bioRxiv preprint doi: https://doi.org/10.1101/564047; this version posted February 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Likelihood-free nested sampling for biochemical reaction networks

Jan Mikelson[1] and Mustafa Khammash [1]

[1]Department of Biosystems Science and Engineering, ETH Zurich, Switzerland.

The development of mechanistic models of biological systems is a central part of Systems Biology. One major challenge in developing these models is the accurate inference of the model parameters. In the past years, nested sampling methods have gained an increasing amount of attention in the Systems Biology community. Some of the rather attractive features of these methods include that they are easily parallelizable and give an estimation of the variance of the final Bayesian evidence estimate from a single run. Still, the applicability of these methods is limited as they require the likelihood to be available and thus cannot be applied to stochastic systems with intractable likelihoods. In this paper, we present a likelihood-free nested sampling formulation that gives an unbiased estimator of the Bayesian evidence as well as samples from the posterior. Unlike most common nested sampling schemes we propose to use the information about the samples from the final prior volume to aid in the approximation of the Bayesian evidence and show how this allows us to formulate a lower bound on the variance of the obtained estimator. We proceed and use this lower bound to formulate a novel termination criterion for nested sampling approaches. We illustrate how our approach is applied to several realistically sized models with simulated data as well as recently published biological data. The presented method provides a viable alternative to other likelihood-free inference schemes such as Sequential Monte Carlo or Approximate Bayesian Computations methods. We also provide an intuitive and performative C++ implementation of our method.

1 Introduction

The accurate modelling and simulation of biological processes such as gene expression or signalling has gained a lot of interest over the last years, resulting in a large body of literature addressing various types of models along with the means for their identification and simulation. The main purpose of these models is to qualitatively or quantitatively describe observed biological dynamics while giving insights into the underlying bio-molecular mechanisms.

bioRxiv preprint doi: https://doi.org/10.1101/564047; this version posted February 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

One important aspect in the design of these models is the determination of the model parameters. Often there exists a mechanistic model of the cellular processes, but their parameters (e.g. reaction rates or initial molecule concentrations) are largely unknown. Since the same network topology may result in different behaviour depending on the chosen parameters [26], this presents a major challenge for modelling and underscores the need for effective parameter estimation techniques.

The models used in Systems Biology can be coarsely classified into two groups: deterministic and stochastic models. Deterministic models usually rely on ordinary differential equations which, given the parameters and initial conditions, can describe the time evolution of the biological system in a deterministic manner. However, many cellular processes like gene expression are subject to random fluctuations [12, 36], which can have important biological functions [43, 49, 31] as well as contain useful information about the underlying molecular mechanisms [39]. The important role of stochastic fluctuations in biological systems has lead to increased interest in stochastic models and methods for their parameter inference[3, 25, 32, 41, 42, 56]. Such stochastic models are usually described in the framework of stochastic chemical reaction networks that can be simulated using Gillespie's Stochastic Simulation Algorithm (SSA) [17]. In recent years, the availability of single-cell trajectory data has drastically increased, providing detailed information about the (potentially stochastic) development of single cells throughout time.

Despite the increasing interest in stochastic systems, performing inference on them is still challenging and the available methods are computationally very demanding (see for instance [3, 20, 53]). Common algorithmic approaches for such cases include various kinds of sequential Monte Carlo methods (SMC) [9, 6], Markov Chain Monte Carlo (MCMC) methods [19, 3, 45], approximate Bayesian computation (ABC) methods [54, 32, 28], iterative filtering [27] and nested sampling (NS) approaches [52, 29, 37, 15]. Furthermore, to reduce computational complexity, several of these inference methods rely on approximating the model dynamics (for instance using the diffusion approximation [18] or linear noise approximation [11]). However, these approximations may not always be justifiable (in the case of low copy numbers of the reactants for example) and might obscure crucial system behaviour. One particular problem that is common to most inference methods is the usually high dimensional parameter space. Most of the sampling-based inference techniques require the exploration of the full parameter space, which is not an easy task as the dimension of the parameter space increases. In this paper, we focus on nested sampling

bioRxiv preprint doi: https://doi.org/10.1101/564047; this version posted February 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

methods and investigate its applicability to stochastic systems. Coming originally from the cosmology community, NS (originally introduced in [52]) has gained increasing popularity and found also applications in Systems Biology (see for instance [1, 5, 10, 29, 46]). Several implementations of NS are available ([14, 22]) and in [29] the authors even provide a NS implementation specifically for a Systems Biology context. Even though the original purpose of NS was to efficiently compute the Bayesian evidence, it has more and more become a viable alternative to MCMC methods for the approximation of the posterior (see for instance [16, 24]).

There are various reasons for the interest in NS which are discussed in detail in [40, 33] and the references within. Some of the rather appealing features of NS is that it performs well for multimodal distributions [16, 22], is easily parallelizable [23, 5] and provides a natural means to compute error bars on all of its results without needing multiple runs of the algorithm [51, 40]. For a comparison of MCMC and NS see for instance [46, 40], for a discussion of other methods to compute the Bayesian evidence using MCMC see [40, 34]. Like standard MCMC methods, NS requires the availability of the likelihood $l(\theta)$ which limits its use to models that allow for the computation of the likelihood such as deterministic models and simple stochastic models. In this paper, we consider an extension to the original NS framework that, similarly to the particle MCMC method [55] and particle SMC [2], allows the use of approximated likelihoods instead of the actual likelihood to be used for NS. In the following we introduce the notation and problem formulation, in section 2 we revisit the basic NS idea and outline some of its features. Section 3 is dedicated to the likelihood-free NS formulation and in section 4 we demonstrate its performance on several chosen examples.

1.1 Chemical Reaction Networks

We are considering a n_x-dimensional Markov Process $X(t)$ depending on a d-dimensional parameter vector θ. We denote with $X_i(t)$ the i^{th} entry of the state vector at time t and with $X(t) = \{X_i(t)\}_{i=1,\dots,n_x}$ the state vector at time t. We will write $X_\tau = X(t_\tau)$ when talking about the state vector at a timepoint t_τ indexed with τ.

In the context of stochastic chemical reaction networks this Markov process describes the abun-

bioRxiv preprint doi: https://doi.org/10.1101/564047; this version posted February 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

dances of n_x species $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_{n_x}$, reacting through n_R reactions $\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_{n_R}$ written as

$$\mathcal{R}_j : \qquad \sum_{i=1}^{n_x} p_{ji} \boldsymbol{X}_i \to \sum_{i=1}^{n_x} q_{ji} \boldsymbol{X}_i,$$

where p_{ji} is the numbers of molecules of species $\boldsymbol{X}_i$ involved in reaction R_j, and q_{ji} is the number of molecules of species $\boldsymbol{X}_i$ produced by that reaction. The random variable $X_i(t)$ corresponds to the number of molecules of species $\boldsymbol{X}_i$ at time t. Each reaction $\mathcal{R}_j$ has an associated propensity. The reaction propensities at a given time t depend on the current state $X(t)$ and on a d-dimensional parameter vector θ.

1.2 General Task

The process $X(t)$ is usually not directly observable but can only be observed indirectly through a n_y-dimensional observation vector

$$Y_\tau \sim p(\cdot|X_\tau, \theta),$$

which depends on the state X_τ and on the d-dimensional parameter vector $\theta \in \Omega$ where $\Omega \subseteq \mathbb{R}^d$ denotes the parameter space. We shall assume that the variable Y is not observed at all times but only on T timepoints $t_1, \ldots, t_T$ and only for M different trajectories. With $\boldsymbol{y}$ we denote the collection of observations at all time points. In the Bayesian approach the parameter vector θ is treated as a random variable with associated prior $\pi(\theta)$. The goal is not to find just one set of parameters, but rather to compute the posterior distribution $\mathcal{P}(\theta|\boldsymbol{y})$ of θ

$$\mathcal{P}(\theta|\boldsymbol{y}) = \frac{1}{Z} l(\boldsymbol{y}|\theta)\pi(\theta),$$

where $l(\boldsymbol{y}|\theta)$ (we will also write $l(\theta)$ if the dependence on $\boldsymbol{y}$ is clear from the context) is the likelihood of θ for the particular observation $\boldsymbol{y}$ and Z is the Bayesian evidence

$$Z = \int_\Omega l(\boldsymbol{y}|\theta) d\pi(\theta). \tag{1.1}$$

bioRxiv preprint doi: https://doi.org/10.1101/564047; this version posted February 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

This has several advantages over a single point estimate as it gives insight into the areas of the parameter space resulting in model behaviour similar to the observations as well as about their relevance for the simulation outcome (a wide posterior indicates non-identifiability for example). For a detailed discussion of Bayesian approaches see for instance [34]. In this paper we follow the Bayesian approach and aim to recover the posterior $\mathcal{P}(\theta|\boldsymbol{y})$. In the following section we briefly outline the basic nested sampling approach.

