



# Deterministic and Stochastic Parameter Estimation for Polymer Reaction Kinetics I: Theory and Simple Examples

Niklas Wulkow,\* Regina Telgmann, Klaus-Dieter Hungenberg, Christof Schütte, and Michael Wulkow

Two different approaches to parameter estimation (PE) in the context of polymerization are introduced, refined, combined, and applied. The first is classical PE where one is interested in finding parameters which minimize the distance between the output of a chemical model and experimental data. The second is Bayesian PE allowing for quantifying parameter uncertainty caused by experimental measurement error and model imperfection. Based on detailed descriptions of motivation, theoretical background, and methodological aspects for both approaches, their relation are outlined. The main aim of this article is to show how the two approaches complement each other and can be used together to generate strong information gain regarding the model and its parameters. Both approaches and their interplay in application to polymerization reaction systems are illustrated. This is the first part in a two-article series on parameter estimation for polymer reaction kinetics with a focus on theory and methodology while in the second part a more complex example will be considered.

equations, that is, a system of ordinary differential equations (ODEs) that describes the temporal change of concentrations of the reactants and products and their properties by modeling individual reactions involving parameters like reaction rate coefficients. Measurement data is given in the sense that some characteristic quantities of the process are measured for a sequence of time points. In PE, one wants to find the parameters for which the solution of the model is as close as possible to the data given. In PE, closeness is typically measured by means of the residual function that measures the distance between model-based prediction and measurement data.

PE for polymer reaction kinetics is used in hundreds of articles. There also is a wide range of literature on PE for ODE systems, including its specific use in polymer reaction engineering,<sup>[1–3]</sup> or systems biology,<sup>[4]</sup> for example. This article complements

## 1. Introduction

Parameter estimation (PE) for chemical kinetics means the process of fitting a mathematical model of the reaction process of interest to given observation data by tuning the parameters of the model. The model is mostly given in the form of reaction rate

these works by a combination of different approaches to the PE problem that normally are dealt with independently: We will discuss (1) the classical approach to PE via minimization of the residual function and (2) the approach utilizing Bayesian PE and uncertainty quantification. We will outline the pros and cons and the different contexts in which these approaches seem appropriate, and will demonstrate their use in application to different realistic scenarios.

It is not the objective of this article to review the literature on PE for polymer reaction processes. Its aim is to demonstrate that classical and Bayesian PE complement each other in ways that allow to deal with the following typical real-world scenario: the model is still under construction, for most model parameters at most rough estimates are available, and the available measurement data is not sufficient, in the sense that not all quantities of interest for the process can be measured, there are too few time points, and/or the measurement quality is low, that is, there is significant measurement error. In this scenario, classical PE often leads to severe problems: the misfit between model and data, the residual, stays significantly large even after many minimization steps, and the resulting fit is unsatisfactory, or the parameters are strongly correlated. We will shed light on the exact reasons for this outcome and discuss how to improve the situation. In this scenario, classical PE suffers from the fact that insufficient and low quality measurement data leads to uncertainty about the optimal parameters.


N. Wulkow

Freie Universität Berlin  
 Arnimallee 14, Berlin 14195, Germany  
 E-mail: niklas.wulkow@zib.de

Dr. R. Telgmann, Dr. M. Wulkow  
 Computing in Technology GmbH  
 Harry-Wilters-Ring 27, Rastede 26180, Germany

Prof. K.-D. Hungenberg  
 Hungenberg Consultant  
 Ortsstrasse 135, Birkenau 69488, Germany

Prof. C. Schütte  
 Zuse Institute Berlin  
 Takustraße 7, Berlin 14195, Germany

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/mats.202100017>

© 2021 The Authors. Macromolecular Theory and Simulations published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/mats.202100017

In general, parameter uncertainty can take several forms: (1) The uncertainty of all parameters is small and their distribution is approximately Gaussian around the optimal values, (2) the uncertainty is large for some parameters meaning that they are hardly informative and the model may even be insensitive with respect to these parameters, and (3) the uncertainty is large because the residual function has more than one main well. Case (1) is the lucky case in which classical PE typically works well despite low-quality data. In case (2), special precautions have to be taken in classical PE in order to avoid the “ill-conditioned” problem of determining the precise value of uninformative parameters (see Section 2.2.1 below). Case (3) poses a severe problem for classical PE that can only be dealt with via many restarts of the minimization procedure for diverse initial conditions. In realistic cases, however, one may have to deal with a mixture of all three cases and may want to get quantitative information on parameter uncertainty and the effects causing it.

Quantitative information on parameter uncertainty typically requires using Bayesian PE instead of classical PE. Bayesian PE, in the form discussed in this article, samples the distribution on parameter space induced by the limited data available, allows to include prior information, and does not only allow to quantify uncertainty but also to identify uninformative parameters as well as multi-modal distributions. However, it is computationally much more demanding than classical PE. Therefore, uncertainty quantification via Bayesian PE can complement classical PE but will not replace it.

Bayesian methods for chemical reaction kinetics have also been discussed extensively in the literature, especially for model inference and model reduction for large chemical reaction networks<sup>[5–8]</sup> or polymer reaction kinetics,<sup>[9]</sup> in relation to specific chemical contexts as catalysis,<sup>[10]</sup> or combustion,<sup>[11]</sup> for large experimental data<sup>[12]</sup> or high dimension cases,<sup>[13]</sup> as well as for more theoretical aspects like approximation quality and sparsity.<sup>[14]</sup> In contrast, applications to polymer chemistry, especially applications to polymer reaction kinetics with a focus on Bayesian PE, were rarely considered: For example, in [15], a Bayesian framework including PE is presented for surfactant-polymer flooding with no focus on polymer reactions. Process design via Bayesian optimization and Bayesian design of experiments has attracted some attention recently (see refs. [16–18], for example), but with no focus on PE.

In this article, Bayesian PE for polymer reaction kinetics is discussed, including uncertainty quantification and its propagation by the model for the real-world scenario of too few and low quality data. We outline the theoretical foundations and introduce a practical sampling algorithm including a novel and efficient step-size scaling procedure. Moreover, we will show how the different tools used in classical PE like efficient evaluation or gradient computation of the residual re-appear (and can be re-used) in Bayesian PE.

The aim of this article is to show that only an interplay between classical and Bayesian PE allows to perform reliable PE for the real-world scenario of insufficient data, allowing to shed light on parameter uncertainty and the effects causing it. The article describes an entire pipeline of tools, from classical PE, via its improvement for ill-conditioned cases, up to uncertainty quantification via Bayesian PE, including efficient sampling and visualization. While many individual ideas contained in this article are not entirely new and have in part been known since many years,

their composition into a unified framework for PE, including uncertainty quantification (PE+UQ) is new in the best sense of a feature article. To our best knowledge, no article exists in which the topic PE+UQ in polymerization is covered in a similar way or in a comparable breadth. Moreover, several of the components were adapted to the context and several others are novel. We believe that this article provides a helpful reference for students and experts involved with non-standard PE problems in both polymer kinetics and other problems where measurements and unknown parameters are strongly unbalanced.

