TAIL APPROXIMATION FOR THE CHEMICAL MASTER EQUATION

Thomas A. Henzinger[1] *and Maria Mateescu*[2*]

[1]IST Austria, Klosterneuburg, Austria
[2]School of Computer and Communication Sciences, EPFL, Switzerland
tah@ist.ac.at, maria.mateescu@epfl.ch

ABSTRACT

The chemical master equation is a differential equation describing the time evolution of the probability distribution over the possible "states" of a biochemical system. The solution of this equation is of interest within the systems biology field ever since the importance of the molecular noise has been acknowledged. Unfortunately, most of the systems do not have analytical solutions, and numerical solutions suffer from the course of dimensionality and therefore need to be approximated. Here, we introduce the concept of tail approximation, which retrieves an approximation of the probabilities in the tail of a distribution from the total probability of the tail and its conditional expectation. This approximation method can then be used to numerically compute the solution of the chemical master equation on a subset of the state space, thus fighting the explosion of the state space, for which this problem is renowned.

1. INTRODUCTION

Models of biochemical reaction networks have traditionally been studied by solving a differential equation called the reaction rate equation. The reaction rate equation approximates the mean behaviour of the network under study and is usually easy to solve through numerical integration. However, in the past decade, it has repeatedly been shown that stochasticity is playing an important role in biological systems, and thus, that the mean behaviour of a system is not providing enough information to understand biological mechanisms [1, 2]. As a consequence, the interest has moved towards the chemical master equation (CME) whose solution gives the probability for a biochemical reaction network to be in a certain state at a certain time. Unfortunately, for most networks of interest, this equation is very hard to solve, due to the large number of states with a strictly positive probability. An important aspect is that stochastic effects appear especially when some molecular species are present in low copy numbers, with significant probability, while for species that are always present in a large copy number their expectation is providing enough information about the system. An approximation method may take advantage of this fact by solving the chemical master equation only for the subspace of the system that corresponds to species in low copy numbers, while using the reaction rate equation for the rest of the space to compute conditional expectations. Connecting the two equations in an accurate manner would be technically difficult because the CME needs more information about the states with large copy numbers than the reaction rate equation is providing. The tail approximation is addressing this issue by approximating the probabilities of the states governed by the reaction rate equation, that are at the boundary with the region governed by the CME. This approximation proceeds stepwise. A first approximation assumes that the tail of the probability distribution has the shape of a geometric distribution, and then several correction steps are applied iteratively. These correction steps take into account the probabilities of the states governed by the CME. After presenting the tail approximation method, we prove its accuracy by distorting via aggregation and then restoring via tail approximation the actual solutions of the CME of two different biological models.

Related work The algorithm that we suggest and that motivates the need for tail approximation as defined here, is a refined version of the hybrid method [3]. In the context of the chemical master equation, interpolation has been already used either in the form of regular polynomial interpolation [4] or using Poisson probabilities [5]. Here, we use an approximation by the geometric distribution, and the resemblance with interpolation comes from the correction step where we use the values of the probabilities in the neighbouring states.

2. CHEMICAL MASTER EQUATION

Consider a biochemical reaction network with n components. This system has as state space $\mathcal{S} = \mathbb{N}^n_{\geq 0}$, where each state in $\mathcal{S}$ is of the form $s = (s_1, \ldots, s_n)$ and gives the number of copies of each component i, denoted by s_i, with $i \in \{1, \ldots, n\}$.

Furthermore, consider that there are m reactions possible in this system and that each reaction R_j, with $j = 1, \ldots, m$, is described by the *propensity function* $\alpha_j : \mathcal{S} \to \mathbb{R}_{\geq 0}$ and the *change vector* $v_j \in \mathbb{Z}^n$.

For such a system we define the Markovian stochastic process $\{X(t), t \geq 0\}$, for which:

$$Pr(X(t + dt) = s + v_j \mid X(t) = s) = \alpha_j(s) \cdot dt,$$

for an infinitesimal dt.

*Authors in alphabetical order.