2 Nested Sampling

Nested sampling is a Bayesian inference technique that was originally introduced by John Skilling in [52] to compute the Bayesian evidence 1.1. NS can be viewed as an importance sampling technique (as for instance discussed in [47]) as it approximates the evidence by generating samples θ_i, weights w_i and likelihoods $l_i = l(\theta_i)$ such that the weighted samples (θ_i, w_i) can be used to obtain numerical approximations of a function f over the prior π

$$\sum_i w_i f(\theta_i) \approx \int f(\theta) d\pi(\theta). \tag{2.2}$$

To compute an approximation $\widehat{Z}$ of the Bayesian evidence 1.1, f is chosen to be the likelihood function l

$$\widehat{Z} = \sum_i w_i l_i \approx \int l(\theta) d\pi(\theta). \tag{2.3}$$

The points θ_i are sampled from the prior distribution constrained to super level sets of the likelihood corresponding to an increasing sequence of thresholds. In this sense it can also be viewed as a sequential Monte Carlo method, where the intermediate distributions are the nested super level sets of the likelihood. This way, samples from NS are concentrated around the higher regions of the likelihood. One can also use the weights $l_i \times w_i$ instead of w_i to approximate functions over the posterior $\mathcal{P}(\theta)$

$$\frac{1}{\widehat{Z}} \sum f(\theta_i) l_i w_i \approx \int f(\theta) d\mathcal{P}(\theta).$$

bioRxiv preprint doi: https://doi.org/10.1101/564047; this version posted February 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2.1 NS algorithm

In the following we briefly outline the NS algorithm. First, a set $\mathcal{L}_0$ of N "live" particles $\{\theta^i\}_{i=1,\dots,N}$ is sampled from the prior π

$$\theta^i \sim \pi(\theta)$$

and their likelihoods $l_i = l(\theta^i)$ are computed. Then the particle with the lowest likelihood

$$\theta_1 = \arg\min \{l(\theta) | \theta \in \mathcal{L}_0\}$$

gets removed from the set of live particles and saved together with its likelihood

$$\epsilon_1 := l(\theta_1)$$

in a set of "dead" particles $\mathcal{D}$. A new particle $\theta^\star$ is then sampled from the prior under the constraint that its likelihood is higher than ϵ_1

$$\theta^\star \sim \pi(\theta | l(\theta) > \epsilon_1). \tag{2.4}$$

This particle is combined with the remaining particles of $\mathcal{L}_0$ to form a new set of live particles $\mathcal{L}_1$ that are now distributed according to the constrained prior $\pi(\theta | l(\theta) > \epsilon_1)$, which we denote as

$$\mathcal{L}_1 \sim \pi(\theta | l(\theta) > \epsilon_1).$$

This procedure is repeated until a predefined termination criteria is satisfied. The result is a sequence of dead points θ_i with corresponding likelihoods ϵ_i that are concentrated in the regions of high likelihood. The Nested Sampling procedure is shown in Algorithm 1.

bioRxiv preprint doi: https://doi.org/10.1101/564047; this version posted February 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Given observations $\boldsymbol{y}$ and a prior $\pi(\theta)$ for θ.
Sample N particles θ^k from the prior π and save in the set $\mathcal{L}_0$, set $\mathcal{D} = \{\emptyset\}$
for i = 1, 2, . . . , m **do**
  Set $\theta_i = \arg\min \{l(\theta) | \theta \in \mathcal{L}_{i-1}\}$ and $\epsilon_i = l(\theta_i)$
  Add $\{\theta_i, \epsilon_i\}$ to $\mathcal{D}$
  Set $\mathcal{L}_i = \mathcal{L}_{i-1} \backslash \theta_i$
  Sample $\theta^\star \sim \pi(\theta | l(\theta) > \epsilon_i)$ and add it to $\mathcal{L}_i$
end for

Algorithm 1: Nested sampling algorithm

2.2 Approximating the Bayesian Evidence

Nested sampling exploits the fact that the Bayesian evidence 1.1 can also be written[1] (see [52]) as a one dimensional integral

$$Z = \int_0^1 L(x)dx,$$

over the prior volume

$$x(\epsilon) := \pi(l(\theta) > \epsilon) = \int_{l(\theta)>\epsilon} d\pi(\theta),$$

where $L(x)$ denotes the likelihood corresponding to the constrained prior with volume x

$$L(x) = \arg\inf_{\epsilon} \{x(\epsilon) \geq x\}. \tag{2.5}$$

We have visualized these quantities on a simple example with a uniform prior on $[0, 1]$ in Figure 1.

The sampling scheme of nested sampling provides a sequence of likelihoods $\epsilon_1 < \epsilon_2 < \ldots < \epsilon_m$, but their corresponding prior volumes $x(\epsilon_i)$ are not known. However, since the ϵ_i are obtained by iteratively removing the lowest likelihood of N uniformly distributed points on the constrained prior $\pi(\theta | l(\theta) > \epsilon_{i-1})$, the prior volume $x(\epsilon_i)$ can be written as

$$x_i := x(\epsilon_i) = t^{(i)} x_{i-1},$$

[1] for this to hold some weak conditions have to be satisfied, see for details [8] and [15]

bioRxiv preprint doi: https://doi.org/10.1101/564047; this version posted February 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

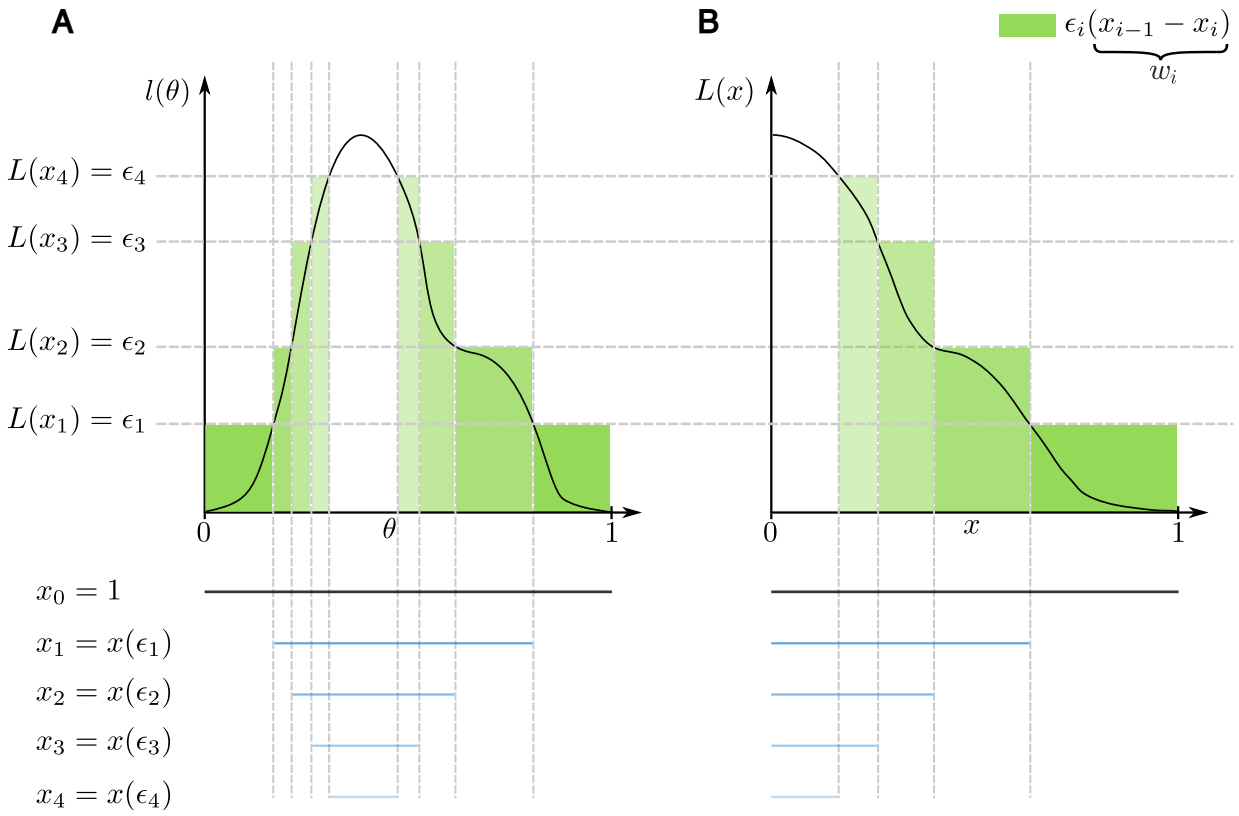

Figure 1: Illustration of the nested sampling approximation with a uniform prior on $[0, 1]$. **A:** The integral over the parameter space $\int_\Omega l(\theta)d\theta$. **B:** The transformed integral $\int_0^1 L(x)dx$ over the prior volume x.

bioRxiv preprint doi: https://doi.org/10.1101/564047; this version posted February 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

where each $t^{(i)}$ is an independent sample of the random variable t which is distributed as the largest of N uniform random variables on the interval $[0, 1]$ and $x_0 = 1$ (For further justification and discussion on this see [52, 15, 8] and the references within). The values $t^{(i)}$ are not known and need to be estimated. Since their distribution is known[2], they can be approximated by their means $\mathbb{E}(t) = \frac{N}{N+1}$ (or by the mean of their logs $\mathbb{E}(\log(t)) = -\frac{1}{N}$), and thus the i^{th} prior volume can be approximated as

$$\widehat{x}_i = \left(\frac{N}{N+1}\right)^i \approx x_i. \tag{2.6}$$

With these prior volumes one can compute the importance weights w_i in equation 2.2 and 2.3 for each of the dead particles θ_i as

$$w_i = (\widehat{x}_{i-1} - \widehat{x}_i)\,. \tag{2.7}$$

These weights correct for the fact that the samples in $\mathcal{D}$ are not drawn uniformly from the prior, but are concentrated in areas of high likelihood. We note that to integrate a function on the parameter space Ω over the prior π, as in equations 2.2, only these weights are needed. To approximate Z, NS uses these weights to integrate the likelihood function $l(\theta)$ over the prior

$$Z = \int_0^1 L(x)dx \approx \sum_{i=1}^{m} L(x_i)(\widehat{x}_{i-1} - \widehat{x}_i) = \sum_{i=1}^{m} \epsilon_i w_i =: \widehat{Z}_{\mathcal{D}}^m \tag{2.8}$$

where m is the number of performed NS iterations and the subscript $\mathcal{D}$ in $\widehat{Z}_{\mathcal{D}}^m$ emphasizes that for NS the evidence estimate is obtained using only the dead points in $\mathcal{D}$. The justification for these weights as well as an in depth discussion and error approximation can be found in [8, 24, 30] and the references therein. This basic idea of nested sampling has seen several modifications and improvements over the years, along with in-depth discussions of various sampling schemes for the constrained prior [14, 22], parallel formulations [22, 23, 5] and several implementations [14, 22, 29].