In order to make real-world testing easily possible, the entire integrated pipeline is made available in the commercial software package PREDICI.<sup>[19]</sup> This article is the first part of a two-paper series. In this part, we will develop the theory and illustrate it on a few very basic examples. In the second part, a comprehensive co-polymerization model from polymer kinetics will be used to show how to really work with the suggested tools.

The article is organized as follows: First, in Section 2, we give an introduction to PE for ODEs, starting with the general setting, then considering classical PE including residual minimization, the potential source of ill-conditioning and improved numerical stability, followed by outlining the key ideas of Bayesian PE including tools for efficient sampling of the posterior distribution, its visualization, resulting uncertainty propagation, and concluded by a comparison between classical and Bayesian PE. Next, in Section 3, we illustrate the interplay between classical and Bayesian PE and demonstrate the performance of the proposed algorithms in application to two examples, an illustrative example of a chemical reaction scheme with four substances, and a more realistic radical polymerization setting. Finally, in Section 4, we draw conclusions and set the results into perspective.

## 2. Parameter Estimation

This section is devoted to the in-depth discussion of two different approaches to parameter estimation (PE) for mathematical models in terms of differential equations.

### 2.1. General Remarks on Parameter Estimation of Chemical Reaction Models

We will concentrate on the PE problem for chemical reaction models that we are going to consider later in this article. Such models can be written as systems of ordinary differential equations (ODE)

$$\frac{dx(t)}{dt} = f(x(t), \theta), \quad x(0) = x_0 \quad (1)$$

where  $x_0$  is a given initial value and  $\theta \in \mathbb{R}^d$  denotes a vector of parameters, such as a set of reaction rate coefficients, that determine the outputs of the model. The general form of (1) also holds for general evolution equations, in particular, the countable system of ordinary differential equations (CODEs) that are used to describe polymerization kinetics.<sup>[19]</sup> The state of the system at time  $t$ , for example, the vector of concentrations of chemical species in a reactor, is denoted by  $x(t)$ . That is,  $x(t)$  in general is a multi-dimensional vector in  $\mathbb{R}^m$  and the right hand side of our

ODE system,  $f$ , maps  $x(t)$  to its temporal derivative, again a vector in  $\mathbb{R}^m$ . Under very mild conditions on  $f$ , the ODE system (1) has a unique solution that is completely specified by the initial condition and the present parameter values. We simply write it in the following form:

$$x(t) = F(t, x_0, \theta), \quad \text{or, componentwise,} \quad x_i(t) = F_i(t, x_0, \theta) \quad (2)$$

where  $x_i(t)$  denotes the  $i$ th component of the vector  $x(t)$  and  $F_i$  the respective component of  $F$ . By integrating Equation (1), in time we find that

$$F(t, x_0, \theta) = x_0 + \int_0^t f(x(s), \theta) ds \quad (3)$$

In general, the solution map  $F$  is not available in explicit form but can only be computed numerically and comes with the (often considerable) computational effort of computing the trajectory of the ODE system (1) from time 0 to time  $t$ . This is especially true for polymerization systems that are solved with respect to full chain-length distributions. Note that in reality complex polymerization systems can result in complex mathematical models, for example, ODEs of a high order or Partial Differential Equations (e.g., population balance equations describing the particle size distribution in emulsion polymerization<sup>[20]</sup>). In the notation of this article, the model formulation is encapsulated solely in the model function  $F$  so that the theory from hereon is applicable regardless of the form or complexity of the system. As we will see later, if the evaluation of the model function is expensive, this naturally makes the use of the numerical methods we introduce expensive.

### 2.1.1. Measurements

Next, we assume that we made measurements of the state of the system at times  $t_j, j = 1, \dots, T$ . We denote these measurements by  $X = (X_1, \dots, X_T)$ . In general, we have to assume that we may not be able to perform measurements for all components of the state vector  $x(t)$  but just for some of them. For the sake of simplicity, we will use notation in which  $x(t_j)$ , the model's state at  $t_j$ , is directly compared to  $X_j$ , for example, by computing the Euclidean distance

$$\|x(t_j) - X_j\|^2 = \sum_{i=1}^m (x_i(t_j) - X_{ji})^2 \quad (4)$$

where  $X_{ji}$  denotes the  $i$ th component of the measurement vector at time  $t_j$ . If measurements are only available for some components of the state vector, then the sum must solely contain these components. For real applications, we should write  $g_i(X(t_j))$  using the state vector  $X(t_j)$  and a function  $g_i$  that maps the state variables to a measurement of type  $i$ . Typical examples are the conversion of a chemical substance or the mean value of a polymer distribution. Both are computed in terms of some variables  $x_i$  of the state vector. However, for the theory developed herein, we will just use  $x_i(t_j)$  as model description of a measurement to make the presentation simpler. Of course, the goal is to develop a model so that each output  $x_i(t_j)$  is as close as possible to  $X_{ji}$ , that is, to minimize the distance in Equation (4).

### 2.1.2. Residual Function

If we knew that the model (1) perfectly reproduced the measurements for a given parameter vector  $\theta^*$ , then we would have a *perfect fit* in the sense that

$$X_j = F(t_j, X_0, \theta^*) \quad \text{for all } j = 1, \dots, T \quad (5)$$

For all other parameter vectors, we can measure the deviation between model and data by the weighted least squares residual function,

$$\begin{aligned} R_X(\theta) &= \frac{1}{mT} \sum_{j=1}^T \sum_{i=1}^m \frac{|X_{ji} - F_i(t_j, X_0, \theta)|^2}{|X_{ji}|^2} \\ &= \frac{1}{mT} \sum_{j=1}^T \|D_j^{-1} [X_j - F(t_j, X_0, \theta)]\|^2 \end{aligned} \quad (6)$$

where the diagonal weight matrix  $D_j$  contains the diagonal entries  $X_{ji}^2, i = 1, \dots, m$ . We choose the entries of  $D$  in this way to measure the quality of fitness of the model  $F$  to each data point  $X_{ji}$  by the same simple relative error model since we do not assume any special knowledge on specific data points. This also makes it easier to translate the formulation to further theoretical sections in this article. Note that the weight matrix could be chosen differently to reflect even a manually chosen weighting of the errors of individual data points (see, e.g., refs. [21, 22]). For example, in many real cases, one will extend the scaling  $D_{ji}$  by measurement inaccuracies or other thresholds  $s_{ji}$  by setting  $D_{ji} = \max(X_{ji}^2, s_{ji})$ . In the perfect fit case, the residual function would have a minimum at  $\theta^*$  with  $R_X(\theta^*) = 0$ . Any evaluation of the residual function comes with the cost of computing the solution of the ODE system, in general by numerical means.