Let $y \in \mathcal{S}$ be the initial state of our model. Then, we define the probability to be in state s at time t as:

$$p^{(t)}(s) = Pr(X(t) = s \mid X(0) = y).$$

The chemical master equation (CME) is a differential equation describing the time evolution of $p^{(t)}(s)$ [6]:

$$\frac{dp^{(t)}}{dt}(s') = \sum_{R_j, s+v_j=s'} p^{(t)}(s)\alpha_j(s) - \sum_{R_j} p^{(t)}(s')\alpha_j(s'),$$

where the first term handles the probabilities that enter s' from various predecessors s, through reaction R_j, and the second one, which is negative, handles the probabilities that exit s' towards various successors.

2.1. Aggregated Solution

As already discussed, stochasticity has an important effect when some of the species are present in low copy numbers. In consistence with this observation and in order to develop fast algorithms, we propose the concept of an *aggregated solution* of the CME.

For a boundary $b \in \mathbb{N}$ let $\hat{\mathcal{S}} = \{0, \ldots, b-1, \top\}^n$, where the value $\top$ represents the set $\{x \in \mathbb{N} | x \geq b\}$. Given a vector $A \in \hat{\mathcal{S}}$ we define $\mathcal{S}_A$ to be a subset of $\mathcal{S}$:

$$\mathcal{S}_A = \{s \in \mathcal{S} \mid s_i = A_i \text{ or } (s_i \geq b \text{ and } A_i = \top)\}.$$

Example 1. *For a two dimensional system, with $b = 3$, the set $\hat{\mathcal{S}}$ is partitioning $\mathbb{N}^2_{\geq 0}$ as shown in Figure 1. The aggregated state $(1, \top)$ represents the subset $\{(1, s) \in \mathcal{S} \mid s \geq 3\}$.*

Definition 1. *The* aggregated solution *of a CME with solution $p^{(t)}$, with respect to the boundary b, is a tuple $\langle \hat{p}^{(t)} : \hat{\mathcal{S}} \to [0,1], \mu^{(t)} : \hat{\mathcal{S}} \to \mathbb{R}^n \rangle$, such that*

$$\hat{p}^{(t)}(A) = \sum_{s \in \mathcal{S}_A} p^{(t)}(s),$$

and

$$\mu^{(t)}(A) = \sum_{s \in \mathcal{S}_A} p^{(t)}(s) \cdot s.$$

We also define the *boundary set* $\mathcal{S}_b$ of an aggregation with boundary b to be:

$$\mathcal{S}_b = \{s \in \mathcal{S} \mid \exists i \text{ s.t. } s_i = b\}.$$

In order to obtain an aggregated solution that is exact we must first compute the probabilities $p^{(t)}(s)$ and then aggregate them. However, if we are only interested in an approximation of the aggregated solution, a direct method might be possible, and that would guarantee a large saving in the computation time due to the state space reduction.

Such a direct method would need to estimate:

1. the probabilities of the states in the boundary set $\mathcal{S}_b$, in order to know how much probability to pass from one macro-state to another one via reactions R_j,

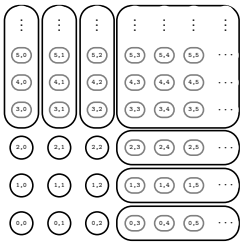

Figure 1. Partitioning of the state space used for an aggregated solution.

2. the conditional expectations:

$$E[X(t) | X(t) \in \mathcal{S}_A],$$

in order to compute the function μ of the aggregated solution.

In this paper, we continue by solving the first of these two points. The second point, computing the conditional expectations, is left as future work and can most likely be solved using a hybrid method [3].

3. TAIL APPROXIMATION

The key observation behind the tail approximation is that almost all probability distributions that describe the stochastic behaviour of real life have a certain "continuity" property and that their tail matches the shape of a geometric distribution. We do not formalize these properties here, but we refer to such real life distributions as "well-behaved", in order to distinguish them from other, possibly random, distributions.