[2] $t \sim \mathcal{B}(N, 1)$ with $\mathcal{B}(a, b)$ being the Beta distribution with parameters a and b.

bioRxiv preprint doi: https://doi.org/10.1101/564047; this version posted February 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2.3 Termination of NS

Assuming that the distribution 2.4 can be efficiently sampled, each iteration of the NS scheme has the same computational complexity (the computationally most expensive step is usually to sample $\theta^\star \sim \pi(\theta|l(\theta) > \epsilon_i)$ and computing its likelihood). The NS algorithm is usually run until the remaining prior volume multiplied by the highest likelihood in this volume is smaller than a predefined fraction of the current BE estimate (see [52]). We write this quantity as

$$\Delta^m_{\max} := \widehat{x}_m \max_{\theta \in \mathcal{L}_m}(l(\theta)) \frac{1}{\widehat{Z}^m_{\mathcal{D}}}.$$

Some other termination criteria have been suggested (for instance in [22]), but since the prior volume decreases exponentially with the number of NS iterations and each iteration takes the same computational time, the choice of the particular termination criterion is not critical.

2.4 Parallelization of NS

The parallelization of NS can be done in a very straight forward manner. Still several different parallelization schemes have been suggested in [22, 23, 5] (for a short overview see section S1). We use a parallelization scheme similar to the one presented in [23], where at each iteration not only the one particle with the lowest likelihood is resampled, but the r lowest particles. The resulting parallel scheme is outlined in Algorithm 2. With r parallel particles the final approximation 2.8 changes to

$$\widehat{Z}^m_{\mathcal{D}} = \sum_{i=1}^{m} \sum_{j=1}^{r} \epsilon_{i,j}(\widehat{x}_{i,j-1} - \widehat{x}_{i,j}), \tag{2.9}$$

with $x_{i,j} = t^{(i)}_j x_{i-1,r}$ and $t^{(i)}_j$ being i^{th} sample of t_j which is the j^{th} largest number among N uniform numbers between 0 and 1 [3] (with the obvious boundary condition $x_{0,r} = 1$). We note that this is slightly different than the parallelization scheme presented in [22, 23, 5], for a brief discussion see S1.

[3] This means $t_j \sim \mathcal{B}(N - j + 1, j)$

bioRxiv preprint doi: https://doi.org/10.1101/564047; this version posted February 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1: Given observations $\boldsymbol{y}$ and a prior $\pi(\theta)$ for θ.
2: Sample N particles θ^k from the prior π and save them in the set $\mathcal{L}_0$, set $\mathcal{D} = \{\emptyset\}$
3: **for** i = 1, 2, . . . , m **do**
4: **for** j = 1, 2, . . . , r **do**
5: Set $\theta_{i,j} = \arg\min \{l(\theta)|\theta \in \mathcal{L}_{i-1}\}$ and $\epsilon_{i,j} = l(\theta_{i,j})$
6: Add $\{\theta_{i,j}, \epsilon_{i,j}\}$ to $\mathcal{D}$
7: remove $\theta_{i,j}$ from $\mathcal{L}_{i-1}$
8: **end for**
9: Set $\mathcal{L}_i = \mathcal{L}_{i-1}$
10: **for** j = 1, 2, . . . , r **do**
11: Sample $\theta^\star \sim \pi(\theta|l(\theta) > \epsilon_{i,r})$ and add it to $\mathcal{L}_i$
12: **end for**
13: **end for**

Algorithm 2: Parallel nested sampling algorithm. The samples drawn in line 11 are all independent and thus can be drawn in parallel.

3 Likelihood-free nested sampling (LF-NS)

In many cases (such as most of the above mentioned stochastic models) the likelihood $l(\theta)$ cannot be directly computed, making approaches like MCMC methods or nested sampling not applicable. Fortunately, many variations of MCMC have been described circumventing this problem by introducing likelihood-free MCMC methods such as [35] or [55] as well as other likelihood-free methods such as ABC [54] or likelihood-free sequential Monte Carlo (SMC) methods [50]. These approaches usually rely on forward simulation of a given parameter vector θ to obtain a simulated data set that can then be compared to the real data or can be used to compute a likelihood approximation $\widehat{l}(\theta) \approx l(\theta)$. In the following we briefly illustrate one way to approximate the likelihood.

3.1 Likelihood approximation using particle filters

A common way to approximate the likelihood through forward simulation is using a particle filter (see for instance [44] or [19]), which iteratively simulates the stochastic system with H particles and then resamples these particles. In the following we illustrate such a particle filter likelihood approximation on a simple birth death model, where one species (mRNA) is produced at rate $k = 1$ and degrades at rate $\gamma = 0.1$. We simulated one trajectory (shown in Figure 2 A) of this system using Gillespies stochastic simulation algorithm (SSA [17]) and, using the finite state projection

bioRxiv preprint doi: https://doi.org/10.1101/564047; this version posted February 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

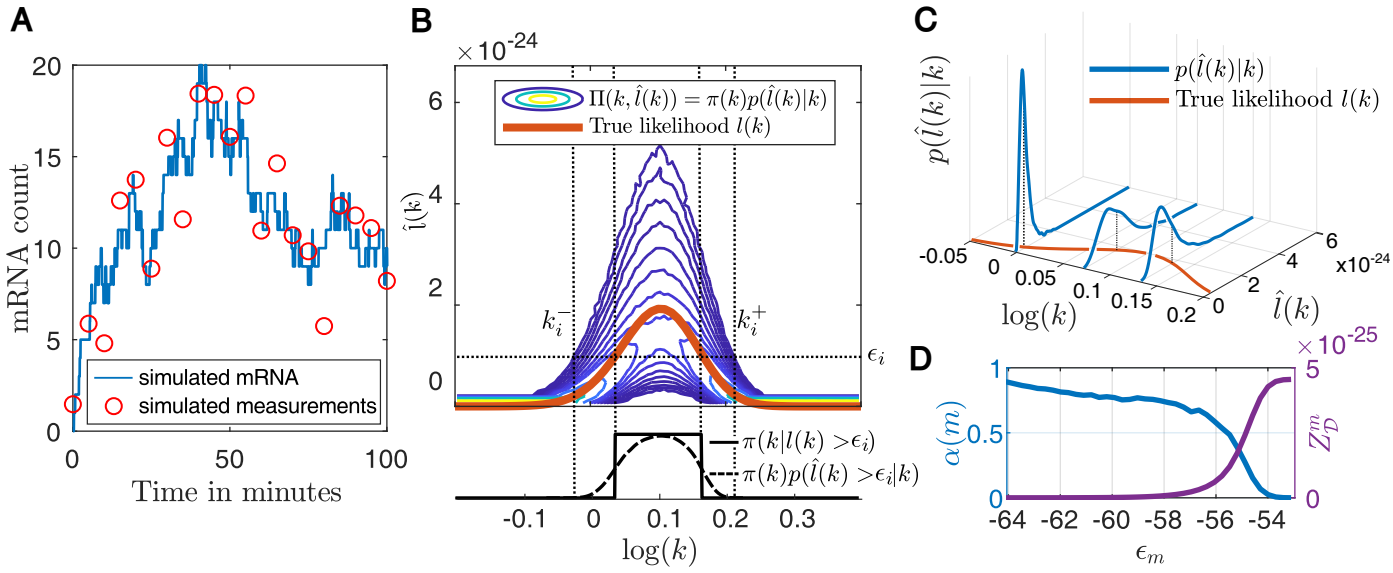

Figure 2: **A:** A simulated trajectory of the birth death system using k = 1 and $\gamma = 0.1$ with 21 equally spaced measurements (taken to be normally distributed around the mRNA count with $\sigma = 2$).**B:** Top: Likelihood for different parameters k (red) and contour lines of the joint distribution $\Pi(k, \log(k)$ of the parameter k and its likelihood approximation $\widehat{l}(k)$, based on 10^6 samples of the likelihood approximation obtained with a particle filter with 100 particles. Bottom: The constrained priors $\pi(k|l(k) > \epsilon)$ and $\pi(k)p(\widehat{l}(k) > \epsilon|k)$ for $\epsilon = 1e-24$. **C:** Example distributions $p(\widehat{l}(k)|k)$ (blue) for $k = 1$, 1.2 and 1.4 and the true likelihood $l(k)$ red. **D:** blue: The ratio $\alpha(m)$ of the probability masses of $\Pi(k, \widehat{l}(k)|p(l(k) > \epsilon_i) > k)$ above and below each likelihood threshold ϵ_i, constrained to those regions of k with $p(\widehat{l}(k) > \epsilon_i) > 0$ (these are the parameter regions in panel B between k_i^- and k_i^+). purple: The evidence as estimated by all particles with likelihood below ϵ_m: $Z_{\mathcal{D}}^m = \int_{\widehat{l}(k) \leq \epsilon_m} \widehat{l}(k) d\Pi(k, \widehat{l}(k))$.

(FSP [38]), computed the likelihood $l(k)$ for different values of k while keeping γ fixed to 0.1. The true likelihood for different k is shown as the solid red line in Figure 2 B and C. We also illustrated the likelihood approximation $\widehat{l}(k)$ using a particle filter ([19]) with $H = 100$ particles for three values of k. For each of the values for k we computed 1000 realizations of $\widehat{l}(k)$ and plotted the empirical distributions in Figure 2 C. Note that $\widehat{l}(k)$ is itself a random variable with distribution $p(\widehat{l}(k)|k)$ and has a mean equal to the true likelihood $\mathbb{E}(\widehat{l}(k)) = l(k)$ (see for instance [44]). We also sampled 10^6 values of k from a log uniform prior and approximate for each k its likelihood with the same particle filter with $H = 100$ particles. We plotted the contour lines of this joint distribution

$$\Pi(k, \widehat{l}(k)) = \pi(k)p(\widehat{l}(k)|k)$$

in Figure 2 B.

In the following we discuss how to utilize such a likelihood approximation to apply the above described NS procedure to cases where the likelihood is not available. Throughout the paper we

bioRxiv preprint doi: https://doi.org/10.1101/564047; this version posted February 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

assume that the likelihood approximation $\widehat{l}(\theta)$ is obtained using a particle filter, but our result hold for any unbiased likelihood estimator.

