2013 European Control Conference (ECC)
July 17-19, 2013, Zürich, Switzerland.

State estimation for gene networks with intrinsic and extrinsic noise: A case study on *E.coli* arabinose uptake dynamics*

Alfonso Carta[1] and Eugenio Cinquemani[2]

***Abstract*— We address state estimation for gene regulatory networks at the level of single cells. We consider models that include both intrinsic noise, in terms of stochastic dynamics, and extrinsic noise, in terms of random parameter values. We take the Chemical Master Equation (CME) with random parameters as a reference modeling approach, and investigate the use of stochastic differential model approximations for the construction of practical real-time filters. To this aim we consider a Square-Root Unscented Kalman Filter (SRUKF) built on a Chemical Langevin Equation (CLE) approximation of the CME. Using arabinose uptake regulation in *Escherichia coli* bacteria as a case study, we show that performance is comparable to that of a (computationally heavier) particle filter built directly on the CME, and that the use of information about parameter uncertainty allows one to improve state estimation performance.**

I. Introduction

A key player of single-cell gene network dynamics is noise [1]. A distinction is usually made between intrinsic noise, i.e. the uncertainty inherent in biochemical events (binding/unbinding of transcription factors, synthesis of mRNA or protein molecules, etc.), and extrinsic noise, such as the variability of individual features over an isogenic population (abundance of aspecific transcription/translation factors, local environmental conditions, etc.) [2].

Gene expression monitoring techniques with single-cell resolution opened the way to the identification of stochastic gene network models. The CME [3], describing the kinetics of the network species in terms of probabilistic reaction events, is a standard tool for the description of intrinsic noise. To account for extrinsic noise, similar to Mixed-Effects (ME) modeling [4], one approach is to describe the parameters of the network dynamics as random variables taking different values in different individuals (see e.g. [5]). Stochastic gene network models are nowadays fundamental tools for understanding the behavior of cells in face of environmental and evolutionary challenges [6]. Most recently, they have also been considered for the real-time computer-based control of gene expression in single cells [7], [8].

This paper investigates state estimation from cell-level measurements for networks with intrinsic and extrinsic noise. State estimation is interesting *per se* for the reconstruction of network states that cannot be measured directly. In addition, it can be used as an intermediate step for identification, and plays a central role toward model-based control.

We start from the CME as the reference ("true") model of a cell network. Inspired by the ME approach [4], [5], we include extrinsic noise in terms of variability of the model parameters. Since CME appears to be impractical for real-time filtering, we propose to use an asymptotic approximation, the Chemical Langevin Equation (CLE) [9], to implement filtering. First, we compare simulations of the CME and CLE models. Then, we use the latter to construct a Square-Root Unscented Kalman Filter (SRUKF) [10], [11], [12]. Using data generated from the true (CME) system, we compare performance of SRUKF with that of a (computationally heavier) particle filter built directly on the CME [13].

We develop our work on the case study of the network regulating the uptake of sugar arabinose in bacteria *Escherichia coli*. While relatively simple, this well characterized system (see [14] and references therein) is representative of the nature and complexity of the genetic feedback mechanisms regulating bacterial response to environmental stress. Different from e.g. [15], [5], where the observations consist of time series of the empirical distribution of gene expression obtained via flow cytometry, we consider the case where the expression in every cell of a small population is observed over time, as it can be obtained e.g. by fluorescence microscopy (see e.g. [7]).

Bayesian inference, such as parameter and state estimation, for biological networks has been considered before, see e.g. [16], [17]. Here, we focus on state estimation under parameter uncertainty. First, we show that, despite the known limitations for small molecule numbers, the CLE is a viable CME approximation for the construction of computationally affordable filters coping with intrinsic noise (stochastic dynamics) and extrinsic noise (random parameters). Second, we show that the use of ME-type models, accounting for parameter variability, may improve state estimation performance.

II. Stochastic modeling of genetic networks

Consider a biochemical reaction network involving n species and m possible reactions among them. For gene regulatory networks, the species are typically proteins, mRNAs, transcription factors, etc., while reactions are e.g. binding/unbinding events, formation of complexes, degradation, and, at a higher level of abstraction, gene expression.

Assume that the reaction volume is uniform. For cells or cell nuclei, this assumption is still accepted in many contexts, as long as spatial resolution is not central. Let $X = (X_1, \ldots, X_n) \in \mathbb{N}^n$, where X_i denotes the number

* The work of A.Carta was supported by ANR GeMCo (French national research agency)

[1]A. Carta is with team BIOCORE at Inria Sophia-Antipolis, France alfonso.carta@inria.fr (Corresponding author)

[2]E. Cinquemani is with team IBIS at Inria - Grenoble Rhône-Alpes, France eugenio.cinquemani@inria.fr

of elements of species i, with $i = 1, \ldots, n$. Let $\nu_j \in \mathbb{Z}^n$ be the stoichiometry of reaction j, with $j = 1, \ldots, m$. That is, element i of ν_j, denoted $\nu_{j,i}$ is the number of elements of species i produced or consumed in reaction j. Assume that reactions occur stochastically with propensities $a_j \in \mathbb{R}_{\geq 0}$ generally depending on X.Then X is the random state vector of a Markovian jump process. For times $t \geq 0$, say, the probability $p(Z,t) = \text{Prob}(X(t) = Z)$, $Z \in \mathbb{N}^n$, obeys the CME [3]

