthesizing stochasticity in biochemical systems. In *Design Automation Conference, 2007. DAC'07. 44th ACM/IEEE*, pages 640–645. IEEE, 2007.
13. M. Kwiatkowska, G. Norman, and D. Parker. Stochastic model checking. In *Formal methods for performance evaluation*, pages 220–270. Springer, 2007.
14. R. Losick and C. Desplan. Stochasticity and cell fate. *science*, 320(5872):65–68, 2008.
15. L. Qian and E. Winfree. Parallel and scalable computation and spatial dynamics with DNA-based chemical reaction networks on a surface. In *DNA Computing and Molecular Programming*, pages 114–131. Springer, 2014.
16. J. M. Schmiedel, S. L. Klemm, Y. Zheng, A. Sahay, N. Blüthgen, D. S. Marks, and A. van Oudenaarden. MicroRNA control of protein expression noise. *Science*, 348(6230):128–132, 2015.
17. D. Soloveichik, M. Cook, E. Winfree, and J. Bruck. Computation with finite stochastic chemical reaction networks. *natural computing*, 7(4):615–633, 2008.
18. D. Soloveichik, G. Seelig, and E. Winfree. DNA as a universal substrate for chemical kinetics. *Proceedings of the National Academy of Sciences*, 107(12):5393–5398, 2010.