3.2 The LF-NS scheme

From here on we assume that the true likelihood $l(\theta)$ is not available, but a realization $\widehat{l}(\theta)$ of the approximated likelihood having the distribution

$$\widehat{l}(\theta) \sim p(\widehat{l}(\theta)|\theta)$$

with

$$\int \widehat{l}(\theta) dp(\widehat{l}(\theta)|\theta) = l(\theta)$$

can be computed.

For NS, the constraint prior $\pi(\theta|l(\theta) > \epsilon_i)$ needs to be sampled. Since in the likelihood-free case, the likelihood $l(\theta)$ is not available and $\widehat{l}(\theta)$ is itself a random variable, the set $\{\theta \in \Omega|\widehat{l}(\theta) > \epsilon_i\}$ (which is the support of the constrained prior) is not defined. To apply the NS idea to the likelihood-free case, we propose to perform the NS procedure on the joint prior

$$\Pi(\theta, \widehat{l}(\theta)) = \pi(\theta)p(\widehat{l}(\theta)|\theta) \tag{3.10}$$

on the set $\Omega \times \mathbb{R}_{>0}$. This joint prior can be sampled by drawing a sample $\theta^\star$ from the prior $\pi(\theta)$ and then drawing one sample $\widehat{l^\star}$ from the distribution of likelihood approximations $p(\widehat{l}(\theta^\star)|\theta^\star)$. With such a sampling scheme we perform the NS steps of constructing the set of "dead" particles $\mathcal{D}$ on the joint prior 3.10. As in standard NS, we sample a set of N "live" particles $\{\theta, \widehat{l}\}$ from $\Pi(\theta, \widehat{l}(\theta))$, then we iteratively remove the particle $\{\theta, \widehat{l}\}$ with the lowest likelihood sample $\widehat{l}$ from the set of live points and add it to the dead points. The LF-NS algorithm is shown in Algorithm 3.

The parallel version of LF-NS is analogous to the parallelization of the standard NS algorithm in Algorithm 2.

bioRxiv preprint doi: https://doi.org/10.1101/564047; this version posted February 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1: Given observations $\boldsymbol{y}$, a prior $\pi(\theta)$ for θ and a likelihood approximation $p(\widehat{l}(\theta)|\theta)$.
2: Sample N particles $\{\theta^k, \widehat{l}^k\}$ from the prior $\Pi(\theta, \widehat{l}(\theta))$ and save it in the set $\mathcal{L}_0$, set $\mathcal{D} = \{\emptyset\}$
3: **for** i = 1, 2, . . . , m **do**
4: Find $i' = \arg\min_k \left(\widehat{l}^k | \{\theta^k, \widehat{l}^k\} \in \mathcal{L}_{i-1}\right)$ and set $\theta_i = \theta^{i'}$ and $\epsilon_i = l^{i'}$
5: Add $\{\theta_i, \epsilon_i\}$ to $\mathcal{D}$
6: Set $\mathcal{L}_i = \mathcal{L}_{i-1} \backslash \{\theta_i, \epsilon_i\}$
7: Sample $\{\theta^\star, \widehat{l}^\star\} \sim \Pi(\theta, \widehat{l}(\theta) | \widehat{l}(\theta) > \epsilon_i)$ and add it to $\mathcal{L}_i$
8: **end for**

Algorithm 3: Likelihood-free nested sampling algorithm

3.3 LF-NS is unbiased

As for standard NS, the sampling procedure for LF-NS guarantees that each set of live points $\mathcal{L}_i$ contains N samples uniformly distributed according to the constrained joint prior $\Pi(\theta, \widehat{l}(\theta) | \widehat{l}(\theta) > \epsilon_i)$, thus removing the sample with the lowest likelihood approximation $\widehat{l}^k$ results in the same shrinkage of prior volume as the standard NS scheme. The prior volumes $x_i = t^{(i)} x_{i-1}$ now correspond to the volumes of the constraint joint priors $\Pi(\theta, \widehat{l}(\theta) | \widehat{l}(\theta) > \epsilon_i)$ and the resulting weights $w_i = \widehat{x}_{i-1} - \widehat{x}_i$ can be used, similarly as in equation 2.2, to integrate functions f over the constrained prior

$$\sum_{i=1}^{m} w_i f(\theta_i, \epsilon_i) \approx \int f(\theta, \widehat{l}(\theta)) d\Pi(\theta, \widehat{l}(\theta).$$

Using $f(\theta_i, \epsilon_i) = \epsilon_i$ we can use this to approximate the Bayesian Evidence

$$\sum_{i}^{m} w_i \epsilon_i \approx \int \widehat{l}(\theta) d\Pi(\theta, \widehat{l}(\theta))$$

$$= \int_{\Omega} \int_{0}^{\infty} \widehat{l}(\theta) \pi(\theta) p(\widehat{l}(\theta)|\theta) d\widehat{l}(\theta) d\theta = \int_{\Omega} \pi(\theta) \int_{0}^{\infty} \widehat{l}(\theta) p(\widehat{l}(\theta)|\theta) d\widehat{l}(\theta) d\theta = \int_{\Omega} l(\theta) \pi(\theta) d\theta = Z,$$

where the last equality relies on the unbiasedness of $\widehat{l}(\theta)$.

While the procedure for LF-NS is very similar to the standard NS algorithm, the new samples $\theta^\star$ have to be drawn from the constraint joint prior $\Pi(\theta, \widehat{l}(\theta) | \widehat{l}(\theta) > \epsilon)$ instead from the constrained prior $\pi(\theta | l(\theta) > \epsilon)$. In the following we discuss the resulting difficulties and show how to overcome

bioRxiv preprint doi: https://doi.org/10.1101/564047; this version posted February 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

them.

3.4 Sampling from the super-level sets of the likelihood

One of the main challenges [7, 40, 4] in the classical NS algorithm is the sampling from the prior constrained to higher likelihood regions $\pi(\theta|l(\theta) > \epsilon)$. A lot of effort has been dedicated to find ways to sample from the constraine prior efficiently, the most popular approaches include slice sampling [22] and ellipsoid based sampling [16].

In the case of LF-NS, at the i^{th} iteration we are sampling not just a new parameter vector $\theta^\star$ but also a realization of its likelihood approximation $\widehat{l}^\star$ from

$$\Pi(\theta, \widehat{l}(\theta)|\widehat{l}(\theta) > \epsilon_i) = \pi(\theta)p(\widehat{l}(\theta)|\theta, \widehat{l}(\theta) > \epsilon_i). \tag{3.11}$$

Since it is in general not possible to sample $\widehat{l}^\star$ from the constraint distribution $p(\widehat{l}(\theta)|\theta, \widehat{l}(\theta) > \epsilon_i)$ directly, we sample $\theta^\star$ from the prior $\pi(\theta)$, then sample $\widehat{l}^\star$ from the unconstrained distribution $p(\widehat{l}(\theta^\star)|\theta^\star)$ and accept the pair $(\theta^\star, \widehat{l}^\star)$ only if $\widehat{l}^\star > \epsilon_i$. While this procedure guarantees that the resulting samples are drawn from 3.11, the acceptance rate might become very low. Each live set $\mathcal{L}_i$ consists of N pairs $(\theta^k, \widehat{l}^k)$ distributed according to 3.11, thus the parameter vectors θ^k in $\mathcal{L}_i$ are distributed according to

$$\theta \sim \int \Pi(\theta, \widehat{l}(\theta)|\widehat{l}(\theta) > \epsilon_i) d\widehat{l}(\theta) = \pi(\theta)p(\widehat{l}(\theta) > \epsilon_i). \tag{3.12}$$

We plotted an example of the distributions 2.4 and 3.12 for the example of the birth-death process in Figure 2 B. The distribution 3.12 has usually an infinite support, although in practice 3.12 will be close to zero for large areas of the parameter space Ω. Similarly to NS, we propose to use the set $\mathcal{L}_i$ to drawn from the areas where 3.12 is non zero. Slice sampling methods ([1, 22]) are unfortunately not applicable for LF-NS since they require a way to evaluate its target distribution at each of its samples. We can still use ellipsoid sampling schemes, but unlike in the case of NS where the target distribution $\pi(\theta|l(\theta) > \epsilon)$ has compact support, the target distribution for LF-NS 3.12 has

bioRxiv preprint doi: https://doi.org/10.1101/564047; this version posted February 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

potentially infinite support framing ellipsoid based sampling approaches rather unfitting. Sampling using MCMC methods (as suggsted in [52]) is expected to work even for target distributions with infinite support, but suffer from the known MCMC drawbacks, as they produce correlated samples and might get stuck in disconnected regions.

To account for the smooth shape of 3.12 we propose to employ a density estimation approach. At each iteration i, we estimate the density $\pi(\theta)p(\widehat{l}(\theta) > \epsilon_i)$ from the live points and employ a rejection sampling approach to sample uniformly from the prior on the domain of this approximation. As density estimation technique, we use Dirichlet Process Gaussian Mixture Model (DP-GMM) [21], which approximates the distribution $\pi(\theta)p(\widehat{l}(\theta) > \epsilon_i)$ with a mixture of Gaussians. DP-GMM uses a hierarchical prior on the mixture model and assumes that the mixture components are distributed according to a Dirichlet Process. The inference of the distribution is an iterative process that uses Gibbs sampling to infer the number and shape of the Gaussians as well as the parameters and hyper parameters of the mixture model. DP-GMM estimations perform comparably well with sparse and high dimensional data and are less sensitive to outliers. Further, since we employ a parallelized LF-NS scheme, the density estimation has to be performed only after the finish of each parallel iteration, making the computational effort of density estimations negligible compared to the computational effort for the likelihood approximation. For a detailed comparison between DP-GMM and kernel density estimation and a further discussion of DP-GMM see [21], for an illustration of DP-GMM, KDE and ellipsoid sampler see section S2. Even though for the presented examples we employ DP-GMM, we note that in theory any sampling scheme that samples uniformly from the prior $\pi(\theta)$ over the support of $\pi(\theta)p(\widehat{l}(\theta) > \epsilon_i)$ will work.