$$\frac{dp(Z,t)}{dt} = \sum_{j=1}^{m} a_j(Z - \nu_j)p(Z - \nu_j, t) - a_j(Z)p(Z,t). \quad (1)$$

whose solution is fully determined given the distribution of $X(0)$. The CME is a linear but infinite-dimensional differential equation. For all but the simplest systems, the exact solution cannot be computed in practice. Simulated sample trajectories can be obtained by the Gillespie and related algorithms [3]. Under appropriate conditions on the process X, typically satisfied for large numbers X_i, the jump process X is well approximated by a continuous process with state $x \in \mathbb{R}^n_{\geq 0}$ that satisfies the so-called Langevin equation [9], i.e. the system of stochastic differential equations

$$\frac{dx_i(t)}{dt} = \sum_{j=1}^{m} \nu_{j,i} a_j\big(x(t)\big) + \nu_{j,i}\sqrt{a_j\big(x(t)\big)}\Gamma_j(t), \quad (2)$$

with $i = 1, \ldots, n$, where, for $j = 1, \ldots, m$, the $\Gamma_j(t)$ are mutually uncorrelated white noise processes. Here x plays the role of a continuous approximation of the molecule count X. Eq. (2) equally describes the evolution of molar concentrations $x = X/(\Omega N_A)$[M], where Ω is the reaction volume, N_A is Avogadro's number and [M] denotes molar (moles/liter) units, provided appropriate rescaling of the reaction propensities and their parameters. From now on we assume $x = X/(\Omega N_A)$ and omit symbol [M] where no confusion may arise.

Inter-individual variability: Similar to ME-modeling in pharmacokinetics [4], variability of reaction dynamics among different cells (extrinsic noise) can be described in terms of inter-individual variability of the parameters of a common kinetic model [5]. Using a Bayesian approach, one assumes that individual parameters are concentrated around a known population average (so-called fixed-effects) but deviate from it by a quantity modeled as a random variable with a given prior. This prior, characteristic of the population, is inferred from a set of representative individuals. Then, deviations of new individuals from the population average are treated as random outcomes from the same prior. In our context, let a_j^θ denote that reaction propensity a_j depends on a parameter vector θ. Let X^ℓ (resp. x^ℓ) and θ^ℓ be the state (resp. the state of the Langevin approximation) and the parameters of the ℓth cell in a population of N cells. Then X^ℓ (resp. x^ℓ) evolves according to the dynamics determined by $a_j^{\theta_\ell}$. To model individual variations from population average, we assume that $\theta^\ell, \ldots, \theta^N$ are independent identically distributed (i.i.d.) outcomes of the random variable θ defined by

$$\theta = \bar{\theta} + \delta, \quad (3)$$

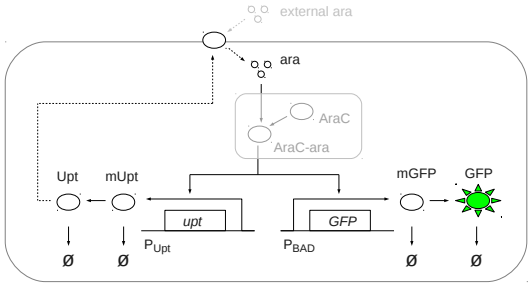


Fig. 1. Arabinose uptake regulatory network in (modified) *E.coli* cells (inspired from [14]). In absence of glucose, global transcriptional regulators enable the expression of arabinose import enzymes, here lumped into a "virtual" enzyme Upt, and metabolization enzyme AraBAD, under the control of promoter P_{BAD}. In the modified strains we refer to, gene *araBAD* is deleted from the DNA and replaced by the P_{BAD}-controlled *GFP* (Green Fluorescent Protein) gene on plasmids. When arabinose is present, it forms complexes with AraC molecules, which promote transcription of the above genes, thus increasing the amount of arabinose transporters and fluorescent molecules. More transporters imply faster arabinose uptake and hence higher transporter synthesis rate, in a positive feedback loop. Parts in grey will not be modeled explicitly.

where $\bar{\theta}$ is fixed and δ is a random variable with known distribution F_δ. Provided suitable definition of F_δ, this model includes the case where some entries of θ are fixed.

Observation model: We consider the case where noisy measurements from individual cells are collected over time. Let $\mathscr{T} = t_k : k = 0, 1, 2, \ldots$, with $t_k < t_{k+1}$ for all k, be a set of measurement times. After standard preprocessing (such as e.g. background removal in fluorescent gene reporter systems) let $y^\ell(t)$ be the measurement at time $t \in \mathscr{T}$ for cell ℓ, with $\ell = 1, \ldots, N$. We assume that

$$y^\ell(t) = CX^\ell(t) + e^\ell(t), \qquad t \in \mathscr{T}, \quad (4)$$

where the output matrix C, typically selecting one of the system states, is known, and the $e^\ell(t)$ are i.i.d. noise samples of appropriate dimension from a Gaussian distribution $\mathscr{N}(0, R)$, with $R > 0$ known. More specific measurement models depend on the details of the experimental setup and are not pursued in this paper. Under the Langevin approximation, we will replace X^ℓ by the process x^ℓ from (2) and rescale quantities accordingly.

III. Case study: *E.coli* arabinose uptake dynamics

We are interested in the network that regulates the uptake of arabinose in *Escherichia coli*, a well characterized bacterium. Upon exhaustion of primary environmental carbon sources (glucose), *E.coli* activates adaptation mechanisms triggering the uptake of less favorable carbon sources such as arabinose (see [14], and references therein). A simplified representation of the system is in Fig.1.

