PARTICLE FILTERING AND THE INVERSE PROBLEM OF BIOCHEMICAL NETWORKS

Petar M. Djurić and Mónica F. Bugallo

Department of Electrical and Computer Engineering
Stony Brook University, Stony Brook, NY 11794 (USA)
phone: +1 631 632 8423, fax: +1 631 632 8494, email: {djuric, monica}@ece.sunysb.edu

ABSTRACT

In this paper we address the inverse problem of biochemical networks composed of first-order reactions. Based on noisy measurements of the number of molecules of some of the species that participate in the reactions of a given network, we propose a method for estimating recursively the numbers of molecules and the stochastic rate constants in the network. The evolution of the number of molecules of the species is modeled by Poisson random processes, and the observations are assumed to be (non)linear functions of the number of molecules. Our method employs particle filtering where we propose particles of the stochastic rate constants from their posteriors. We demonstrate the performance of the proposed method with several examples.

1. INTRODUCTION AND BACKGROUND

Inter- and intra-cellular biochemical processes are very complex in nature, and their accurate modeling is a very challenging task [1], [2], [3]. They can be studied using system theory where, in general, the systems are represented by a network of coupled chemical reactions. These networks can reproduce genetic networks, signal transduction networks, and metabolic pathways describing processes such as transcription, gene repression, translation, degradation, transport and auto-regulation. From a set of simple biological systems one can construct much more complex systems which contain feedbacks and nonlinearities and can model events on a broader scale [1], [4].

Most of the contemporary computational approaches to studying biological systems are deterministic in nature. They represent biochemical reactions by differential equations which are numerically solved. However, it is well known that one can improve on the modeling of the dynamics of biological systems by using the laws of probability theory. This is especially true when the numbers of molecules of some species in the reactions are small. In fact, in important scenarios of signal transduction and gene expression, the deterministic methods can produce misleading results, and the stochastic approaches are the only ones for studying them [5], [6], [7]. Examples include the lambda phage switch [5], the observed individual differences among genetically identical bacteria due to random gene expression [8], and the presence of multistability of an observed system due to noise random dynamic switching between stationary states [9]. The relationship between the deterministic and stochastic approaches has been addressed in [10]. In this paper, we work on a method that belongs to the stochastic approaches.

There are two general classes of problems related to biochemical networks and they are known as the forward and inverse types of problems. In a forward problem the objective is to simulate the evolution of a biochemical network given its description, the initial number of molecules of the various species, and all the stochastic rate constants. For simple networks this is a rather easy task, and to that end one can use, for example, the method from [11]. The inverse problem is about estimating unknowns in the system (the biochemical network) from observations that are usually number of molecules of some of the species in the network. This can be quite a challenging task. Some methods that have been used for this purpose are Monte Carlo-based and include Markov chain Monte Carlo sampling [12].

In this paper we present results on the inverse problem of biochemical networks with first-order reactions. We model the number of individual reactions in a given interval $\Delta t$ by Poisson processes. The parameters of these processes are functions of the stochastic rate constants of the reactions that are assumed unknown. We have measurements of some of the species in the system and we want to use them to estimate the time-varying numbers of molecules of the species participating in the reactions.\footnote{1We note here that the technology for routinely obtaining such measurements does not exist yet.} In addition, we also do not know the values of the stochastic rate constants of the reactions in the system.

We apply particle filtering (PF) for their recursive estimation [13], [14]. In particular, we are interested in the joint posterior of the unknowns, which is approximated by a random measure composed of particles that represent the unknowns on the system. As priors for the vector of stochastic constant rates, we use a product of Gamma probability distributions. We point out that the nature of the problem does not require real time processing of the data. This means that one may employ, for instance, smoothing in order to improve the estimates of some of the unknowns.

The paper is organized as follows. In the next section we present the mathematical formulation of the problem. In Section 3, we describe the proposed solution and show its implementation on a specific example. Section 4 contains

\footnote{This work has been supported by the National Science Foundation under Awards CCR-0220011 and CCF-0515246 and the Office of Naval Research under Award N00014-06-1-0012.}
computer simulations that show the performance of the proposed method. With Section 5, we conclude the paper with some final remarks.