We cannot assume to have a perfect fit for various reasons, for example, our model might simply be imprecise. Nevertheless, parameter vectors for which the residual function  $R_X$  is smaller will be preferable to one with larger values of  $R_X$ . The parameter values for which  $R_X$  attains a (global) minimum will belong to the best fit given the model and will be called optimal parameter, given model and data. As we will see, it is crucial to be aware that there may be more than one minimum, so-called local minima, and that the vicinity of a global minimum may contain regions in which the values of  $R_X$  only slightly deviate from its minimal value, at least in comparison to the error of computing the residual function numerically. These are exactly the problems that make PE notoriously hard.

### 2.1.3. Measurement Errors

Until now, we have not considered any measurement error. In general, the measurements will have a relative error, that is, the measured value  $\tilde{X}_{ji}$  of component  $i$  at time  $t_j$  will be related to the real value  $X_{ji}$  via

$$\frac{|\tilde{X}_{ji} - X_{ji}|}{|X_{ji}|} = \varepsilon_{ji}, \quad \text{i.e.,} \quad \tilde{X}_{ji} = X_{ji}(1 + \varepsilon_{ji}) \quad (7)$$

where  $\varepsilon_{ji}$  is an unknown measurement error that is often modeled by a normally distributed random number with mean 0 and variance  $\sigma_{ji}^2$  with some small  $\sigma_{ji}$  that has to be provided as part of the measurement process. Consequently, even for a perfect fit we only can hope for

$$X_{ji} = F_i(t_j, X_0, \theta^*) + X_{ji} \varepsilon_{ji} \text{ for all } i = 1, \dots, T \quad (8)$$

such that the residual function will no longer vanish at  $\theta^*$ , but yield the value

$$R_X(\theta^*) = \frac{1}{mT} \sum_{j=1}^T \sum_{i=1}^m |\varepsilon_{ji}|^2 = R_X^* \quad (9)$$

Therefore, all other parameter vectors  $\theta$  with  $R_X(\theta) = R_X^*$  are as good as  $\theta^*$  with regard to fitting the model to the data. For the sake of simplicity, let us consider the uniform case  $\sigma_{ji} = \sigma$ . Then, the square root of  $R_X^*$  is normally distributed with mean 0 and variance  $\sigma^2$ . Because of this, the probability distribution on the parameter space is given by

$$\exp\left(-\frac{1}{2\sigma^2} R_X(\theta)\right) \quad (10)$$

The exponential function comes in because of the assumed normal distribution of the measurement error. This probability distribution is called the likelihood in the sense that all parameters leading to residual values around 0 with standard deviation  $\sigma$  are likely candidates for the “true” parameters while ones with higher residual values are exponentially unlikely. This concept survives if we do not assume a perfect fit. Then parameter values in a  $\sigma$ -vicinity of a minimum of the residual function are likely, while ones with higher residual values are unlikely.

In case that the  $\sigma_{ji}$  are not identical across all  $i, j$ , it is straightforward to modify the likelihood accordingly by dividing each term in Equation (6) by the corresponding  $\sigma_{ji}^2$  instead of  $\sigma^2$ . In part 2 of this series, we will devote more space to the measurements and their accuracies in the context of polymerization kinetics.

*Remark 1.* Note that we take the perspective that the measurement  $X_{ji}$  is given and together with the measurement error induces a distribution on what the true values  $\tilde{X}_{ji}$  is. This leads to the Equations (8) and (9). One could alternatively assume  $X_{ji} = F_i(t_j, X_0, \theta^*)(1 + \varepsilon_{ji})$ , so that the measurement results from a perturbation of the true model, and define the diagonal entries of the weight matrix in (6) by  $D_{ji} = F_i(t_j, X_0, \theta)$  to arrive at (9).

#### 2.1.4. Classical versus Bayesian PE

These short considerations already explain the different approaches to the PE problem: Classical PE is focused on computing minima of the residual function in a reliable and numerically efficient way. However, even local minimization of  $R_X$  may be troublesome because there may be extended regions in parameter space where the residuum is numerically indistinguishably small which makes numerical computations ill-conditioned. In contrast, Bayesian PE tries to explore the likely parameter

regimes stochastically. Numerically, this is a much more demanding task that avoids the problems of classical PE but often may be computationally infeasible. If feasible, it generates information about the uncertainty of model-based predictions caused by uncertainty about the parameter values. The following two sections will outline these aspects in more detail.

### 2.2. Classical Parameter Estimation

While the generation of synthetic data with a model with parameters—or of true measurements with a chemical experiment in the laboratory—is usually called the forward problem, the inference of the parameters from the data is referred to as the inverse problem:

Forward problem: Model, parameter  $\theta$

→ Measurements/Results  $X$

(11)

Inverse problem: Model, Measurements/Results  $X$

→ Parameter  $\theta$

In classical parameter estimation, solving the inverse problem means, one searches for the global minimum  $\theta^*$  of the residual function.

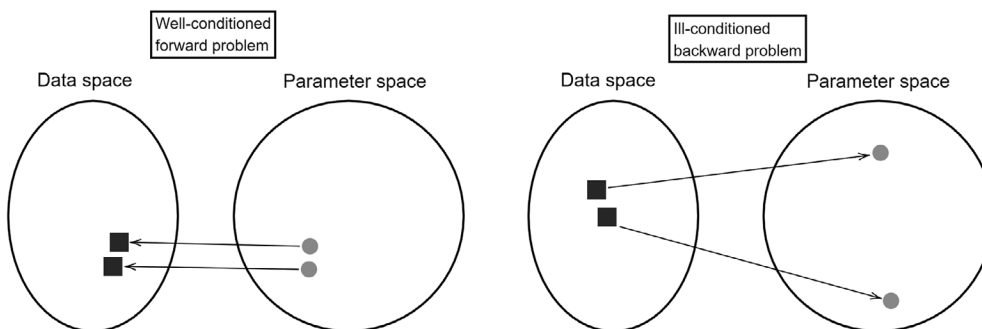
In many cases, such an inverse problem is ill-conditioned. This means that even slight variations to the data can cause a significantly different result for the estimated parameter. Since we assume the data to be subject to measurement errors, these perturbations naturally occur and can cause misleading results regarding the parameter estimation. These effects can be significant and, it must be emphasised, very severely alter the result of the parameter estimation, even if the forward problem is well-conditioned (which means slight changes to the parameters only lead to small perturbations of the model result). This is illustrated in **Figure 1**. An example can be found in ref. [1], pages 227–230.