We consider a model inspired by [14]. The model consists of the $n = 5$ species ara (arabinose), mUpt (Upt messenger RNA), Upt (Upt protein), mGFP (GFP messenger RNA), GFP (GFP protein), interacting via the $m = 12$ "lumped" reactions reported in Fig.2 with the corresponding propensities $a(x) = \big[a_1(x), \ldots, a_{12}(x)\big]$, where $x = \big[x_1, x_2, x_3, x_4, x_5\big]$ denotes amounts of ara, mUpt Upt, mGFP and GFP, in the

Synthesis	Rate a_j	Degradation	Rate a_j
$\varnothing \xrightarrow{a_1}$ ara	$v_1 x_3$	ara $\xrightarrow{a_2} \varnothing$	$\gamma_1 x_1$
$\varnothing \xrightarrow{a_3}$ mUpt (basal)	v_2^0	mUpt $\xrightarrow{a_5} \varnothing$	$\gamma_2 x_2$
$\varnothing \xrightarrow{a_4}$ mUpt (regulated)	$v_2 \dfrac{x_1^3}{K_u^3 + x_1^3}$		
mUpt $\xrightarrow{a_6}$ mUpt+Upt	$v_3 x_2$	Upt $\xrightarrow{a_7} \varnothing$	$\gamma_3 x_3$
$\varnothing \xrightarrow{a_8}$ mGFP (basal)	v_4^0	mGFP $\xrightarrow{a_{10}} \varnothing$	$\gamma_4 x_4$
$\varnothing \xrightarrow{a_9}$ mGFP (regulated)	$v_4 \dfrac{x_1^3}{K_u^3 + x_1^3}$		
mGFP $\xrightarrow{a_{11}}$ mGFP+GFP	$v_5 x_4$	GFP $\xrightarrow{a_{12}} \varnothing$	$\gamma_5 x_5$

Fig. 2. Reactions of the stochastic model of the system of Fig.1 and corresponding propensities. An arrow from (to) symbol $\varnothing$ means synthesis (degradation, including dilution effects due to cell growth).

$\bar{\theta}$	Values (CME)	Values (CLE)
v_1	$120\, min^{-1}$	$120\, min^{-1}$
v_2^0	$0.05\,\#\, min^{-1}$	$3.9643 \cdot 10^{-11}\, M\, min^{-1}$
v_2	$4.95\,\#\, min^{-1}$	$3.9643 \cdot 10^{-9}\, M\, min^{-1}$
K_u	$58541.79\,\#$	$4.6416 \cdot 10^{-5}\, M$
v_3	$4.16\, min^{-1}$	$4.16\, min^{-1}$
v_4^0	$0.05\,\#\, min^{-1}$	$3.9643 \cdot 10^{-11}\, M\, min^{-1}$
v_4	$4.95\,\#\, min^{-1}$	$3.9643 \cdot 10^{-9}\, M\, min^{-1}$
v_5	$5\, min^{-1}$	$5\, min^{-1}$

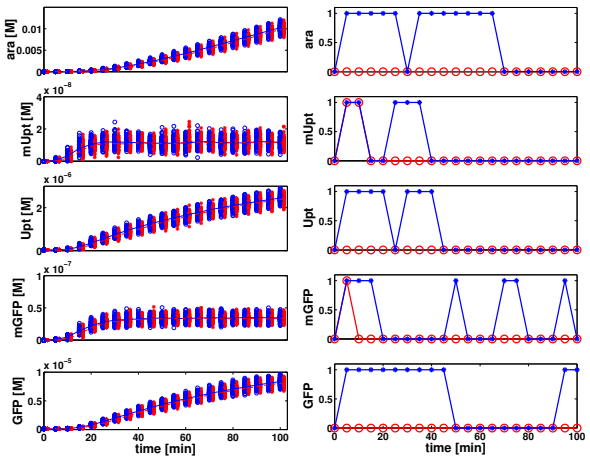
Fig. 3. Nominal parameter values for the CME (in molecule number units – symbol # denotes number of molecules) and the CLE (in concentration units – normalization factor $1/N_A\Omega \simeq 7.9 \cdot 10^{-10}$M/#). Degradation rates, equal for CME and CLE models, are $(\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5) = (0.0139, 0.347, 0.0139, 0.116, 0.0139) min^{-1}$.

same order. Nominal parameter values $\bar{\theta}$, mostly derived from [14], are listed in Fig.3. In this model, the rate of transcription of the Upt and GFP genes (resp. a_4 and a_9) is described as a switch-like (Hill) function of the internal arabinose abundance, with a threshold parameter that depends on the concentration of unmodeled regulators (notably AraC, see Fig. 1 and [14]). The observed variable is a fluorescence level proportional to the amount of GFP, i.e. $C = [0\ 0\ 0\ 0\ K]$, for simplicity we take $K = 1$. The CME and CLE models follow from replacing the stoichiometries $\nu_1, \ldots, \nu_{12}$ and the propensities of the model of Fig. 2 into (1) and (2). In particular, the CLE can be written in the matrix form

$$\dot{x} = V\ a(x) + H\ V\ \text{diag}\left(\sqrt{a(x)}\right)\ \Gamma \qquad (5)$$

where $V = [\nu_1, ..., \nu_{12}]_{5\times 12}$, diag $\left(\sqrt{a(x)}\right)$ is the diagonal matrix having the square root of the entries of vector $a(x)$ on the diagonal, $\Gamma = [\Gamma_1, ..., \Gamma_m]^T$ and $H = 1/\sqrt{N_A\Omega}$.