3.5 A lower bound on the estimator variance

Unlike for NS, for LF-NS, even if at each iteration the proposal particle $\theta^\star$ is sampled from the support of $\pi(\theta)p(\widehat{l}(\theta) > \epsilon_i)$, it will only be accepted with probability $p(\widehat{l}(\theta^\star) > \epsilon_i)$. This means that depending on the variance of the likelihood estimation $p(\widehat{l}(\theta)|\theta)$ and the current likelihood threshold ϵ_i the acceptance rate for LF-NS will change and with it the computational cost. We illustrated this on the example for the birth-death model above. For each of the 10^6 samples $\{k_i, \widehat{l}_i\}$ from $\Pi(k, \widehat{l}(k))$

bioRxiv preprint doi: https://doi.org/10.1101/564047; this version posted February 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

we set $\epsilon_i = l_i$ and considered the particles $k_i^- = \min(k_j : l_j \geq \epsilon_i)$ and $k^+ = \max(k_j : l_j \geq \epsilon_i)$ (illustrated in Figure 2 B). The particles $\{k_j\}$ in between k_i^- and k_i^+ give a numerical approximation of the support of $\pi(\theta)p(\widehat{l}(\theta) > \epsilon_i)$. We denote with S_i^+ all the pairs $\{k_j, \widehat{l}_j\}$ with k_j between k_i^- and k_i^+ with a likelihood above ϵ_i and with S_i^- the pairs with a likelihood below ϵ_i

$$S^+ = \{j : l_j > \epsilon_i, k_i^+ \geq k_j \geq k_i^-\} \quad \text{and} \quad S^- = \{j : l_j \leq \epsilon_i, k_i^+ \geq k_j \geq k_i^-\}$$

and computed the ratio of the number of their element

$$\alpha(m) = \frac{|S_m^+|}{|S_m^-| + |S_m^+|}.$$

The values of $\alpha(m)$ give us an idea what the acceptance rate for LF-NS looks like in the best case where the new particles $k^\star$ are sampled from the support of $\pi(k)p(\widehat{l}(k) > \epsilon_m)$. We plotted $\alpha(m)$ in Figure 2 D as well as the evidence $Z_{\mathcal{D}}^m = \int_{\widehat{l}(k) \leq \epsilon_m} \widehat{l}(k)\Pi(k, \widehat{l}(k))$. We see that $\alpha(m)$ decreases to almost zero as $Z_{\mathcal{D}}^m$ approaches Z. The shape of α_m will in general be dependent on the variance of the likelihood approximation $\widehat{l}(\theta)$. For a further discussion on the acceptance rate for different particle filter settings see section S3.

Due to this possible increase in computational time, it is important to terminate the LF-NS algorithm as soon as possible. We propose to use for the Bayesian evidence estimation not only the dead particles $\mathcal{D}$, but also the current live points $\mathcal{L}_m$. This possibility has been already mentioned in other places (for instance in [8, 24, 30]) but is usually not applied, since the contribution of the live particles decreases exponentially with the number of iterations.[4] Since for standard NS each iteration is expected to take the same amount of time, most approaches simply increase the number of iterations to make the contribution of the live particles negligibly small.

[4]We point out that while in classical nested sampling the contribution of the live points can indeed be made arbitrarily small, the resulting estimator (employing only the dead points) is strictly speaking not unbiased since it approximates the Bayesian evidence not over the full prior volume but only up to the final x_m, which is the quantity $Z_{\mathcal{L}}^m$ in equation 3.13

bioRxiv preprint doi: https://doi.org/10.1101/564047; this version posted February 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The Bayesian evidence can be decomposed as

$$Z = \underbrace{\int_0^{x_m} L(x)dx}_{=:Z_{\mathcal{L}}^m} + \underbrace{\int_{x_m}^1 L(x)dx}_{=:Z_{\mathcal{D}}^m} \tag{3.13}$$

where x_m is the prior volume for iteration m. The first integral $Z_{\mathcal{L}}^m$ is the part that can be approximated through the N live samples at any given iteration, while the integral $Z_{\mathcal{D}}^m$ is approximated through the dead samples. Writing

$$\overline{L}_m := \frac{1}{N} \sum_{\{\theta,\widehat{l}\}\in\mathcal{L}_m} \widehat{l} \approx \int \widehat{l}(\theta) d\Pi(\theta, \widehat{l}(\theta) | \widehat{l}(\theta) > \epsilon_m)$$

for the estimator of the integral of the likelihoods in the live set, we propose the following estimator for Z

$$\widehat{Z}_{\text{tot}}^m = \underbrace{\sum_{i=1}^m \epsilon_i w_i}_{=\widehat{Z}_{\mathcal{D}}^m \approx Z_{\mathcal{D}}^m} + \underbrace{\widehat{x}_m \overline{L}_m}_{=\widehat{Z}_{\mathcal{L}}^m \approx Z_{\mathcal{L}}^m}, \tag{3.14}$$

where $\widehat{Z}_{\mathcal{D}}^m$ approximates the finite sum $\widetilde{Z}_{\mathcal{D}}^m = \sum_{i=1}^m \epsilon_i (x_{i-1} - x_i)$ by replacing the random variables x_i with their means $\widehat{x}_i$. Since $\widehat{x}_m \overline{L}_m$ is an unbiased estimator of $Z_{\mathcal{L}}^m$ and $\widehat{Z}_{\mathcal{D}}^m$ is an unbiased estimator of $Z_{\mathcal{D}}^m$, the estimator $\widehat{Z}_{\text{tot}}^m$ is an unbiased estimator of the Bayesian evidence Z for any m. In particular, this implies that terminating the LF-NS algorithm at any iteration m will result in an unbiased estimate for Z. However, terminating the LF-NS algorithm early on will still result in a very high variance of the estimator. Since the error of replacing the integral Z with the finite sum $\widetilde{Z}_{\mathcal{D}}^m$ is negligible compared to the error resulting from replacing x_i with $\widehat{x}_i$ (see [13] or [8]), this variance is a result of the variances in x_i and the variance in the Monte Carlo estimate $\overline{L}_m$.[5] In the following we formulate a lower bound σ_{min}^{2m} on the estimator variance $\sigma_{\text{tot}}^{2m} = \text{Var}\left(\widetilde{Z}_{\mathcal{D}}^m + \widehat{Z}_{\mathcal{L}}^m\right)$ at iteration m, show that this lower bound is monotonically increasing in m and propose to terminate the LF-NS

[5] As pointed out in [24], when using nested sampling approximations to approximate the integral of arbitrary functions f over the posterior, an additional error is introduced by approximating the average value of $f(\theta)$ on the contour line of $l(\theta) = \epsilon_i$ with the value $f(\theta_i)$.

bioRxiv preprint doi: https://doi.org/10.1101/564047; this version posted February 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

algorithm as soon as the current estimator variance differs from this lower bound by no more than a predefined threshold δ.

Treating the prior volumes x_i and the Monte Carlo estimate $\overline{L}_m$ as random variables, the variance σ^{2m}_{tot} of the NS estimator at iteration m can be estimated at each iteration without additional computational effort (see section S4 and [30]). This variance depends on the variance of the Monte Carlo estimate $\overline{L}_m$ and is monotonically increasing in $\text{Var}\left(\overline{L}_m\right)$ (see section S5). We define the term σ^{2m}_{min} which is the same variance $\text{Var}\left(\widetilde{Z}^m_{\mathcal{D}} + \widehat{Z}^m_{\mathcal{L}}\right)$ but under the additional assumption that the Monte Carlo estimate has variance 0: $\text{Var}\left(\overline{L}_m\right) = 0$. Clearly we have for any m (see section S5)

$$\sigma^{2m}_{\text{tot}} \geq \sigma^{2m}_{\text{min}}.$$

More importantly, as we show in section S5.2, σ^{2m}_{min} is monotonically increasing in m

$$\sigma^{2m'}_{\text{min}} \geq \sigma^{2m}_{\text{min}}, \quad \forall m' \geq m.$$

This allows us to bound the lowest achievable estimator variance $\sigma^2_{\text{min}} = \sup_{m\to\infty} \sigma^{2m}_{\text{min}}$ from below

$$\sigma^2_{\text{min}} \geq \sigma^{2m}_{\text{min}}.$$

The terms for σ^{2m}_{tot} and σ^{2m}_{min} both contain the unknown value L_m which can be approximated using its Monte Carlo estimate $\overline{L}_m$ giving us the estimations of the above variances $\widehat{\sigma}^{2m}_{\text{tot}}$ and $\widehat{\sigma}^{2m}_{\text{min}}$. We use these variance estimates to formulate a termination criteria by defining

$$\Delta^m_{LFNS} := \frac{\sqrt{\widehat{\sigma}^{2m}_{\text{tot}}} - \sqrt{\widehat{\sigma}^{2m}_{\text{min}}}}{\widehat{Z}^m_{\text{tot}}}$$

and terminate the algorithm as soon as $\Delta^m_{LFNS} < \delta$ for some predefined δ. This termination criteria seems intuitive since it terminates the LF-NS algorithm as soon as a continuation of the algorithm is not expected to make the final estimator significantly more accurate. As a final remark we note that the final estimator $\widehat{Z}^m_{\text{tot}}$ as well as the termination criteria using Δ^m_{LFNS} can of course also be

bioRxiv preprint doi: https://doi.org/10.1101/564047; this version posted February 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