2. MATHEMATICAL FORMULATION OF THE PROBLEM

We observe a set of noisy measurements that represent functions of the number of molecules of species from a biochemical network of interest. Here we only consider a network composed of first-order reactions. We proceed by way of extensions of the number of molecules of species from a biochemical network. We may know the initial value of the number of molecules of the species $X$ and the state vectors $X(t)$ or $X(t)$ after the state vector $X(t)$ will be a molecule of species $X$ or of species $X$. The constants $c_1$, $c_2$, and $c_3$ are the stochastic rate constants of the network [12].

We represent the state of the system at time instant $t$ by the state vector $x_t = [x_1(t), x_2(t), x_3(t)]^T$, where each element of the vector denotes the number of molecules of the respective species. With time, the state vector evolves in a random way. In particular, if the actual time interval between the time instants $t$ and $t+1$ is $\Delta t$, then we define the probabilities that a molecule of species $X_1$ or $X_2$ after $\Delta t$ by $p_{11}$ and $p_{12}$. Similarly, we define the probabilities that a molecule of species $X_2$, $X_3$, or of species $X_3$ after $\Delta t$ by $p_{21}$, $p_{22}$, and $p_{23}$. Clearly, we also have $p_{33} = 1$. Each of these probabilities are modeled as functions of the relevant stochastic rate constants.

In the above biochemical network, we assume that the unknowns are the stochastic rate constants $c = [c_{12}, c_{21}, c_{23}]^T$ and the state vectors $x_0$, $y_t$, where $T+1$ is the total number of available measurements. We may know the initial value of the state vector $x_0$ or we have some measurements about it from which we can construct a prior given by $p(x_0)$. For the stochastic rate constants we also have a prior denoted by $p(c)$.

The noisy measurements are modeled by

$$y_t = g(x_t) + v_t$$

where $g(\cdot)$ is some known function and $v_t$ is observation noise with a known probability distribution. The objective is to use the measurements $y_0$, and the model of the system described by the probability distributions $p(x_t|x_{t-1}, c_{t-1})$ and $p(y_t|x_t)$ and estimate all the unknowns in the system.

A good model for approximating the transitions of $x_t$ is the Poisson model. According to this model, in our example the number of molecules of the species $X_1$ that in $\Delta t$ become molecules of species $X_2$ is given by

$$p(\Delta x_{12}, t) = \frac{\lambda_{12} \Delta t}{\Delta x_{12}, t} e^{-\lambda_{12}, t}$$

(3)

3 Note that $t$ in our notation represents an integer.

\[ p(\Delta x_{12}, t) = \frac{\lambda_{12} \Delta t}{\Delta x_{12}, t} e^{-\lambda_{12}, t} \]

Similarly, we have

$$p(\Delta x_{21}, t) = \frac{\lambda_{21} \Delta t}{\Delta x_{21}, t} e^{-\lambda_{21}, t}$$

(4)

$$p(\Delta x_{23}, t) = \frac{\lambda_{23} \Delta t}{\Delta x_{23}, t} e^{-\lambda_{23}, t}$$

(5)

where

$$\lambda_{12}, t = c_1 x_{1,t-1} \Delta t.$$

In our work, we use the Poisson model.

In general, we have $N$ species in the network, $x_t = [x_1(t), x_2(t), \ldots, x_N(t)]^T$ and $L$ stochastic rate constants $c = [c_{12}, c_{21}, \ldots, c_{L}]^T$. Based on measurements of some of the species $y_t$, we want to estimate the evolution of the system in time and to estimate the unknown vector of constants $c$.

3. PROPOSED SOLUTION

From the previous section, it is clear that we have a standard problem of tracking a state vector $x_t$ in time, given a set of observations $y_t$. The probability distributions that correspond to the state and observations equations of the system are given by $p(x_t|x_{t-1}, c)$ and $p(y_t|x_t)$. As stated in the introduction, we propose to solve this problem by using the PF methodology.

In mathematical terms, we are interested in the joint posterior distribution given by $p(x_0, c_0 | y_0)$.