To get an insight into the accuracy of the CLE approximation of the CME, we analyze numerical simulations of the two processes. To simulate the CLE (5) we used a modified version of the Euler-Maruyama method [18] (with sampling time of 0.1[min]) which shuts down a reaction channel when the amount of any its reactants reaches zero, for preserving non-negativity of the system state [19]. To simulate the CME (1) we used a customized version of software *StochKit* [20]. Simulations are started from the state $X^- = (0, 0, 43, 0, 155)$ which is (up to integer round-off) the expected state of equilibrium before the arabinose uptake



(a) (b)

Fig. 4. (a) Comparison of 100 simulations of the CME model (red, stars) and the CLE approximation (blue, circles) at times $\mathscr{T}$. At each time point, samples from the two simulations are plotted next to each other (slightly off the corresponding time point) for visual comparison. Lines indicate simulation mean for CME (red dashed) and CLE (blue solid). (b) Results of the Kolmogorov-Smirnov test for the CME and CLE simulations in Fig. 4(a) (blue), and for the same data after equalization of the means (red). The test is applied, separately for every state entry, to the simulated states at times $\mathscr{T}$. A value of 0 (resp. 1) means that the hypothesis of equivalence of the distributions has been accepted (resp. rejected) with 95% confidence level.

mechanism kicks in. In practice, this value is computed by approximately solving the equation $0 = Va(x^-)$ subject to $x_1^{-1} = 0$, with $x^- = X^-/(\Omega N_A)$, with respect to $X^- \in \mathbb{N}^n$. The simulated states $X(t)$ and $x(t)$ are recorded at times $\mathscr{T} = \{t_k = k \cdot T :\ k = 0, \ldots, 20\}$, with $T = 5$[min]. Results from 100 simulated trajectories are reported in Fig. 4(a). Distributions look similar, but the Langevin approximation appears to be slightly biased. At every time k, the hypothesis of equal distributions was tested by applying a standard two-sample Kolmogorov-Smirnov test first on the simulated states, then on the same data but with their means equalized. Results in Fig. 4(b) show that, while the hypothesis is often rejected before bias correction, this is no longer the case after mean equalization, except for small molecule numbers (see e.g. mUpt and mGFP around time 10) where the CLE is known to perform worse [9]. We will discuss the implications of this for filtering in the next section.

IV. Gene network state estimation

Consider a cell population model specified as in Section II. Given an initial distribution $p(\cdot, 0)$ at initial time $t_0 = 0$, the dynamics of X^ℓ in every individual $\ell = 1, \ldots, N$ are described by the propensities $a_j^{\theta_\ell}$ of the reactions ν_j, with $j = 1, \ldots, m$, and the individual parameters θ_ℓ follow (3) with assigned distribution F_δ. Let measurements (4) be available for all individuals. We consider the following real-time state estimation problem in all individuals.

Problem 1: Let $\mathscr{Y}^\ell(t) = \{y^\ell(t_k) :\ t_k \le t\}$. For $t \ge 0$, compute $\mathbb{E}[X^1(t), \ldots X^N(t) | \mathscr{Y}^1(t), \ldots, \mathscr{Y}^N(t)]$.

Expectation is taken with respect to the process laws and the random deviations of the individual parameters from population average. Since these quantities are assumed to

be statistically independent across individuals, the problem splits into N estimation problems, one per individual. Focusing on the problem of filtering, the problem reformulates as follows.

Problem 2: For every $\ell = 1, \ldots, N$, and $k \in \mathbb{N}$, compute $\hat{X}^\ell[k|k] = \mathbb{E}[X^\ell(t_k)|\mathscr{Y}^\ell(t_k)]$.

Being the problem identical for every individual, from now on we drop superscript "ℓ" from the notation and look at the generic individual (process) X with parameters θ obeying (3).

Given the known difficulty of solving the CME, the problem of computing $\hat{X}^\ell[k|k]$ appears quite challenging and is further complicated by the randomness of parameters θ. Possible but computationally demanding approaches include Markov Chain Monte Carlo (MCMC) sampling [17] and particle filtering [13], [16]. We consider that computationally more effective solutions may be obtained based on the CLE approximation (2). We saw in Section III that, at least in our case study, the CLE provides a viable approximation of the CME, with the exception of a small bias. For state estimation purposes, the closed-loop correction of the estimates as new measurements become available is expected to compensate for this. Since, conditionally on θ, process x in (2) is an approximation of process X, it is natural to see the augmented process $\xi = (x, \theta)$ as an approximation of (X, θ), and to approximate the solution of Problem 2 by computing $\hat{x}[k|k] = \mathbb{E}[x(t_k)|Y(t_k)]$ in place of $\hat{X}[k|k] = \mathbb{E}[X(t_k)|Y(t_k)]$. Following a well-known approach (see e.g. [21]), the dynamics of ξ can be written by treating θ as an invariant state $\theta(t)$ with random initial condition. Combining $\dot{\theta}(t) = 0$ with Eq. (5) leads to the augmented Langevin system

$$\dot{\xi}(t) = \begin{bmatrix} Va(\xi(t)) \\ 0 \end{bmatrix} + \begin{bmatrix} HV\text{diag}\left(\sqrt{a(\xi(t))}\right)\Gamma \\ 0 \end{bmatrix} \quad (6)$$

with a priori distribution of $\theta(t)$ at time t_0 given by (3). Based on this, we address the following problem, slightly more general than the computation of $\mathbb{E}[x(t_k)|Y(t_k)]$.