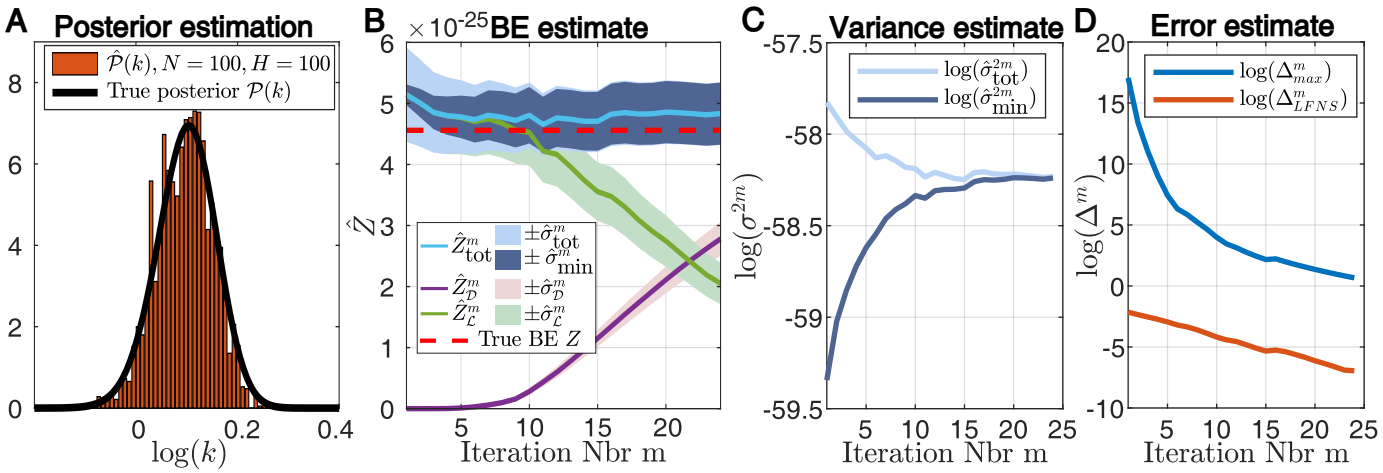

Figure 3: **A**: Histogram of the posterior $\mathcal{P}(k)$ estimate obtained with LF-NS using $N = 100$ and $H = 100$. The true posterior is indicated in black. **B**: Development of the estimation of the Bayesian evidence using the estimation based solely on the dead points $\widehat{Z}_{\mathcal{D}}$, the estimate approximation from the live points $\widehat{Z}_{\mathcal{L}}$ and the estimation based on both $\widehat{Z}_{\text{tot}}$. The corresponding standard errors are indicated as the shaded areas. The true Bayesian evidence is indicated with the dashed red line. **C**: Estimate of the current variance estimate $\widehat{\sigma}^{2m}_{\text{tot}}$ and the lower bounds for the lowest achievable variance $\widehat{\sigma}^2_{\text{min}}$. **D**: Developments of the different error estimations for each iteration.

applied in the standard NS case.

4 Examples

We test our proposed LF-NS algorithm on three examples for stochastic reaction kinetic models. The first example is the birth death model, already introduced in section 3.1, the second example is the Lac-Gfp model used for benchmarking in [32] and the third example is a transcriptional model from [48] with corresponding real data. In the following examples all priors are chosen as uniform or log-uniform in the bounds indicated in the posterior plots.

4.1 The stochastic birth-death Model

We first revisit the example of section 3.1 to compare our inference results to the solution obtained by FSP. We use the same data as in section 3.1 and use the same log-uniform prior. We run our LF-NS algorithm as described above using DP-GMM for the sampling. We used $N = 100$ LF-NS particles, $H = 100$ particle filter particles and sample at each iteration $r = 10$ particles. We ran the LF-NS algorithm until Δ^m_{LFNS} is smaller than 0.001. We show the obtained posterior in Figure 3 A. Figure 3 B shows the obtained estimates of the Bayesian evidence, where the shaded areas

bioRxiv preprint doi: https://doi.org/10.1101/564047; this version posted February 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

indicate the standard error at each iteration. The dashed red line indicates the true BE computed from 10^6 samples from $\Pi(\theta, \widehat{l}(\theta))$. The estimates of the lower $\widehat{\sigma}^{2m}_{\text{min}}$ and upper bound $\widehat{\sigma}^{2m}_{\text{tot}}$ for the lowest achievable estimator variance σ^2_{min} are shown in Figure 3 C and we can clearly see how they converge to the same value. For our termination criteria we show the quantities Δ^m_{max} and Δ^m_{LFNS} in Figure 3 D.

4.2 The Lac-Gfp model

We demonstrate how our algorithm deals with a realistic sized stochastic model, by inferring the posterior for the parameters of the Lac-Gfp model illustrated in Figure 4 A. This model has been already used in [32] as a benchmark, although with distribution-data. Here we use the model to simulate a number of trajectories and illustrate how our approach infers the posterior of the used parameters. This model is particularly challenging in two aspects. First, the number of parameters is 18, making it a fairly large model to infer. Secondly, the model exhibits switch-like behaviour which makes it very hard to approximate the likelihood of such a switching trajectory (see section S6.2 and particularly Figure S 3 for further details). We used $N = 500$ LF-NS particles, $H = 500$ particle filter particles and sample at each iteration $r = 50$ particles.

The measured species in this example is fluorescent Gfp (mGFP) where it is assumed that each Gfp-molecule emits fluorescence according to a normal distribution. We used one trajectory to infer the posterior, whose marginals are shown in Figure 4 E. The solid blue lines indicate the parameters used to simulate the data. Figure 4 B shows the estimated Bayesian evidence with corresponding standard errors for each iteration. Figure 4 D shows the corresponding estimations of the bounds of the lowest achievable variance. As we see, the estimated Bayesian evidence, as well as the estimated variance bounds, do several jumps in the process of the LF-NS run. These jumps correspond to iterations in which previously unsampled areas of the parameter space got sampled with a new maximal likelihood. In Figure 4 C we plotted the acceptance rate of the LF-NS algorithm for each iteration as well as the cumulative computational time[6]. The inference for this model took well over 12 hours and as we see, the computational time for each iteration seems to increases exponentially,

[6]The computation was performed on 48 cores of the Euler cluster of the ETH Zurich.

bioRxiv preprint doi: https://doi.org/10.1101/564047; this version posted February 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

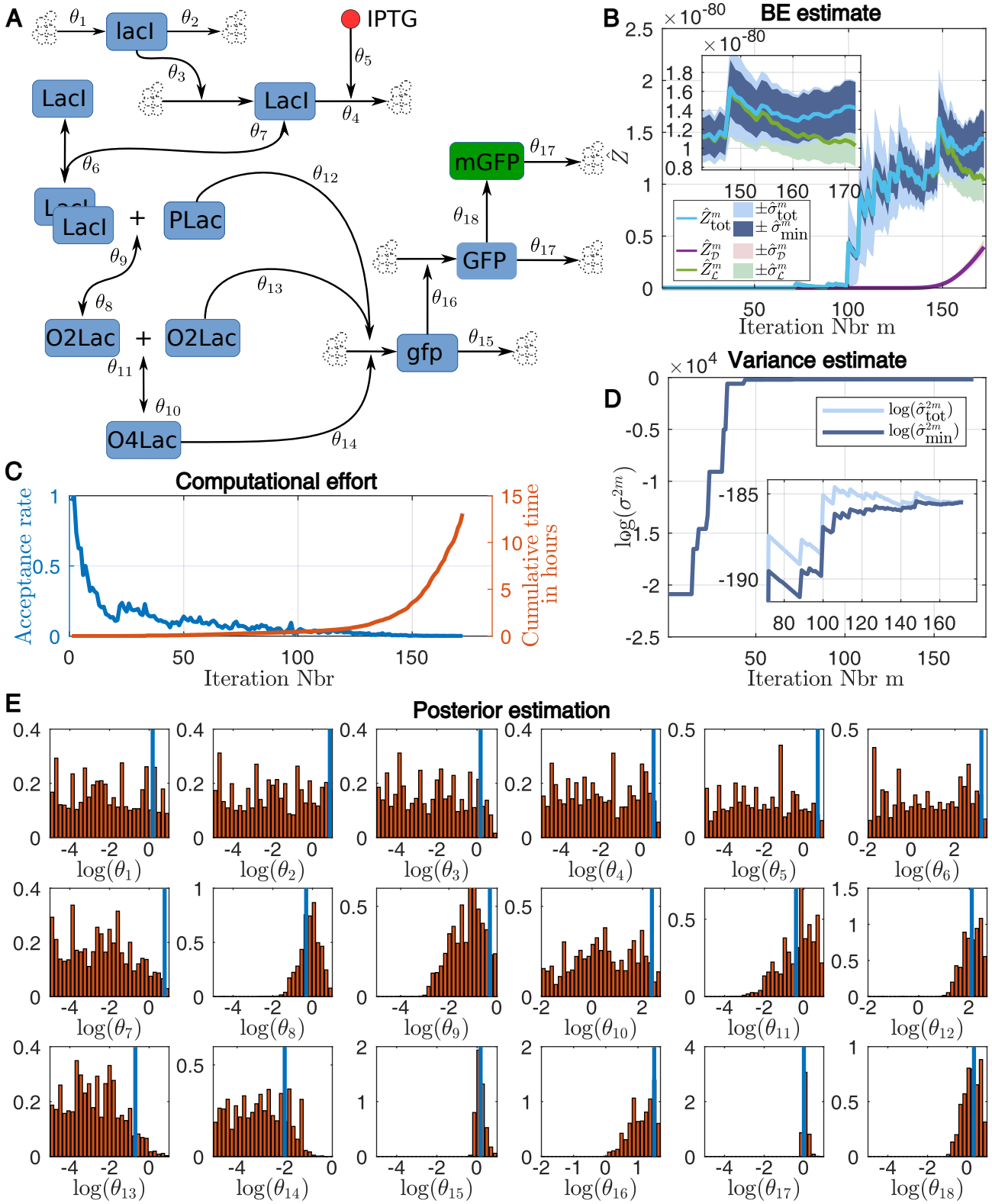

Figure 4: **A**: Schematic of the Lac-Gfp Model where the final measurement is the mature GFP (mGFP) and the input is IPTG (assumed to be constant 10μM). **B**: Development of the estimation of the Bayesian evidence using the estimation based solely on the dead points $\widehat{Z}_{\mathcal{D}}$, the estimate approximation from the live points $\widehat{Z}_{\mathcal{L}}$ and the estimation that uses both $\widehat{Z}_{\text{tot}}$. The corresponding standard errors are indicated as the shaded areas. **C**: The acceptance rate of the LF-NS algorithm for each iteration (blue) and the cumulative time needed for each iteration in hours (red). The computation was performed on 48 cores in parallel on the Euler cluster of the ETH Zurich. **D**: Estimate of the current variance estimate $\widehat{\sigma}^{2m}_{\text{tot}}$ and the lower bounds for the lowest achievable variance $\widehat{\sigma}^{2}_{\text{min}}$. **E:** Marginals of the inferred posterior distributions of the parameters based on one simulated trajectory. The blue lines indicate the parameters used for the simulation of the data.

bioRxiv preprint doi: https://doi.org/10.1101/564047; this version posted February 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

as the acceptance rate decreases. The low acceptance rate is expected, since the number of particle filter particles $H = 500$ results in a very high variance of the particle filter estimate (see Figure S3 B). Clearly, for this example, the early termination of LF-NS is essential to obtain a solution within a reasonable time.