There are various numerical methods to find a local minimum of a function, in our case of the residual function. Most of these methods focus on finding critical points where the derivative of the function is 0. These points typically are local minima. The global minimum is selected as the local minimum with the lowest residual function value. There exists an abundance of literature on numerical parameter estimation methods to which we refer for a complete overview of the topic, such as refs. [2, 23–27]. We will concentrate on one particular method to both illustrate the basic methodology of such methods and introduce an effective way to overcome typical problems in parameter estimation. This method will also be used in the experiments later on in this article.

#### 2.2.1. Gauss-Newton Method with Essential Directions

A prominent way to find the zeros of a function is Newton's method applied to the derivative of  $R_X(\theta)$ . Let us write the terms

$$\frac{X_{ji} - F_i(t_j, X_0, \theta)}{X_{ji}} \quad (12)$$



**Figure 1.** Left: Well-conditioned forward problem. Similar parameters lead to similar simulation data. Right: Ill-conditioned inverse problem. Similar simulation data can be caused by very different parameters.

for all  $i, j$  into a single vector  $E_X(\theta) = [E_1(\theta), \dots, E_N(\theta)]^T$  with  $N = mT$  entries so that  $\frac{1}{mT} \|E_X(\theta)\|_2^2 = R_X(\theta)$ . Its gradient defines the Jacobian matrix  $J(\theta) = J_{ij}(\theta) \in \mathbb{R}^{mT \times d}$  via

$$J_{lj}(\theta) = \frac{\partial E_l(\theta)}{\partial \theta_j}, \quad l = 1, \dots, N, \quad j = 1, \dots, d \quad (13)$$

where for now  $\theta_j$  denotes the  $j$ th entry of a parameter  $\theta$ .

This approach leads to the Gauss–Newton method.<sup>[1,28]</sup> It denotes an iterative scheme in which, from an initial parameter guess  $\theta_0$ , subsequent iterates  $\theta_k, k = 1, 2, \dots$  are generated. Please note that the index of the vector  $\theta_k$  now denotes the iteration step such that its  $j$ th entry is denoted  $\theta_{k,j}$ . Each step of the iteration consists of the following two sub-steps:

1. Finding  $\Delta\theta_k$  that minimizes  $\|J(\theta_k)\Delta\theta_k + E_X(\theta_k)\|_2$  (14)
2. Setting  $\theta_{k+1} = \theta_k + \Delta\theta_k$

A direct way to solve this minimization problem is by demanding that

$$J(\theta_k)\Delta\theta_k = -E_X(\theta_k) \quad (15)$$

Using the Moore–Penrose pseudo-inverse  $J^+(\theta_k)$  (this is a substitute for the inverse of the matrix if the true inverse does not exist or the matrix is not square),<sup>[29,30]</sup> this gives

$$\Delta\theta_k = -J(\theta_k)^+ E_X(\theta_k) \quad (16)$$

The pseudo-inverse  $J(\theta_k)^+$  can be computed using Singular Value Decomposition. To this end, one first represents  $J = J(\theta_k)$  as

$$J = USV^T \quad (17)$$

where  $U \in \mathbb{R}^{N \times d}$  and  $V \in \mathbb{R}^{d \times d}$  are orthogonal matrices, while  $S \in \mathbb{R}^{d \times d}$  is a diagonal matrix, carrying the singular values  $s_1 \geq s_2 \geq \dots \geq s_d$  on the diagonal. If  $s_d > 0$ ,  $J^+$  is given by

$$J^+ = VS^{-1}U^T \quad (18)$$

For handling the specific case  $s_d \approx 0$  or  $s_d = 0$ , see below. With this, we can write the Gauss–Newton step in Equation (16) as

$$(\Delta\theta_k)_j = - \sum_{l=1}^d V_{jl} \frac{1}{s_l} U_l^T E_X(\theta_k), \quad j = 1, \dots, d \quad (19)$$

where  $U_l$  denotes the  $l$ th column of  $U$ .

**Damping:** When computing the pseudo-inverse of  $J(\theta)$ , we essentially solve an inverse problem (a linear system of equations) which can be ill-conditioned. In this case, errors in the data may perturb the Jacobian  $J(\theta)^+$ . In addition, correlations between parameters will result in unreasonably large step lengths. Consequently, the objective function at the end point of the step may show a potentially huge increase instead of a monotonous decrease. To counter this, one has to apply a damping strategy by multiplying the Gauss–Newton step with a damping factor  $\lambda_k \leq 1$ , so that

$$\theta_{k+1} = \theta_k + \lambda_k \Delta\theta_k \quad (20)$$

The damping factor is updated from step to step using a monotonicity test based on the objective function. There are many sophisticated strategies available. A summary is given in ref. [31].

**Essential Directions:** In many cases, one has to deal with hidden parameter correlations. These correlations are not easily detected, but rather hidden in the model structure. The parameters are also not fully correlated in a way that one can express one parameter by a function of some others, which would merely be a modeling defect. Instead, they are only locally dependent in the sense that a change of one parameter can nearly (with respect to the residual) be adjusted by some other parameters. By that, the parameter estimation practically does not lead to a unique solution. In other words: only some of the directions in parameter space are really essential whereas the others depend on the essential directions. These are called the flat directions.

Next, we outline a technique for detecting such parameter correlations and for dealing with them. This technique is rarely discussed in the context of classical PE, but proves to be essential for practical purposes, for more details please see ref. [1]. A similar approach using rank decisions to identify sensitive parameters is presented in the text book.<sup>[4]</sup>

As we will see below, parameter correlations and their detection and removal are deeply related to the so-called condition



number of the underlying Jacobian matrix  $J$  at the respective step of the iteration, which is given by

$$\kappa(J) = \frac{s_1}{s_d} \quad (21)$$

that is, by the ratio between the biggest and the smallest singular value. If  $s_d = 0$ , then the condition is infinite.