Problem 1 (Tail approximation)**.** *For an aggregated solution $\langle \hat{p}, \mu \rangle$ with boundary $b \in \mathbb{N}$, of an unknown probability distribution p, approximate the probabilities $p(s)$ of the states s that belong to the boundary set $\mathcal{S}_b$.*

We first solve the tail approximation problem for systems of dimension one, $\mathcal{S} = \mathbb{N}_{\geq 0}$

The tail approximation is done in two stages. First, we make a coarse approximation using a shifted geometric distribution and then we iteratively correct this value with respect to the probability values $\hat{p}(x)$ with $x < b$.

Recall that the geometric distribution with mean M is a discrete probability distribution defined as:

$$g_M(x) = p \cdot (1-p)^x, \text{ where } p = \frac{1}{M+1}.$$

For the first stage of our approximation, we define the *shifted geometric distribution* $\bar{g}_{M,b} : \mathbb{N}_{\geq b} \to [0,1]$ to be: $\bar{g}_{M,b}(x) = g_{M-b}(x-b)$. This function is used to roughly approximate $p(b)$ by $\tilde{p}(b) = \hat{p}(\top) \cdot \bar{g}_{\mu(\top),b}(b)$, where $\hat{p}(\top)$ is the probability for $x > b$, and $g_{\mu(\top),b}$ estimates the probability of being in state b conditioned on $x \geq b$, from the aggregated expectation $\mu(\top)$.

The correction stage of our approximation is based on the observation that for a well-behaved probability distribution the relative errors of our approximation at points b

and $b-1$ are almost equal:

$$\frac{\tilde{p}(b)}{p(b)} \approx \frac{\tilde{p}(b-1)}{p(b-1)}.$$

Therefore, from the above approximation and because $\hat{p}(b-1) = p(b-1)$, we obtain the first corrected approximation $\tilde{p}_{c_1}$:

$$p(b) \approx \tilde{p}_{c_1}(b) = \tilde{p}(b) \cdot \frac{\hat{p}(b-1)}{\tilde{p}(b-1)}.$$

This new value of the approximation can be updated in a second correction step in which the value $\hat{p}(b-2)$ is taken into account. In Section 4 we present results for up to three correction steps.

Let us extend the vector $\hat{p}$ with $\hat{p}(-1) = 0$, and let z be the largest state for which $z < b$ and $p(z) = 0$. The correction stage of our approximation can have up to $b - z - 1$ steps.

3.1. Multiple dimensions

Here we are interested in the approximation of the value $p(s)$ with $s \in \mathbb{N}^n_{\geq 0}$. First, we define the one dimensional projection $p_{|i,s} : \mathbb{N}_{\geq 0} \to [0,1]$ to be a sub-stochastic probability distribution such that $p_{|i,s}(x)$ is the probability for the i-th component to have value x and all the other components $i' \neq i$ to have the values $s_{i'}$. Formally, let $s_{|i,x}$ denote the vector s in which the i-th component has been set to x. Then, the probability projection is defined as $p_{|i,s}(x) = p(s_{|i,x})$.

For a state s with only one component i for which $s_i = \top$ we apply the 1-dimension tail approximation method on the projection $\langle \hat{p}_{|i,s}, \mu_i \rangle$.[1]

Finally, if we have more than one dimensions for which $s_i = \top$, we make an independence assumption in order to reduce the problem to one dimension. For 2-dimensions, this assumption is:

$$p^{(t)}(b,b) \approx \hat{p}^{(t)}(b,\top) \cdot \hat{p}^{(t)}(\top,b).$$

4. CASE STUDIES

In this section we give statistical evidence that empirically proves the level of accuracy of our approximation. For this, we start with the actual solution of the CME of two different biochemical reaction networks, from which we compute the aggregated solution with respect to a boundary b (step in which we loose information). After that, we use the tail approximation in order to restore the probabilities of the boundary states $\mathcal{S}_b$ from the aggregated solution. Finally, we compare the restored probabilities with those in the probability distribution we have started with.