We note here that the vector $c$ is composed of constants. In PF, it is well known that we have to take special care when the problem involves estimation of constants (see for example, [15], [16]). Here we treat $c_2$ as a random vector whose distribution is represented by its posterior.

We express the joint posterior as

$$p(x_0, c_0 | y_0) \propto p(y_0 | x_0, c_0) p(x_0 | x_0-1, c_0)$$

$$\times p(c_0 | x_0) p(x_0-1 | c_0)$$

(6)

where $\propto$ symbolizes proportionality. When we apply the importance sampling principle and we use an importance function that satisfies

$$\pi(x_0, c_0 | y_0) = \pi(x_0, c_0)$$

$$\times \prod_{k=1}^{t} \pi(x_k, c_k | x_{k-1}, c_{k-1}, y_{0:k})$$

(7)

we obtain the recursions that are used in PF.

A critical issue in our implementation would be the ability to propose particles for $c_2$. We assume that the stochastic rate constants are independent and the initial particles $c_0^{(m)}$ of the streams are obtained from Gamma distributions, denoted by $Ga(\alpha, \beta)$, where $\alpha$ and $\beta$ are its parameters. At
When we apply this prior, we obtain a posterior distribution \( p(\mathbf{c}|\mathbf{x}_{0:t}) \) which is also a product of one-dimensional Gamma distributions but with different parameters, \( \alpha_{t,1} \) and \( \beta_{t,1} \), \( l = 1, 2, \ldots, L \). The update of the parameters \( \alpha_{t,1} \) and \( \beta_{t,1} \) from \( \alpha_{t,0} \) and \( \beta_{t,0} \) can readily be obtained (see also the example below). More importantly, as we proceed, the posterior of \( p(\mathbf{c}|\mathbf{x}_{0:t}) \) remains of the same form (a product of Gamma distributions) with time-varying parameters. This implies that we can propose particles of \( \mathbf{c} \) from the posterior. It is also important to observe that given the vectors of constants \( \mathbf{c}_{t-1} \), we can generate the particles of \( \mathbf{x}_t \) from \( p(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{c}_{t-1}) \).

In PF, the posterior is approximated by the random measure

\[
\chi_t = \{\mathbf{x}_t^{(m)}, \mathbf{c}_t^{(m)}, w_t^{(m)})\}_{m=1}^M
\]

where \( m \) is an index for particles, \( M \) is the total number of particles, and \( w_t^{(m)} \) is the weight of the \( m \)-th particle \( \{\mathbf{x}_t^{(m)}, \mathbf{c}_t^{(m)}\} \). The weights of the particles are updated by

\[
w_{t-1}^{(m)} \propto \frac{p(\mathbf{y}_t|\mathbf{x}_t^{(m)})p(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{c}_{t-1})p(\mathbf{c}_t|\mathbf{x}_{0:t})}{\pi(\mathbf{x}_{t-1},\mathbf{c}_{t-1})\pi(\mathbf{c}_t|\mathbf{x}_{0:t})}.
\]

Thus, if

\[
\pi(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{c}_{t-1}) = p(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{c}_{t-1})
\]

\[
\pi(\mathbf{c}_t|\mathbf{x}_{0:t}) = p(\mathbf{c}_t|\mathbf{x}_{0:t})
\]

for the update we get

\[
w_{t-1}^{(m)} \propto w_{t-1}^{(m)} p(\mathbf{y}_t|\mathbf{x}_t^{(m)}).
\]

In summary, if we can sample the particles \( \mathbf{x}_t^{(m)} \) from \( p(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{c}_{t-1}) \) and \( \mathbf{c}_t^{(m)} \) from \( p(\mathbf{c}_t|\mathbf{x}_{0:t}) \), the update of the weights simplifies considerably. In the next subsection, we provide an example where we show the details of the implementation of the method.

### 3.1 Example

Consider the reaction

\[ X_1 \overset{c_1}{\rightarrow} X_2 \overset{c_2}{\rightarrow} X_3. \]

We assume that we know the number of molecules of the species \( X_1 \), \( X_2 \), and \( X_3 \) at time instant \( t = 0 \), that is, \( x_0 \) and that the priors of the stochastic rate constants are modeled by the Gamma probability distribution \( \text{Ga}(\alpha, \beta) \).