Problem 3: Compute $\hat{\xi}[k|k] = \mathbb{E}[\xi(t_k)|\mathscr{Y}(t_k)]$, $\forall k \in \mathbb{N}$.

In the next section we will provide one solution based on a version of the so-called Unscented Kalman Filter (UKF) [11]. The goodness of the resulting estimates $\hat{x}[k|k]$ relative to a direct solution of Problem 2 will be assessed in simulation by comparison with a particle Filter (PF), the Bootstrap filter, built on model (1). This PF is implemented in accordance with [13, §1.3.3] , i.e. sampling the particle dynamics by Gillespie simulation of (1), via customized *StochKit* [20] software and Matlab.

A. The Square-Root Unscented Kalman Filter

To solve Problem 3, we present a continuous-discrete SRUKF. A square-root version of UKF has been chosen to improve numerical stability and also to speed up filtering via the direct use of the matrix roots from UKF. The filter is built for system (6) with measurements (4) rewritten as $y(t_k) = \bar{C}\xi(t_k) + e(t_k)$, with $\bar{C} = [C\ 0]$, where the $0s$ account for the extension of the state.

The SRUKF utilizes a deterministic sampling approach for the approximate computation of $\hat{\xi}[k|k]$. Let L be the dimension of ξ. So-called sigma vectors $\mathcal{X}_i$, with $i = 0, \ldots, 2L$, are chosen after each measurement update based on a square-root decomposition of the estimation covariance P and mean $\mu = \hat{\xi}[k|k]$, and used in a two steps (*prediction* and *correction*) algorithm to compute weighted mean and covariance approximating the true conditional distribution of ξ at each t_k. Let $A(t) = \text{chol}(P(t))$, that is A is computed as the lower triangular Cholesky factor of the covariance P. Following [12], one defines $\mathcal{X}_0 = \mu$, $\mathcal{X}_i = \mu + (\sqrt{c}A)_i$ for $i = 1, ..., L$ and $\mathcal{X}_i = \mu - (\sqrt{c}A)_i$ for $i = L+1, ..., 2L$, and corresponding weights $W_0^{(\mu)} = \lambda/c$, $W_i^{(\mu)} = W_i^{(c)} = \lambda/(2c)$, for $i = 1, ..., 2L$, and $W_0^{(c)} = W_0^{(\mu)} + (1 - \alpha^2 + \beta)$. The parameters $c = \alpha^2(L + \kappa)$ and $\lambda = c - L$ are scaling parameters with positive constants α, β and κ used to tune the SRUKF (in our applications we set $\alpha = 0.17$, $\kappa = 200$, $\beta = 2$). Define $\mathcal{X} = [\mathcal{X}_0, \ldots, \mathcal{X}_{2L}]$, $w_m = \left[W_0^{(\mu)} \ ... \ W_{2L}^{(\mu)}\right]^T$, $W = H\,\text{diag}\left(W_0^{(c)} \ ... \ W_{2L}^{(c)}\right) H^T$, with $H = (I - [w_m \ ... \ w_m])$,

$$\underbrace{\begin{bmatrix} Va(\mathcal{X}_1) & \cdots & Va(\mathcal{X}_{2L}) \\ 0 & \cdots & 0 \end{bmatrix}}_{\triangleq F(\mathcal{X})}, \quad \underbrace{\begin{bmatrix} HV\text{diag}\left(\sqrt{a(\mathcal{X}_i)}\right) \\ 0 \end{bmatrix}}_{\triangleq G(\mathcal{X}_i)}.$$

a) Prediction: From time t_{k-1} to t_k, the SRUKF prediction equations [12], [10] can be written in terms of sigma vectors as $B(t) = \sqrt{c}\begin{bmatrix} 0 & A(t)\Phi(M(t)) & -A(t)\Phi(M(t)) \end{bmatrix}$, $M(t) = A^{-1}(t)[\mathcal{X}(t)WF^T(\mathcal{X}(t)) + F^T(\mathcal{X}(t))W\mathcal{X}(t) + \sum_{i=0}^{2L} W_i^{(\mu)} G(\mathcal{X}_i(t))G^T(\mathcal{X}_i(t))]A^{-T}(t)$,

$$d\mathcal{X}_i(t)/dt = F(\mathcal{X}(t))w_m + B_i(t), \quad i = 0, ..., 2L,$$

where $\Phi(\cdot)$ is a function defined as: $\Phi_{ij}(M) = M_{ij}\ \forall i > j$, $\Phi_{ij}(M) = 1/2M_{ij}\ \forall i = j$, $\Phi_{ij}(M) = 0$ otherwise. In practice, the above equationd are integrated numerically as follows. Choosing a discretization interval of δt (we set $\delta t = 0.005$[min]) and dividing the interval between measurements into $J = (t_k - t_{k-1})/\delta t$ subintervals, one computes $\mathcal{X}_i(t + \delta t) = \mathcal{X}_i(t) + [F(\mathcal{X}(t))w_m + B_i(t)]\,\delta t$, $i = 0, ..., 2L$ iteratively from t_{k-1} to t_k. At each iterate, one extracts $A(t)$ from the current $\mathcal{X}(t)$ and updates $F(\mathcal{X}(t))$ and $B(t)$ accordingly. This eventually yields new $\mathcal{X}(t_k)$ and $A(t_k)$, from which the a priori moments are given by $\hat{\xi}[k|k-1] = \mathcal{X}(t_k)w_m$, $P[k|k-1] = A(t_k)A^T(t_k)$.