4.3 A stochastic transcription model

As a third example we use a transcription model recently used in [48], where an optogenetically inducible transcription system is used to obtain live readouts of nascent RNA counts. The model consists of a gene that can take two configurations "on" and "off". In the "on" configuration mRNA is transcribed from this gene and can be individually measured during this transcription process (see [48] for details). We modelled the transcription through $n = 8$ subsequent RNA species that change from one to the next at a rate λ. This is done to account for the observed time of 2 minutes that one transcription event takes. With such a parametrization the mean time to move from species RNA_1 to the degradation of RNA_n is $\frac{n}{\lambda}$. An illustration of the model is shown in Figure 5 A. For the inference of the model parameters we chose five trajectories of real biological data, shown in Figure 5 C. Clearly, the system is inherently stochastic and requires corresponding inference methods. We ran the LF-NS algorithm for $N = 500$ and $H = 500$ on these five example trajectories. The resulting marginal posteriors are shown in Figure 5 B, we also indicated the model ranges considered in [48]. These ranges were chosen in [48] in an ad-hoc manner but, apart from the values for k_{off} seem to fit very well with our inferred results. In Figure 5 D and E we show the development of the evidence approximation as well as the corresponding standard errors and the development of the upper and lower bound estimation for the lowest achievable variance σ^2_{min}. Similarly to the Lac-Gfp example, we see that the development of the BE estimate is governed by random spikes which again are due to the sampling of particles with a new highest likelihood.

bioRxiv preprint doi: https://doi.org/10.1101/564047; this version posted February 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

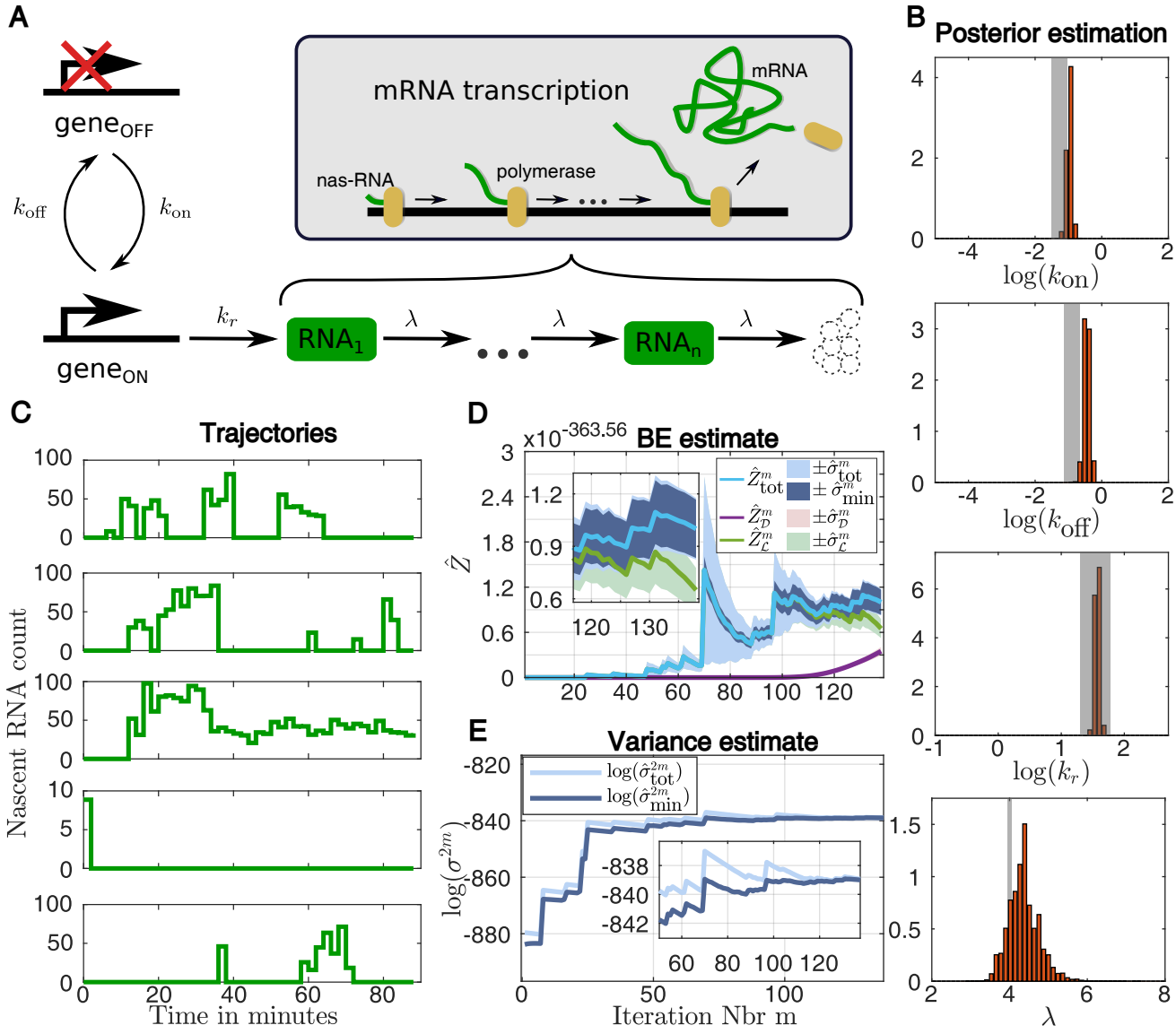

Figure 5: **A**: A schematic representation of the gene expression model. The model consists of a gene that switches between an "on" and an "off" state with rates k_{on} and k_{off}. When "on" the gene is getting transcribed at rate k_r. The transcription process is modelled through n RNA species that sequentially transform from one to the next at rate λ. The observed species are all of the intermediate RNA_i species. **B**: The marginal posterior distribution of the parameters of the system. The shaded areas indicate the parameter ranges that were considered in [48]. **C**: The five trajectories used for the parameter inference. **D**: Development of the estimation of the Bayesian evidence using the estimation based solely on the dead points $\widehat{Z}_{\mathcal{D}}$, the estimate approximation from the live points $\widehat{Z}_{\mathcal{L}}$ and the estimation that uses both $\widehat{Z}_{\text{tot}}$. The corresponding standard errors are indicated as the shaded areas. **E**: Estimate of the current variance estimate $\widehat{\sigma}^{2m}_{\text{tot}}$ and the lower bounds for the lowest achievable variance $\widehat{\sigma}^{2}_{\text{min}}$.

bioRxiv preprint doi: https://doi.org/10.1101/564047; this version posted February 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

5 Discussion

We have introduced a likelihood-free formulation of the well known nested sampling algorithm and have shown that it is unbiased for any unbiased likelihood estimator. While the utilization of NS for systems without an available likelihood is straight forward, one has to take precautions to avoid infeasibly high computational times. Unlike for standard NS it is crucial to include the estimation of the live samples to the final BE estimation as well as terminate the algorithm as soon as possible. We have shown how using a Monte Carlo estimate over the live points not only results in an unbiased estimator of the Bayesian evidence Z, but also allows us to derive a formulation for a lower bound on the achievable variance in each iteration. This lower bound at each iteration has been shown to be a lower bound for the best achievable variance and has allowed us to formulate a novel termination criterion that stops the algorithm as soon as a continuation can at best result in an insignificant improvement in accuracy. While the formulation of the variances and its lower bound were derived having a parallel LF-NS scheme in mind, they can equally well be used in the standard NS case and can be added effortlessly to already available toolboxes such as [14] or [22]. We emphasize that the lower variance bound approximation $\widehat{\sigma}^{2m}_{\text{min}}$ is neither a strict error term, as it only gives information of the variance of the estimator, nor a strict lower bound of the estimator variance since it contains the unknown term L_m. Instead, it gives an estimate of the lowest achievable estimator variance that depends on the Monte Carlo estimate of the likelihood average over the live points $\overline{L}_m$. This can be seen Figure 4 D and Figure 5 E, where the lower bound estimate $\widehat{\sigma}^{2m}_{\text{min}}$ does not only make jumps, but also decreases after each jump (the actual lower bound estimate σ^{2m}_{min} is monotonically increasing in m as shown in section S5.2). Our suggested LF-NS scheme has three different parameters that govern the algorithm behaviour. The number of LF-NS particles N determines how low the minimal variance of the estimator can get, where low values for N result in a rather high variance and high values for N result in a lower variance. The number of particles for the particle filter H determines how wide or narrow the likelihood estimation is and thus determines the development of the acceptance rate of the LF-NS run, while the number of LF-NS iterations determines how close the variance of the final estimate comes to the minimal variance. We have demonstrated the applicability of our

bioRxiv preprint doi: https://doi.org/10.1101/564047; this version posted February 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

method on several models with simulated as well as real biological data. Our LF-NS can, similarly to ABC, pMCMC or SMC models deal with stochastic models with intractable likelihoods and has all of the advantages of classic NS. We believe that particularly the variance estimation that can be performed from a single LF-NS run proves to be useful as well as the straight forward parallelization.

References

[1] Stuart Aitken and Ozgur E Akman. Nested sampling for parameter inference in systems biology: application to an exemplar circadian model. *BMC systems biology*, 7(1):72, 2013.

[2] Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle markov chain monte carlo for efficient numerical simulation. In *Monte Carlo and quasi-Monte Carlo methods 2008*, pages 45–60. Springer, 2009.