Often, the singular values of  $J$  show a clear gap in their magnitudes at an index  $k_{\text{ess}}$ , meaning that  $s_1, \dots, s_{k_{\text{ess}}}$  are clearly bigger than  $s_{k_{\text{ess}}+1}, \dots, s_d$ . With this, we can closely approximate  $J(\theta)$  by

$$J(\theta) \approx U \underbrace{\begin{pmatrix} s_1 & & & & 0 \\ & s_2 & & & \\ & & \dots & & \\ & & & s_{k_{\text{ess}}} & \\ & & & & 0 \\ 0 & & & & & \dots \\ & & & & & & 0 \end{pmatrix}}_{=: \tilde{S}} V^T \quad (22)$$

This allows us to fix the condition of the minimization problem in Equation (14) to  $\frac{s_1}{s_{k_{\text{ess}}}}$ . The Gauss–Newton step is then given by

$$\Delta\theta_k = -U\tilde{S}V^T E_X(\theta_k) \quad (23)$$

The explicit matrix notation (23) shows that we have transformed the parameter space of the problem into a new space where the essential directions are just the main axes. By dropping the non-essential coordinates there, performing the Gauss–Newton step and then going back to the original space, we have automatically performed an adjustment of the flat directions based on the progress of the essential directions. It is important to note that the parameters of the flat directions are not insensitive or even fixed. There are examples where all pairs of two out of three parameters (with one parameter fixed in a reasonable range) are perfectly sensitive and essential, but the fit problem with all three parameters has only two essential directions. In all classical PEs shown in this publication, we have applied the detection and treatment of essential directions. We call this method reduced-direction approach.<sup>[32]</sup> Flat directions are usually identified by being assigned to singular values that lead to a condition number of about 100 or larger (if there is no other clear gap between the singular values). In a 2D parameter space, the condition can be visualized by the ratio of the half-axes of an ellipsoid (with the lengths of these half-axes corresponding to the singular values). It is obvious that in cases of large condition numbers of, say,  $\kappa(J) = 100$  or larger, this ellipsoid almost degenerates to a straight line exhibiting a clear parameter correlation. Finally, it is important to note that the condition of a problem and the number and type of essential directions is strictly dependent on the problem setup. Even a slight change of the scaling or a different weighting can alter these structures. All analyses of this kind are only performed locally at single points in parameter space that are reached by the iteration. Therefore, it is very important to also create a global view of the problem.

### 2.3. Bayesian Parameter Estimation

Bayesian parameter estimation aims at avoiding the possibly ill-conditioned inverse problem by a stochastic reformulation of the problem such that only forward problems need to be solved in order to find good parameter values. It does not result in a single optimal parameter value but, in contrast, in information on the probability of certain parameter values. To this end, one includes some a priori known uncertainty about measurement errors into the parameter estimation. The Bayesian inverse problem is structured as

Model, Measurements  $X$ , Measurement precision  $\sigma$

$$\rightarrow \text{Probability for each } \theta \quad (24)$$

The underlying assumption is that the uncertainty in the given data and model imperfection translates to an uncertainty in the estimated parameter. From a quantification of data uncertainty, Bayesian PE deduces a probability distribution on the parameter space. If this distribution is peaked, that is, concentrates around a single maximum, then the parameter uncertainty is small and the parameter at the maximum highly informative. If, in contrast the distribution is spread out, perhaps with many local maxima, then parameter uncertainty is large.

The central relation of Bayesian PE is

$$p(\theta|X) \propto L(X|\theta)p_{\Theta}(\theta) \quad (25)$$

$p(\theta|X)$  denotes the so-called posterior density, interpreted as the probability distribution on parameter space resulting from the data values  $X$  observed.  $L$  denotes the likelihood that the observed data  $X$  come from the model with parameters  $\theta$ . The density  $p_{\Theta}$ , called the prior density, is used to encode all the prior knowledge that we may have on the parameter values. The relation states that the posterior density is proportional, that is, equal up to a constant, to the product of the likelihood and the prior density.

#### 2.3.1. Derivation and Fundamentals of Bayesian Modelling

Bayesian PE starts from the fundamental Bayesian identity that directly results from the definition of conditional probabilities:

$$\mathbb{P}[\theta \in A | X \in B] \cdot \mathbb{P}(X \in B) = \mathbb{P}[X \in B | \theta \in A] \cdot \mathbb{P}(\theta \in A) \quad (26)$$

where as above,  $X = (X_1, \dots, X_T)$  denotes observation data,  $\theta$  the parameter vector, and  $A$  and  $B$  subsets of the parameter space and the data space, respectively. Both sides of this equality are identical to the joint probability that  $X \in B$  and  $\theta \in A$ .

Next, one assumes that the respective probability distributions on data and parameter space exhibit probability densities, meaning that

$$\mathbb{P}[\theta \in A | X \in B] \cdot \mathbb{P}(X \in B) = \int_A \int_B p(\theta|X)p_X(X) d\theta dX \quad (27)$$

$$\mathbb{P}[X \in B | \theta \in A] \cdot \mathbb{P}(\theta \in A) = \int_B \int_A L(X|\theta)p_{\Theta}(\theta) dX d\theta \quad (28)$$

where  $L$  now denotes the density related to the probability of the data, given the parameters. Since this is true for all  $A, B$ , we get the Bayesian identity (26) in terms of densities:

$$p(\theta|X)p_X(X) = L(X|\theta)p_\Theta(\theta) \quad (29)$$

Assuming  $p_X(X) > 0$ , (29) yields

$$p(\theta|X) = \frac{p_\Theta(\theta)}{p_X(X)} L(X|\theta) \quad (30)$$

We will see later that the posterior density  $p(\theta|X)$  allows us to do uncertainty quantification on the parameter estimation and in the resulting model based predictions. At this point, however, the key fact is that computation of the posterior density is typically well-conditioned even if the inverse problem is ill-conditioned.<sup>[33]</sup>

*Specifics on The Likelihood:* Under the assumption that the noise  $\varepsilon_{ji}$  in (8) is normally distributed with variance  $\sigma^2$  we find, as above,

$$L(X|\theta) \propto \exp\left(-\frac{1}{2\sigma^2} R_X(\theta)\right) \quad (31)$$

where  $\alpha$  means identity up to some normalization factor such that  $\int L(X|\theta)dX = 1$ . This holds because

$$\begin{aligned} L(X|\theta) &\propto \prod_{ij} \exp\left(-\frac{1}{2\sigma^2} \frac{|X_{ji} - F_i(t_j, X_0, \theta)|^2}{|X_{ji}|^2}\right) \\ &= \exp\left(-\frac{1}{2\sigma^2} \sum_{ij} \frac{|X_{ji} - F_i(t_j, X_0, \theta)|^2}{|X_{ji}|^2}\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2} R_X(\theta)\right) \end{aligned} \quad (32)$$

The first relation holds because every data point  $X_{ji}$  is assumed to be normally distributed around  $F_i(t_j, X_0, \theta)$  with variance  $\sigma^2$ . The likelihood of all data points is then the product of these terms. The factor  $mT$  from Equation (6) can be omitted here since it only changes the normalization constant. Note that the residual that is used in classical PE is directly transformed into a probability.