b) Measurement update: For a new measurement $y(t_k)$, $\hat{\xi}[k|k]$ and $P[k|k]$ are computed from the above $\hat{\xi}[k|k-1]$ and $P[k|k-1]$ by a standard Kalman update step according to our linear measurement model.

c) SRUKF Initialization: We set $\hat{\xi}[0|-1] = \mathbb{E}[\xi(t_0)]$ and $P[0|-1] = Var[\xi(t_0)]$, where the statistics of ξ at t_0 are determined by the priors on $x(t_0)$ and θ.

V. State estimation: Simulation results for the *E.coli* arabinose uptake system

A. Comparison of SRUKF and PF

To evaluate the CLE approximation of the CME for filtering performance, we compare the (CLE-based) SRUKF

with a (CME-based) PF using $P = 1000$ particles. This comparison is carried out with model parameters fixed to the nominal values of Table 3. Data are generated by simulating the "true" CME model. We consider two scenarios.

Scenario 1: We consider simulations always starting from state X^- of Sec. III), and initialize the filters at X^- and x^-, assuming this state is known. This choice is relevant to experiments where the beginning of arabinose uptake occurs at a known time, e.g. at the delivery of arabinose in a glucose-poor medium. We simulated 100 trajectories $X(t)$ and produced corresponding fluorescence measurements $y(t)$ at times $\mathscr{T}$ by corrupting the simulated values of X_5 with zero-mean Gaussian noise with standard deviation σ fixed to $4 \cdot 10^{-7}$[M], which is approximately 10% of the mean observed value of $X_5/(N_A\Omega)$. On each trajectory, we ran SRUKF and PF from the true initial state and null variance.

Scenario 2: We assume that the system has started the arabinose uptake mechanism earlier than expected. This is relevant to experiments where the start of arabinose uptake is somewhat undetermined, e.g. for bacteria placed in an arabinose-rich medium where the switch to arabinose depends on depletion of environmental glucose. To simulate this we considered the same 100 simulations of the previous dataset, discarding the first two measurement times (thus taking $t = 10$[min] as the initial time) and extending the simulations to include two additional measurements. For each trajectory, we ran both SRUKF and PF initialized as in the first scenario, in the (wrong) belief that the system is observed starting from an equilibrium state.

Let $\hat{x}[k|k]$ and $\hat{X}[k|k]$ be the estimates of the SKRUF and of the PF, respectively. Denote with $\mu[k]$ the *a priori* mean of the process X (computed from 10000 separate Gillespie simulation runs). Note that $\mu[k]$ can be thought of as the *a priori* state estimator. To compare performance, from the 100 simulations we compute empirically the estimation error time series (in the concentrations domain)

$$\begin{aligned} \bar{e}[k|k] &= \mathbb{E}\big[\,|\,X[k]/(N_A\Omega) - \hat{x}[k|k]\,|\,\big] && \text{(SRUKF)},\\ \bar{e}_{PF}[k|k] &= \mathbb{E}\big[\,|\,X[k] - \hat{X}[k|k]\,|\,\big]/(N_A\Omega) && \text{(PF)},\\ \bar{e}_{\mu}[k] &= \mathbb{E}\big[\,|\,[X[k] - \mu[k]\,|\,\big]/(N_A\Omega) && \text{(\textit{a priori})}. \end{aligned}$$

For reference, we also compute $\bar{x}[k] = \mathbb{E}[X(k)]/(N_A\Omega)$, the empirical mean of the 100 simulated trajectories on which filtering is performed.

In Scenario 1 filters performed nearly identically: For all k and all state entries $i = 1,\dots,5$, $|\,\bar{e}_i[k|k] - \bar{e}_{PF,i}[k|k]\,|\,/\bar{x}_i[k] \leq 10^{-2}$. However, both filters improved upon $\mu[k]$ only in the 5th (observed) state component (on average over time $\bar{e}_{\mu,5}/\bar{x}_5 = 0.083$ whereas $\bar{e}_5/\bar{x}_5 \simeq \bar{e}_{PF,5}/\bar{x}_5 \simeq 0.054$). The SRUKF appears to compensate for the bias of the CLE approximation (Sec. III), while neither SRUKF nor PF decreased the uncertainty on the unobserved states. In Scenario 2, Fig.5 shows plots of the relative estimation errors $\bar{e}[k|k]/\bar{x}[k]$, $\bar{e}_{PF}[k|k]/\bar{x}[k]$ and $\bar{e}_{\mu}[k]/\bar{x}[k]$ over time. Both filters improve significantly upon $\mu[k]$ in state entries 1, 3 and 5, and, to a lesser extent, in entries 2 and 4, showing that measurements are successfully exploited to compensate for the wrong belief on the initial state. Again, SRUKF and PF behave very similarly. We interpret these results as follows. While a more advanced PF (e.g a larger number of particles P) could be considered, at least for our case study, the modeling error introduced by approximating the CME with the CLE has no practical effect on estimation performance. In addition, in our implementation, the SRUKF is much faster than PF (filtering one trajectory takes more than 15min for PF and less than 2min for the SRUKF on a 64bits 3.20GHz 6-core 6Gb-RAM Linux workstation). Although the PF computational burden may be reduced e.g. by the use of tau-leaping [3], these results make CLE-based SRUKF an appealing alternative for real-time applications.