We implement the PF scheme as follows:

**Initialization.** We initialize the particle streams by generating at time instant \( t = 0 \), \( c_1^{(m)} \sim \text{Ga}(\alpha_1, \beta_1) \) and \( c_2^{(m)} \sim \text{Ga}(\alpha_2, \beta_2) \), and by setting \( x_0^{(m)} = x_0 \), for \( m = 1, 2, \ldots, M \).

In addition we set the parameters of the Poisson distributions by

\[
\lambda_{1,0}^{(m)} = x_{1,0} c_{1,0} \Delta t
\]

\[
\lambda_{2,0}^{(m)} = x_{2,0} c_{2,0} \Delta t.
\]

Finally, to each particle stream, we assign the same weights, \( w_0^{(m)} = 1/M \).

Given the random measure \( \chi_{t-1} = \{\mathbf{x}_{t-1}^{(m)}, \mathbf{c}_{t-1}^{(m)}, w_{t-1}^{(m)})\}_{m=1}^M \), we update it to \( \chi_t = \{\mathbf{x}_{t}^{(m)}, \mathbf{c}_{t}^{(m)}, w_{t}^{(m)})\}_{m=1}^M \) by implementing the following steps:

**Update the random measure.** We implement the following steps:

1. Generate the particles that contribute to the changes in the state vector \( \mathbf{x}_t \) according to

\[
\Delta x_{2,t}^{(m)} \sim \mathcal{P}(\lambda_{1,2}^{(m)})
\]

\[
\Delta x_{23,t}^{(m)} \sim \mathcal{P}(\lambda_{23}^{(m)})
\]

where \( \mathcal{P}(\cdot) \) stands for the Poisson distribution. Update the state vector by

\[
\mathbf{x}_t^{(m)} = \mathbf{x}_{t-1}^{(m)} + \Delta x_{1,t}^{(m)} + \Delta x_{2,t}^{(m)} + \Delta x_{23,t}^{(m)}.
\]

2. Generate \( \mathbf{c}_t^{(m)} \) by

\[
c_{1,t}^{(m)} \sim \text{Ga}(\alpha_{1,t}^{(m)}, \beta_{1,t}^{(m)})
\]

\[
c_{2,t}^{(m)} \sim \text{Ga}(\alpha_{2,t}^{(m)}, \beta_{2,t}^{(m)})
\]

where

\[
\alpha_{1,t}^{(m)} = \alpha_{1,t-1}^{(m)} + \Delta x_{12,t}^{(m)}
\]

\[
\beta_{1,t}^{(m)} = \beta_{1,t-1}^{(m)} + x_{1,t}^{(m)} \Delta t
\]

\[
\alpha_{2,t}^{(m)} = \alpha_{2,t-1}^{(m)} + \Delta x_{23,t}^{(m)}
\]

\[
\beta_{2,t}^{(m)} = \beta_{2,t-1}^{(m)} + x_{2,t}^{(m)} \Delta t.
\]

3. Update the parameters of the Poisson distributions by

\[
\lambda_{1,t}^{(m)} = x_{1,t}^{(m)} c_{1,t}^{(m)} \Delta t
\]

\[
\lambda_{2,t}^{(m)} = x_{2,t}^{(m)} c_{2,t}^{(m)} \Delta t.
\]

4. Compute the weights of the particles \( \mathbf{x}_t^{(m)}, \mathbf{c}_t^{(m)} \) by

\[
w_{t}^{(m)} \propto w_{t-1}^{(m)} p(\mathbf{y}_t|\mathbf{x}_t^{(m)}).
\]

5. Resample if necessary.

It is important to point out that at every time instant \( t \) we have the complete marginal posterior of \( \mathbf{c} \) from every particle stream. If we are interested in point estimates of \( \mathbf{c} \), we can use the mean of the weighted particles \( \mathbf{c}_t^{(m)} \) or we can compute the average of the maximum a posteriori estimates obtained from each particle stream.
4. COMPUTER SIMULATIONS

In this section, we provide some examples related to the studied problem.