*Specifics on  $p_X$ :* In order to get an interpretation of the data density  $p_X$ , we first integrate equation (29) with respect to  $\theta$  on both sides, yielding

$$p_X(X) \int p(\theta|X)d\theta = 1 \quad (33)$$

Therefore,  $p_X$  merely plays the role of the normalizing factor of the posterior understood as a density over parameter space. Because of this identity, one often simply writes the relation (29) in the form given by Equation (25).

*Specifics on The Prior:* The density  $p_\Theta$ , called the prior density, is used to encode all the prior knowledge that we may have on the parameter values. When introducing the so-called potential function

$$\mu(\theta) = -\ln p_\Theta(\theta) \quad (34)$$

(that takes the value  $\infty$  where the density is zero), and using the expression for the likelihood, we can express the posterior by

$$p(\theta|X) = \frac{1}{Z} \exp(-S_X(\theta)), \quad S_X(\theta) = \frac{1}{2\sigma^2} R_X(\theta) + \mu(\theta) \quad (35)$$

Although with this, we merely reformulate the posterior density by introducing additional notation at this point, the term  $S_X$  will be of help later on.

For the form of the prior, there are two typical cases:

**Uniform prior:** One may know the parameters are positive values and perhaps even that their values must come from a certain interval. In this case,  $p_\Theta$  is constant in a certain subset  $A$  of the parameter space and zero outside, that is,

$$\mu(\theta) = \begin{cases} c & \text{if } \theta \in A \\ \infty & \text{otherwise} \end{cases} \quad (36)$$

$c > 0$  is a constant value resulting from normalization,  $\int_A p_\Theta(\theta)d\theta = 1$ . It simply means, we know the parameters are in a certain interval but have no additional information about them.

**Gaussian prior:** Often one assumes that the parameters come from a normal distribution around some mean value  $\theta_0$  with an appropriate covariance matrix  $\Sigma$ . In this case,

$$p_\Theta(\theta) = \frac{1}{Z_\Theta} \exp\left(-\frac{1}{2}(\theta - \theta_0)^T \Sigma^{-1}(\theta - \theta_0)\right) \quad (37)$$

If the prior takes this Gaussian form then,

$$S_X(\theta) = \frac{1}{2\sigma^2} R_X(\theta) + \frac{1}{2}(\theta - \theta_0)^T \Sigma^{-1}(\theta - \theta_0) + \ln(Z_\Theta) \quad (38)$$

Of particular interest is the maximal posterior parameter estimate

$$\theta_{max} = \operatorname{argmax}_\theta p(\theta|X) = \operatorname{argmin}_\theta S_X(\theta) \quad (39)$$

denoting the most probable parameter value based on both the data and the prior estimate. In the case that the prior is constant over an interval and the optimal parameter estimate of classical PE from (8) is in this interval, then it is equal to the maximal posterior parameter estimate. If the prior is not uniform and has its global maximum at a different value than the estimate from classical PE, then this is generally not the case. In fact, then there exists a conflict between our prior, data-independent knowledge, which suggests that the optimal parameters are close to one certain value, and the evidence that the data provide. The posterior distribution takes both into account. Its specific shape then depends on the exact specifications of the prior, the likelihood and the data. For example, it could well be a bi-modal distribution with peaks at the maxima of likelihood and prior.

### 2.3.2. Computing the Posterior and Related Expectation Values

In the literature, two main algorithmic concepts for computing the posterior dominate. The first and most simple one is grid-based: one defines a grid of parameter values  $\theta_k$ ,  $k = 1, \dots, M$ ,

evaluates the likelihood and the prior at these points and sets

$$p(\theta_k|X) \approx \frac{L(\theta_k|X)p_\Theta(\theta_k)}{\sum_{k=1}^M L(\theta_k|X)p_\Theta(\theta_k)} \quad (40)$$

This requires  $M$  evaluations of trajectories of the ODE system (one for each point as part of the evaluation of the residual  $R_X$ ; remember that the likelihood  $L$  is directly computed from  $R_X$  as in (31)). Whenever the parameter space is high dimensional, the number of points of a typical grid gets too large (it grows with  $n^d$  if we take  $n$  grid points per dimension for each of the  $d$  dimensions). For such cases, sparse grid or sparse approximation techniques offer a solution if the dimension is not too high.<sup>[14,34]</sup> Expectation values (like the mean or the variance) of the posterior distribution then are computed by

$$\langle A \rangle = \int A(\theta)p(\theta|X)d\theta \approx \sum_{k=1}^M A(\theta_k)p(\theta_k|X) \quad (41)$$

for some function  $A = A(\theta)$  (with  $A(\theta) = \theta$  for the mean and  $A(\theta) = \theta^T \theta$  for variance).

The second approach samples the posterior distribution. This means, one generates a set of parameters, the samples, that are distributed according to the desired posterior distribution.

To do this, there is a variety of techniques based on Monte Carlo methods<sup>[35]</sup> and corresponding multilevel approaches.<sup>[36]</sup> A particularly prominent variant is the so-called Langevin sampler,<sup>[37,38]</sup> or its pre-conditioned or underdamped versions,<sup>[39,40]</sup> that generates a sequence of parameter points  $(\theta_1, \dots, \theta_n)$  according to the following iterative scheme: In each step, one first computes the proposal  $\tilde{\theta}_{k+1}$  for the next parameter point via

$$\tilde{\theta}_{k+1} = \theta_k - \Delta t \text{grad}S_X(\theta_k) + \sqrt{2\Delta t} r_k \quad (42)$$

where  $r_j$  is a random number generated from the standard  $m$ -dimensional normal distribution with mean 0 and variance 1,  $\Delta t$  some sufficiently small stepsize, and  $\text{grad}S_X$  the gradient of the function  $S_X$  from Equation (37). Next, one determines the acceptance probability  $\alpha$  according to the Metropolis–Hastings algorithm<sup>[41]</sup>:

$$\alpha = \min \left\{ 1, \frac{e^{-S_X(\tilde{\theta}_{k+1})} q(\theta_k, \tilde{\theta}_{k+1})}{e^{-S_X(\theta_k)} q(\tilde{\theta}_{k+1}, \theta_k)} \right\} \quad (43)$$

with

$$q(\theta', \theta) = \exp \left( -\frac{1}{4\Delta t} \|\theta' - \theta + \Delta t \text{grad}S_X(\theta)\|^2 \right) \quad (44)$$

This choice of the acceptance probability yields that for the sequence of parameters, the detailed balance property holds; that is, in expectation, there are as many jumps from a parameter  $\theta$  to  $\tilde{\theta}$  as in the other direction. Under mild additional conditions, this has a consequence that the distribution of parameters in the sequence converges to  $\frac{1}{Z} e^{-S_X(\theta)}$  as in Equation (35).