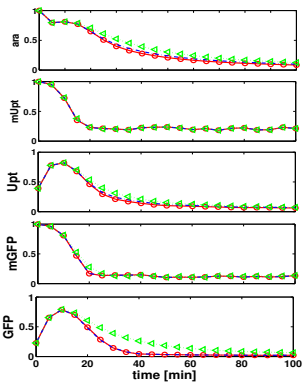


Fig. 5. Comparison of estimation errors $\bar{e}[k|k]/\bar{x}[k]$ (red, solid), $\bar{e}_{PF}[k|k]/\bar{x}[k]$ (blue, dash-dotted) and $\bar{e}_{\mu}[k]/\bar{x}[k]$ (green, dotted).

B. Performance of the SRUKF in presence of extrinsic noise

We now evaluate performance of the SRUKF when some parameters are random. We again consider two scenarios.

Scenario 1: All parameters are known and fixed to the nominal values except for K_u. This is the threshold of the sigmoidal function that determines the switch-like behavior of the system via regulation of mUpt synthesis, and may represent e.g. individual-dependent concentrations of unmodeled regulators (such as AraC, see Fig. 1 and [14]).

Scenario 2: We fix all parameters to nominal values except v_4, the maximal regulated synthesis rate of mGFP. Variability of v_4 may represent e.g. different number of promoters for GFP (unequal number of plasmids carrying the reporter) in different cells.

In both cases, we consider 100 values of the variable parameter sampled from a Gaussian distribution with mean equal to the nominal value and standard deviation equal to 20% of the mean (for consistency, we ensure that all samples are positive). For each of the 100 parameter values, we simulate the system once starting from X^- (which, by its definition, does not depend on K_u and v_4) and record state and noisy measurements at times $\mathscr{T}$, with noise distributed as in the previous section. SRUKF is run on each simulated trajectory initialized with the true statistics of the parameters and null-variance initial state estimate $X^-/(N_A\Omega)$. To assess the performance gain in using the information on parameter variability, we also run a SRUKF using the wrong belief that parameters are all fixed to nominal values, and refer to it as nominal SRUKF (nSRUKF). Let $\Xi[k]$ denote either $(X[k], K_u)$ or $(X[k], v_4)$, depending on the scenario. Let $\hat{\xi}[k|k]$ and $\hat{x}[k|k]$ denote the estimates from SRUKF and

nSRUKF, in the same order. Let $\mu[k]$ denote the *a priori* mean of process $\Xi[k]$ with nominal parameters (*a priori* estimator, with last entry now fixed to the nominal parameter value). From the 100 filtering results we compute the (empirical) statistics

$$\begin{aligned}
\bar{e}[k|k] &= \mathbb{E}\big[\,|\,\Xi[k]/(N_A\Omega) - \hat{\xi}[k|k]\,|\,\big] && \text{(SRUKF)},\\
\bar{e}_n[k|k] &= \mathbb{E}\big[\,|\,X[k]/(N_A\Omega) - \hat{x}[k|k]\,|\,\big] && \text{(nSRUKF)},\\
\bar{e}_\mu[k] &= \mathbb{E}\big[\,|\,[\Xi[k] - \mu[k]\,|\,\big]/(N_A\Omega) && (\textit{a priori}).
\end{aligned}$$

Fig. 6 reports plots of the estimation errors $\bar{e}[k|k]$ and $\bar{e}_n[k|k]$ relative to $\bar{e}_\mu[k]$. In both scenarios, both SRUKF and nSRUKF improve upon the prior knowledge on $X[k]$ in at least some components. In Scenario 1, uncertainty about K_u is reduced around times $20-30$min, where the threshold is crossed by the increasing concentrations of intracellular arabinose, more markedly for SRUKF. This leads to a transient improvement of the estimation of arabinose concentration. For the remaining times, where the specific value of K_u is inessential (saturation of the nonlinearity), the contribution of filtering is not apparent. In Scenario 2, SRUKF clearly outperforms nSRUKF in the estimation of mGFP and GFP concentrations, i.e. the states more directly related to v_4. Overall, results show that exploiting the prior on extrinsic noise (parameter variability) not only enables estimation of the individual parameter value, but also improves estimation of unobserved states. This supports the use of ME-type models for state estimation and control applications.