[3] Richard J Boys, Darren J Wilkinson, and Thomas BL Kirkwood. Bayesian inference for a discretely observed stochastic kinetic model. *Statistics and Computing*, 18(2):125–135, 2008.

[4] Brendon J Brewer, Livia B Pártay, and Gábor Csányi. Diffusive nested sampling. *Statistics and Computing*, 21(4):649–656, 2011.

[5] Nikolas S Burkoff, Csilla Várnai, Stephen A Wells, and David L Wild. Exploring the energy landscapes of protein folding simulations with bayesian computation. *Biophysical journal*, 102(4):878–886, 2012.

[6] Olivier Cappé, Simon J Godsill, and Eric Moulines. An overview of existing methods and recent advances in sequential monte carlo. *Proceedings of the IEEE*, 95(5):899–924, 2007.

[7] Nicolas Chopin and C Robert. Contemplating evidence: properties, extensions of, and alternatives to nested sampling. Technical report, Citeseer, 2007.

[8] Nicolas Chopin and Christian P Robert. Properties of nested sampling. *Biometrika*, 97(3):741–755, 2010.

bioRxiv preprint doi: https://doi.org/10.1101/564047; this version posted February 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[9] Arnaud Doucet, Nando De Freitas, and Neil Gordon. An introduction to sequential monte carlo methods. In *Sequential Monte Carlo methods in practice*, pages 3–14. Springer, 2001.

[10] Richard Dybowski, Trevelyan J McKinley, Pietro Mastroeni, and Olivier Restif. Nested sampling for bayesian model comparison in the context of salmonella disease dynamics. *PloS one*, 8(12):e82317, 2013.

[11] Johan Elf and Måns Ehrenberg. Fast evaluation of fluctuations in biochemical networks with the linear noise approximation. *Genome research*, 13(11):2475–2484, 2003.

[12] Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002.

[13] M Evans. Discussion of nested sampling for bayesian computations by john skilling. *Bayesian Statistics*, 8:491–524, 2007.

[14] F Feroz, MP Hobson, and M Bridges. Multinest: an efficient and robust bayesian inference tool for cosmology and particle physics. *Monthly Notices of the Royal Astronomical Society*, 398(4):1601–1614, 2009.

[15] F Feroz, MP Hobson, E Cameron, and AN Pettitt. Importance nested sampling and the multinest algorithm. *arXiv preprint arXiv:1306.2144*, 2013.

[16] Farhan Feroz and MP Hobson. Multimodal nested sampling: an efficient and robust alternative to markov chain monte carlo methods for astronomical data analyses. *Monthly Notices of the Royal Astronomical Society*, 384(2):449–463, 2008.

[17] Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361, 1977.

[18] Daniel T Gillespie. The chemical langevin equation. *The Journal of Chemical Physics*, 113(1):297–306, 2000.

bioRxiv preprint doi: https://doi.org/10.1101/564047; this version posted February 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[19] Andrew Golightly and Darren J Wilkinson. Bayesian parameter inference for stochastic biochemical network models using particle markov chain monte carlo. *Interface Focus*, 1(6):807–820, 2011.

[20] Andrew Golightly and Darren J Wilkinson. Bayesian inference for markov jump processes with informative observations. *arXiv preprint arXiv:1409.4362*, 2014.

[21] Dilan Görür and Carl Edward Rasmussen. Dirichlet process gaussian mixture models: Choice of the base distribution. *Journal of Computer Science and Technology*, 25(4):653–664, 2010.

[22] WJ Handley, MP Hobson, and AN Lasenby. Polychord: next-generation nested sampling. *Monthly Notices of the Royal Astronomical Society*, 453(4):4384–4398, 2015.

[23] R Wesley Henderson and Paul M Goggans. Parallelized nested sampling. In *AIP Conference Proceedings*, volume 1636, pages 100–105. AIP, 2014.

[24] Edward Higson, Will Handley, Mike Hobson, Anthony Lasenby, et al. Sampling errors in nested sampling parameter estimation. *Bayesian Analysis*, 2018.

[25] Andreas Hilfinger and Johan Paulsson. Separating intrinsic from extrinsic fluctuations in dynamic biological systems. *Proceedings of the National Academy of Sciences*, 108(29):12167–12172, 2011.

[26] Piers J Ingram, Michael PH Stumpf, and Jaroslav Stark. Network motifs: structure does not determine function. *BMC genomics*, 7(1):108, 2006.

[27] Edward L Ionides, C Bretó, and Aaron A King. Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 103(49):18438–18443, 2006.

[28] Nick Jagiella, Dennis Rickert, Fabian J Theis, and Jan Hasenauer. Parallelization and high-performance computing enables automated statistical inference of multi-scale models. *Cell Systems*, 2017.

bioRxiv preprint doi: https://doi.org/10.1101/564047; this version posted February 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[29] Rob Johnson, Paul Kirk, and Michael PH Stumpf. Sysbions: nested sampling for systems biology. *Bioinformatics*, 31(4):604–605, 2015.

[30] Charles R Keeton. On statistical uncertainty in nested sampling. *Monthly Notices of the Royal Astronomical Society*, 414(2):1418–1426, 2011.

[31] Caroline H Ko, Yujiro R Yamada, David K Welsh, Ethan D Buhr, Andrew C Liu, Eric E Zhang, Martin R Ralph, Steve A Kay, Daniel B Forger, and Joseph S Takahashi. Emergence of noise-induced oscillations in the central circadian pacemaker. *PLoS biology*, 8(10):e1000513, 2010.

[32] Gabriele Lillacci and Mustafa Khammash. The signal within the noise: efficient inference of stochastic gene regulation models using fluorescence histograms and stochastic simulations. *Bioinformatics*, 29(18):2311–2319, 2013.

[33] Thomas Liphardt. *Efficient computational methods for sampling-based metabolic flux analysis*. PhD thesis, ETH Zurich, 2018.

[34] David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[35] Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328, 2003.

[36] Harley H McAdams and Adam Arkin. Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences*, 94(3):814–819, 1997.

[37] Pia Mukherjee, David Parkinson, and Andrew R Liddle. A nested sampling algorithm for cosmological model selection. *The Astrophysical Journal Letters*, 638(2):L51, 2006.

[38] Brian Munsky and Mustafa Khammash. The finite state projection algorithm for the solution of the chemical master equation. *The Journal of chemical physics*, 124(4):044104, 2006.

bioRxiv preprint doi: https://doi.org/10.1101/564047; this version posted February 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[39] Brian Munsky, Brooke Trinh, and Mustafa Khammash. Listening to the noise: random fluctuations reveal gene network parameters. *Molecular systems biology*, 5(1), 2009.

[40] Iain Murray. *Advances in Markov chain Monte Carlo methods*. PhD thesis, Citeseer, 2007.

[41] Gregor Neuert, Brian Munsky, Rui Zhen Tan, Leonid Teytelman, Mustafa Khammash, and Alexander van Oudenaarden. Systematic identification of signal-activated stochastic gene regulation. *Science*, 339(6119):584–587, 2013.

[42] Ertugrul M Ozbudak, Mukund Thattai, Iren Kurtser, Alan D Grossman, and Alexander van Oudenaarden. Regulation of noise in the expression of a single gene. *Nature genetics*, 31(1):69–73, 2002.

[43] Johan Paulsson, Otto G Berg, and Måns Ehrenberg. Stochastic focusing: fluctuation-enhanced sensitivity of intracellular regulation. *Proceedings of the National Academy of Sciences*, 97(13):7148–7153, 2000.

[44] Michael Pitt, Ralph Silva, Paolo Giordani, and Robert Kohn. Auxiliary particle filtering within adaptive metropolis-hastings sampling. *arXiv preprint arXiv:1006.1914*, 2010.

[45] Michael K Pitt, Ralph dos Santos Silva, Paolo Giordani, and Robert Kohn. On some properties of markov chain monte carlo simulation methods based on the particle filter. *Journal of Econometrics*, 171(2):134–151, 2012.

[46] Nick Pullen and Richard J Morris. Bayesian model comparison and parameter inference in systems biology using nested sampling. *PloS one*, 9(2):e88419, 2014.

[47] Christian P Robert and Darren Wraith. Computational methods for bayesian model choice. In *AIP Conference Proceedings*, volume 1193, pages 251–262. AIP, 2009.

[48] Marc Rullan, Dirk Benzinger, Gregor W Schmidt, Andreas Milias-Argeitis, and Mustafa Khammash. An optogenetic platform for real-time, single-cell interrogation of stochastic transcriptional regulation. *Molecular cell*, 70(4):745–756, 2018.

bioRxiv preprint doi: https://doi.org/10.1101/564047; this version posted February 28, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[49] Michael Samoilov, Sergey Plyasunov, and Adam P Arkin. Stochastic amplification and signaling in enzymatic futile cycles through noise-induced bistability with oscillations. *Proceedings of the National Academy of Sciences*, 102(7):2310–2315, 2005.

[50] Scott A Sisson, Yanan Fan, and Mark M Tanaka. Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007.

[51] John Skilling. Nested sampling's convergence. In *AIP Conference Proceedings*, volume 1193, pages 277–291. AIP, 2009.

[52] John Skilling et al. Nested sampling for general bayesian computation. *Bayesian analysis*, 1(4):833–859, 2006.

[53] Vassilios Stathopoulos and Mark A Girolami. Markov chain monte carlo inference for markov jump processes via the linear noise approximation. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20110541, 2013.

[54] Tina Toni, David Welch, Natalja Strelkowa, Andreas Ipsen, and Michael PH Stumpf. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202, 2009.

[55] Darren J Wilkinson. Parameter inference for stochastic kinetic models of bacterial gene regulation: a bayesian approach to systems biology. In *Proceedings of 9th Valencia International Meeting on Bayesian Statistics*, pages 679–705, 2010.

[56] Christoph Zechner, Michael Unger, Serge Pelet, Matthias Peter, and Heinz Koeppl. Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings. *Nature methods*, 11(2):197–202, 2014.