Last, in each step, one sets  $\theta_{k+1} = \tilde{\theta}_{k+1}$  with probability  $\alpha$  and  $\theta_{k+1} = \theta_k$  otherwise. The acceptance step guarantees that the sampler cannot enter regions of the parameter space where  $S_X = \infty$ , which may exist due to uniform priors, for example.

The sequence generated via this kind of MALA is automatically distributed according to the posterior<sup>[37]</sup> such that expectation values are simply given by

$$\langle A \rangle \approx \frac{1}{n} \sum_{k=1}^n A(\theta_k) \quad (45)$$

by the Birkhoff Ergodic Theorem.<sup>[42]</sup>

These sampling techniques automatically concentrate points in regions of the parameter space where the posterior density is significantly large. Expectation values converge with  $n^{-1/2}$  in the number of points generated, almost independent of the dimension. However, each step of the Langevin sampler requires the evaluation of the gradient of the residual function, which may be very expensive whenever the dimension of the data space is high. Note that for multi-modal posteriors, the expressiveness of the expectation value can be low since it will simply lie somewhere between the different modes without sending a interpretable message. In this case, it becomes especially important to get a global view of the representation compared to simply considering the expectation value.

The optimal acceptance rate in order to yield a fast convergence typically lies between 50% and 70%, as is discussed in refs. [37, 43]. It does not hurt if the acceptance ratio lies slightly outside of this interval. Truly worrying would be acceptance ratios of close to 100% because this would indicate that the step size was too small to efficiently sample from the entire state space of interest and below  $\approx 20\%$  because in this case the algorithm is likely to remain in a state unreasonably long.

We provide additional explanations on the algorithm in Appendix A.1.

### 2.3.3. Prescaling of the Step Size

In the case that parameters are in different orders of magnitude, the choice of a reasonable step size  $\Delta t$  can pose a difficult task: a choice that is suitable for one parameter might either be too big for another, yielding proposals to frequently come from regions where  $S_X$  is high so that the proposal is accepted only with very low probability, or too low so that an infeasibly high number of steps is required in order for the distribution of samples to converge because the steps taken become minuscule in the directions of parameters with values on higher orders of magnitude (see Figure 2).

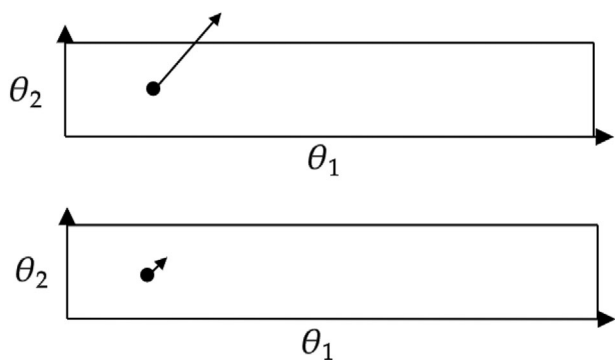
We therefore introduce a novel step size prescaling technique, which chooses different step sizes in each parameter direction by proposing a prescaling of the step size for each parameter beforehand. We replace the step size  $\Delta t$  by a diagonal matrix  $P$  that contains the step sizes for each direction in its diagonal. The proposal step then has the form

$$\tilde{\theta}_{k+1} = \theta_k - P \text{grad}S_X(\theta_k) + \sqrt{2P} r_k \quad (46)$$

where  $r_k$  is a vector with random entries drawn from a standard normal distribution.

In Appendix A.2, we provide an algorithmic way on how to suitably set the diagonal entries of  $P$ . Note that realistic examples very often require a proper prescaling.





**Figure 2.** Example of difficulty when the parameters are of very different sizes. The parameter spaces here are bounded via a uniform prior that is positive only inside the box. Top: Step size that is suitable for parameter 1 but not for parameter 2 since the proposal lies outside the parameter space. Bottom: Step size that is suitable for parameter 2 yields that many samples are necessary to move through the space of parameter 1.

#### 2.3.4. Marginal Densities

Whenever  $\theta$  is more than 2D, we can hardly visualize the full density any more. Instead, we will be interested in the posterior densities for each of these individual parameters or a subset of two of them. These densities will be called marginal densities: let  $\theta$  be multidimensional and be split into  $\theta = (u, v)$  and let the set of all possible parameters be denoted by  $\Theta = \Theta^U \times \Theta^V$  (so that  $u$  lies in  $\Theta^U$  and  $v$  in  $\Theta^V$ ). Then the marginal distribution over  $u$  is defined as

$$p_U(u|X) := \int_{\Theta^V} p(u, v|X) dv \quad (47)$$

In other words, one integrates the posterior  $p$  with respect to all entries of  $\theta$  except for the ones in  $u$  (see **Figure 3**).

#### 2.3.5. Visualization of Densities with Kernel Density Estimation

Sometimes a global representation of the posterior is needed, for example, for the sake of visual inspection. Then point-wise values need to be interpolated in some appropriate form. To this end, we suggest Kernel density estimation (KDE).<sup>[44,45]</sup> Although there exist various other approaches to visualising the distribution, we want to illustrate KDE here in order to provide a complete work flow from the classical PE problem of minimizing the residual, the augmentation into the Bayesian PE, sampling from the posterior and the visualization.

KDE can be understood as a continuous form of a histogram. Having sampled parameters  $\theta_1, \dots, \theta_n$  from the posterior with, for example, MALA, we could create a histogram of the distribution by counting the number of samples that lie in each of a chosen set of boxes. However, weaknesses of histograms include a severe loss of precision when choosing boxes too big and need for a high number of samples when choosing boxes small. In KDE, we approximate the posterior at every point  $\theta$  by

$$\hat{p}_{KDE}(\theta) := \frac{1}{nh_1 \dots h_d} \sum_{k=1}^n K(H^{-1}(\theta - \theta_k)) \quad (48)$$

where  $K$  is a closeness measure and  $H$  is a diagonal matrix that contains so-called bandwidths  $h_1, \dots, h_d > 0$  as diagonal entries. As a consequence,  $\hat{p}_{KDE}(\theta)$  will be high in regions with many samples since the values for  $K(H^{-1}(\theta - \theta_k))$ , which are added together, will be high if many samples are close to  $\theta$ . This is consistent with the fact that by construction of the MALA algorithm, many samples should come from regions where the posterior is high. In total, we approximate  $p(\theta)$  through a continuous weighting according to closeness of samples to  $\theta$ . The similarity to histograms lies in the fact that we do keep note of how many samples are close to a point  $\theta$  in the parameter space but instead of simply counting how many samples lie inside a certain box, we include the exact distance between each sample and  $\theta$ . More details on the choice of the function  $K$  and the bandwidths can be found in Appendix A.3.