We consider two systems: the predator-prey [7] and the exclusive switch [8]. First, we compute the solution of the CME associated to each of these two systems (at time points $t = 1, 2, \ldots, 5$ for predator-prey, and $t = 10, 20, \ldots, 100$ for the exclusive switch) using our previous algorithm [9] and tool [10]. For each of the obtained probability distributions $p^{(t)}$, for each of the boundaries $b \in \{10, 50, 100\}$, and for all states s in the boundary set $\mathcal{S}_b$, we consider all projections $p_{|i,s}$ of the solution $p^{(t)}$. It is these projections that we first aggregate and then restore using tail approximation.

[1]The tail approximation can be applied on sub-stochastic vectors as well.

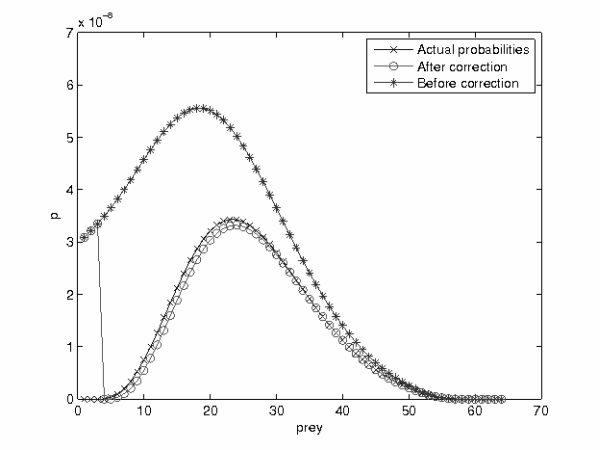

Figure 2. For all boundaries $b \in \{1, \ldots, 64\}$, we show the comparison between the actual probability vector $p(b)$, the first approximation $\tilde{p}(b)$ and the corrected approximation $\tilde{p}_{c_1}(b)$. The probability vector p is taken from the solution at time $t = 3$ of a predator-prey system and it gives the probabilities over the number of preys for a fixed number of 46 predators.

For three correction steps $k = 1, 2, 3$, the column **max. abs. err.** of Table 1 gives the maximal absolute error computed as

$$\max_{s \in \mathcal{S}_b} \Big(\mid p_{c_k}(s) - p(s) \mid \Big).$$

The column **rel. err.** gives the relative error for the state with maximal absolute error. The relative error is computed as:

$$1 - \min\Big(\frac{\tilde{p}_{c_k}(s)}{p(s)}, \frac{p(s)}{\tilde{p}_{c_k}(s)}\Big).$$

Finally, Table 1 also reports the percentage of projections for which the absolute error, as defined above, is greater than 10^{-7}. In some cases this percentage is high because the chosen boundary is too small.

In Figure 2 we show how the tail approximation performs at all possible boundaries b of a predator-prey model for the projection $p_{|\text{prey,predator=46}}$. For this figure, the original probability distribution has been aggregated with respect to all possible boundaries b for the number of preys. For small boundaries, the region of the probability distribution at the right of the boundary does not have a geometric shape and thus the errors are larger. Even more, for very small boundaries, the correction step can not be applied because the value of p is 0, and the approximation is completely unacceptable. This is one of the major problems that an algorithm using this approximation needs to

Table 1. Results.

Model	b	Max abs. err.			rel. err.			abs. err.$> 1\times10^{-7}$		
		1st corr.	2nd corr.	3rd corr.	1st corr.	2nd corr.	3rd corr.	1st corr.	2nd corr.	3rd corr.
pred. prey	10	2×10^{-3}	1×10^{-3}	5×10^{-4}	8×10^{-2}	3×10^{-2}	1×10^{-2}	46%	46%	43%
	50	2×10^{-4}	2×10^{-5}	1×10^{-6}	2×10^{-1}	2×10^{-2}	9×10^{-4}	4%	4%	3%
	100	4×10^{-8}	1×10^{-9}	6×10^{-11}	1×10^{-3}	6×10^{-5}	2×10^{-6}	0%	0%	0%
ex. switch	10	7×10^{-3}	1×10^{-2}	1×10^{-3}	4×10^{-1}	4×10^{-1}	1×10^{-1}	92%	95%	94%
	50	1×10^{-3}	1×10^{-4}	2×10^{-5}	1×10^{-1}	1×10^{-2}	1×10^{-3}	46%	46%	12%
	100	3×10^{-4}	3×10^{-5}	3×10^{-6}	4×10^{-2}	4×10^{-3}	4×10^{-4}	33%	33%	14%