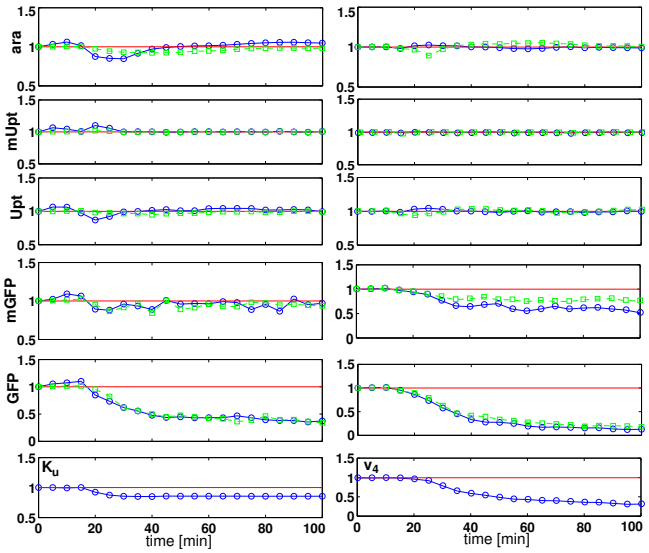


Fig. 6. Estimation error ratios $\bar{e}_i[k|k]/\bar{e}_{\mu,i}[k]$ (blue, solid) and $\bar{e}_{n,i}[k|k]/\bar{e}_{\mu,i}[k]$ (green, dashed) for the state entries $i = 1,\dots,5$ (top to bottom, first five rows) and $\bar{e}_6[k|k]/\bar{e}_{\mu,6}[k]$ (unknown parameter, last row). The smaller the value, the more accurate the state estimate. Reference value 1 (no improvement relative to prior knowledge) is indicated by a horizontal line (red). Left: Variability on K_u; Right: Variability on v_4.

VI. Conclusions

We investigated filtering of single-cell biochemical regulatory networks with intrinsic and extrinsic noise. Simulation results on a relevant case study show that approximation of the reference CME model via CLE is an appealing approach to construct practical real-time state estimators. Moreover, the use of prior information on parameter uncertainty led to improved estimation results, showing the potential of extrinsic noise modeling for state estimation and control applications. Directions of investigation include extensive performance comparisons and applications to real data.

Acknowledgment

We thank Prof. Michel Page for his help with *StochKit*.

References

[1] M. Thattai and A. van Oudenaarden, "Intrinsic noise in gene regulatory networks," *PNAS*, vol. 98, no. 15, p. 8614?8619, 2001.

[2] J. Paulsson, "Models of stochastic gene expression," *Physics of Life Reviews*, vol. 2, no. 2, pp. 157–175, 2005.

[3] H. E. Samad, M. Khammash, L. Petzold, and D. Gillespie, "Stochastic modelling of gene regulatory networks," *Intl J Rob Nonl Control*, pp. 691–711, 2005.

[4] M. Davidian and D. M. Giltinan, "Nonlinear models for repeated measurement data: An overview and update," *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 8, no. 4, pp. 387–419, 2003.

[5] J. Hasenauer, S. Waldherr, M. Doszczak, N. Radde, P. Scheurich, and F. Allgöwer, "Identification of models of heterogeneous cell populations from population snapshot data," *BMC Bioinformatics*, 2011.

[6] J. Paulsson, "Control, exploitation and tolerance of intracellular noise," *Nature*, vol. 420, pp. 231–237, 2002.

[7] J. Uhlendorf, A. Miermont, T. Delaveau, G. Charvin, F. Fages, S. Bottani, G. Batt, and P. Hersen, "Long-term model predictive control of gene expression at the population and single-cell levels," *PNAS*, Aug. 2012.

[8] A. Milias-Argeitis, S. Summers, J. Stewart-Ornstein, I. Zuleta, D. Pincus, H. El-Samad, M. Khammash, and J. Lygeros, "In silico feedback for in vivo regulation of a gene expression circuit," *Nature Biotechnology*, no. 29, pp. 1114–1116, 2011.

[9] D. Gillespie, "The chemical langevin equation," *Journal of Chemical Physics*, 2000.

[10] H. Singer, "Continuous-discrete unscented kalman filtering," FernUniversität Hagen, Germany, Tech. Rep. 384, 2006.

[11] S. Julier and J. Uhlmann, "Unscented filtering and nonlinear estimation," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 401–422, 2004.

[12] S. Sarkka, "On unscented kalman filtering for state estimation of continuous-time nonlinear systems," *Automatic Control, IEEE Transactions on*, vol. 52, no. 9, pp. 1631–1641, 2007.

[13] A. Doucet, N. de Freitas, and N. Gordon, Eds., *Sequential Monte Carlo Methods in Practice*. New York: Springer, 2001.

[14] J. Megerle, G. Fritz, U. Gerland, K. Jung, and J. Rädler, "Timing and dynamics of single cell gene expression in the arabinose utilization system," *Biophysical journal*, vol. 95, no. 4, pp. 2103–2115, 2008.

[15] C. Zechner, J. Ruess, P. Krenn, S. Pelet, M. Peter, J. Lygeros, and H. Koeppl, "Moment-based inference predicts bimodality in transient gene expression," *PNAS*, 2012.

[16] D. Wilkinson, *Stochastic Modelling for Systems Biology*. Chapman & Hall/CRC, 2006.

[17] C. Gillespie and A. Golighty, "Bayesian inference for the chemical master equation using approximate models," in *Proceedings of the WCSB*, ULM Germany, 2012.

[18] P. Kloeden, E. Platen, and H. Schurz, *Numerical solution of SDE through computer experiments*. Springer, 1994, vol. 1.

[19] S. Dana and S. Raha, "Physically consistent simulation of mesoscale chemical kinetics: The non-negative FIS-α method," *Journal of Computational Physics*, vol. 230, no. 24, pp. 8813–8834, 2011.

[20] K. R. Sanft, S. Wu, M. Roh, J. Fu, R. K. Lim, and L. R. Petzold, "Stochkit2: software for discrete stochastic simulation of biochemical systems with events," *Bioinformatics*, vol. 27, no. 17, pp. 2457–2458, 2011. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/27/17/2457.abstract

[21] D. Dochain, "State and parameter estimation in chemical and biochemical processes: a tutorial," *Journal of Process Control*, vol. 13, pp. 801–818, 2003.