4.1 Example 1: $X_1 \xrightarrow{c_1} X_2$

We carried out a simple simulation experiment to illustrate the performance of a particle filter that jointly tracks the evolution of the species and estimates the stochastic rate constant $c_1$. In the experiment, the number of molecules of $X_1$ were observed with error. The transition of the state was modeled by $x_{1,t} \sim p(x_{1,t}|x_{1,t-1}, c_{t-1})$ where $p(\cdot)$ was a Poisson distribution.\(^4\) The measurement of the number of molecules $x_{1,t}$ was modeled by

$$y_t = g(x_{1,t}) + v_t$$

where $g(\cdot)$ was a function of the number of molecules (non-linear measurements from fluorescence spectroscopy experiments \cite{17}) and $v_t$ was noise (or error) which was modeled as a Gaussian with zero mean and variance $\sigma_v^2 = 100$.

We considered a system whose initial number of molecules of $X_1$ and $X_2$ were set to 100 and 0, respectively, and where the stochastic rate constant was $c_1 = 0.01$. The experiment was run for $T = 500$ sec and a time step of $\Delta t = 0.5$ sec. The priors for the state and the stochastic rate constant were set to $x_{0|m} = x_{1,0}$ and $c_{1,0}^{(m)} \sim \text{Ga}(1,50)$ for $m = 1, 2, \ldots, 200$, respectively. Fig. 1 (left) shows the evolution of the species in the system in a single simulation run obtained by the the particle filter. It is apparent that the PF algorithm tracks the evolution of the species very accurately and remains locked to the true value (the curve representing the true values and the curve depicting the estimates are almost indistinguishable). Fig. 1 (middle) depicts the estimate of the stochastic constant rate by the particle filter. In Fig. 1 (right) we show the histogram of the estimated $c_1$ in 1000 runs.

4.2 Example 2: $X_1 \xrightarrow{c_1} X_2 \xrightarrow{c_2} X_3$, known $c_2$

For this scenario, the state of the system was $x_t = [x_{1,t}, x_{2,t}, x_{3,t}]^T$. Assuming that we only observe a noisy function of $X_1$ molecules, it is evident that estimating $c_2$ is not possible since it is non-observable. We considered the joint estimation of $x_t$ and $c_1$ for an analogous experiment as in the previous example with initial number of molecules of $X_3 = 0$, and constants $c_1 = 0.02$ and $c_2 = 0.01$. For this case the prior of $c_1$ was slightly modified to $\text{Ga}(1,50)$. Fig. 2 shows the obtained results. Similar conclusions as in the previous case can be drawn.

4.3 Example 3: $X_1 \xrightarrow{c_1} X_2 \xrightarrow{c_2} X_3$, unknown $c_1$ and $c_2$

We considered the case when noisy measurements of $X_1$ and $X_3$ are available, and therefore we can also estimate $c_2$. We modeled the observation $y_t = [y_{1,t}, y_{2,t}]^T$ following equation (2) where $v_t \sim \mathcal{N}(0, 100 I_2)$ with $I_2$ denoting the identity matrix of size $2 \times 2$. We considered $c_{1,0}^{(m)} \sim \text{Ga}(1,50)$

\(^4\)Note that the number of molecules of $X_2$ can be calculated in a straightforward manner once $x_{1,t}$ is obtained.
and $c_{m}^{(m)} \sim Ga(1,80)$. The results shown in Fig. 3 are as expected and in accordance with the previous ones.

5. CONCLUSIONS

The inverse problem of biochemical networks involves estimation of unknowns in the networks based on measurements that represent time series of functions of the numbers of molecules of some of the species. In this paper, we proposed to solve an inverse problem of a network with first-order reactions by particle filtering. The method tracks the changing numbers of molecules of the species in the system and estimates the posterior distributions of the stochastic rate constants of the reactions. We were able to sample particles of the stochastic rate constants from their posteriors directly, which significantly simplified the implementation. Thereby, we did not have to apply methods that treat the constants in a special way in order to avoid the quick degeneracy of the random measures produced by the particle filter. We provided simulation results that show the performance of the method.

REFERENCES