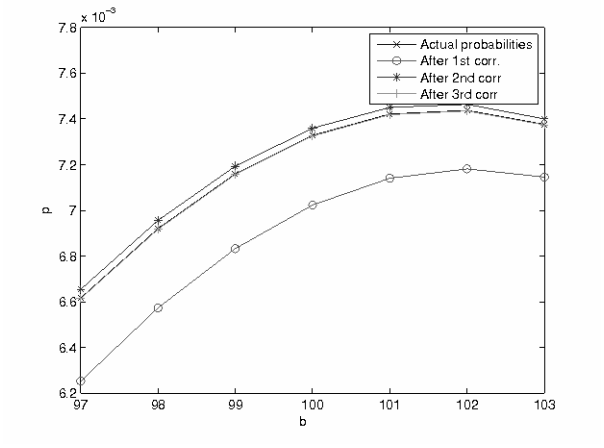

Figure 3. Approximation of state probabilities in the exclusive switch model. With each correction step the approximation is closer to the real probabilities.

solve. However, for larger boundaries, the approximation is very accurate and the correction step is performing well.

Figure 3 illustrates the case in which more than one correction steps are needed in order to obtain an accurate result. As future work, we hope to design a fix-point algorithm that would detect how many correction steps are necessary for a given tolerance.

5. CONCLUSION

We have presented a way to approximate probabilities of a biochemical reaction network at the boundary between low and large copy numbers. The proposed approximation is simple but we have proved that the correction step of the approximation is powerful in obtaining an accurate value. Future work will include developing an algorithm that uses the tail approximation in order to reduce the state space of the CME, proving the convergence of the correction steps and ensuring an error bound of the tail approximation.

6. ACKNOWLEDGMENTS

We thank Brian Munsky, Mathieu Tracol and Verena Wolf for helpful discussions, and Ghid Maatouk for reading previous drafts of this paper.

7. REFERENCES

[1] N. Fedoroff and W. Fontana, "Small numbers of big molecules," *Science*, vol. 297, pp. 1129–1131, 2002.

[2] J. Paulsson, "Summing up the noise in gene networks.," *Nature*, vol. 427, no. 6973, pp. 415–418, 2004.

[3] T. A. Henzinger, L. Mikeev, M. Mateescu, and V. Wolf, "Hybrid numerical solution of the chemical master equation," in *Computational Methods in Systems Biology*, 2010, pp. 55–65.

[4] L. Ferm and P. Ltstedt, "Adaptive solution of the master equation in low dimensions," in *Applied Numerical Mathematics*, 2007.

[5] B. Munsky, *The Finite State Projection Approach for the Solution of the Master Equation and its Applications to Stochastic Gene Regulatory Networks*, Ph.D. thesis, University of California, Santa Barbara, 2008.

[6] D. T. Gillespie, "A rigorous derivation of the chemical master equation," *Physica A*, vol. 188, pp. 404–425, 1992.

[7] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions.," *Journal of Physical Chemistry*, vol. 81, no. 25, pp. 2340–2361, 1977.

[8] A. Loinger, A. Lipshtat, N. Q. Balaban, and O. Biham, "Stochastic simulations of genetic switch systems," *Physical Review E*, vol. 75, no. 2, pp. 021904, 2007.

[9] F. Didier, T. A. Henzinger, M. Mateescu, and V. Wolf, "Fast adaptive uniformization of the chemical master equation," in *Proceedings of the 2009 International Workshop on High Performance Computational Systems Biology*, 2009, pp. 118–127.

[10] F. Didier, T. A. Henzinger, M. Mateescu, and V. Wolf, "Sabre: A tool for stochastic analysis of biochemical reaction networks," in *Proceedings of the 2010 Seventh International Conference on the Quantitative Evaluation of Systems*, 2010, pp. 193–194.