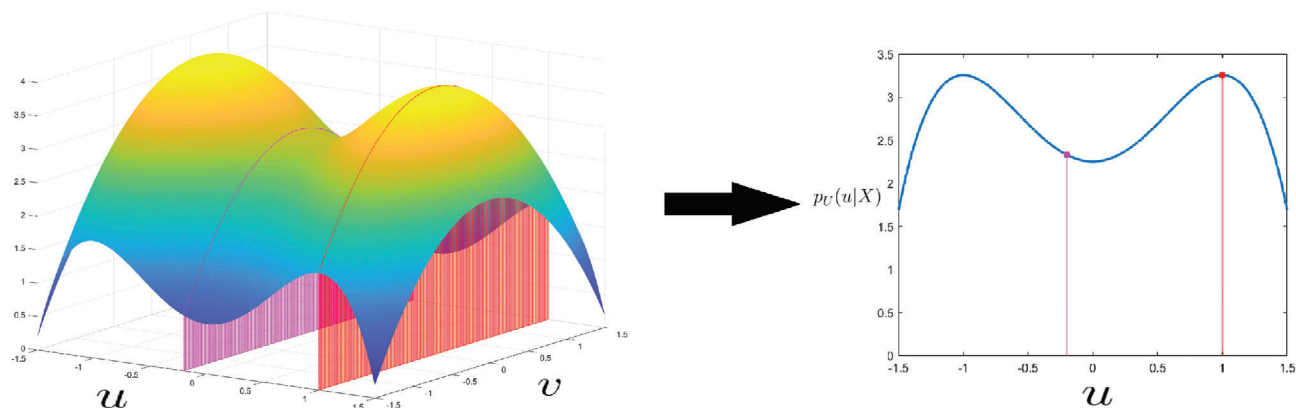
#### 2.4. Comparison between Classical and Bayesian PE

So far we have discussed two fundamentally different approaches to PE. In classical PE, we seek a solution of the optimization problem in order to find the parameter that, in combination with the model, best explains the data at hand. In Bayesian PE, we assign a probability distribution to parameter space, the posterior distribution, assuming that the data and model might not be precise. In this way, we can quantify how reliably we can determine the parameters and draw conclusions for the range of possible forward simulations of the model. Although both approaches can be tackled by various means, there are specific state-of-the-art methods for both, which we have presented in our variants of the Gauss–Newton optimization and MALA.

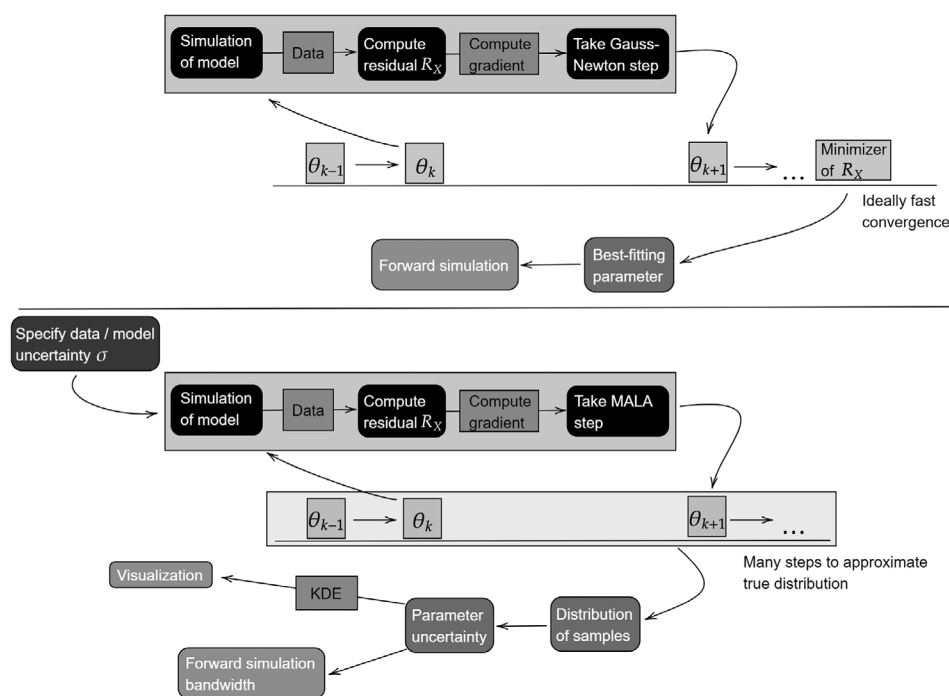
We have to observe, however, that in spite of their different aims, algorithmically both of these approaches are in fact quite similar. As is conceptualized in **Figure 4**, both methods rely on the approximation of the gradient of the residual function that compares data and the output of the model for a given parameter. While in classical PE, we construct a sequence of which we hope that it quickly converges to a minimum of the residual, in the Bayesian approach with MALA, we generate a sequence (of a chosen length) which should visit parameters in a frequency that is proportional to the probability with which they denote the optimal parameter. From then on, the real structural differences become apparent: In the Bayesian approach, we can then quantify and visualize the parameter uncertainty and instead of using the optimal parameters for forward simulation of the model, compute the propagation of parameter uncertainty by the model, for example, by computing the 90% percentile of the forward trajectories (this last aspect will be illustrated in Section 3).

### 3. Examples

We now showcase the interplay of classical PE with the Bayesian approach to parameter estimation in combination with the prescaled MALA on two examples and explain which conclusions can be made from the results in practice. The first example is quite simple and meant to illustrate the basic steps while the second is more complex. All simulations have been performed in the commercial program package Predici, v11, 2021.<sup>[19]</sup>



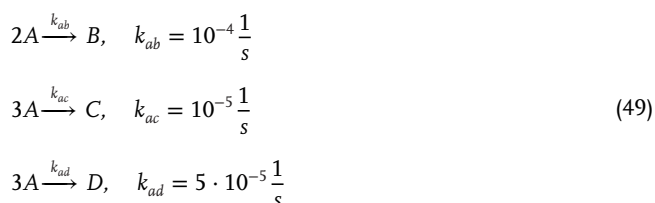
**Figure 3.** Left: 2D distribution  $p(u, v|X)$  for two scalar parameters  $u$  and  $v$ . Right: Marginal distribution over  $u$  was computed by integrating along the axis of  $v$  for each value of  $u$ .



**Figure 4.** Structures of classical PE using Gauss–Newton (top) and Bayesian PE using MALA (bottom). Both approaches are very similar in the way they create a sequence of states. While in classical PE, one tries to find the minimizer of the residual as quickly as possible, MALA generates a long sequence of states (as part of the overall result) whose distribution converges to the true distribution.

### 3.1. Example 1: Simple Four-Substance First Order Reaction

Consider the reaction scheme

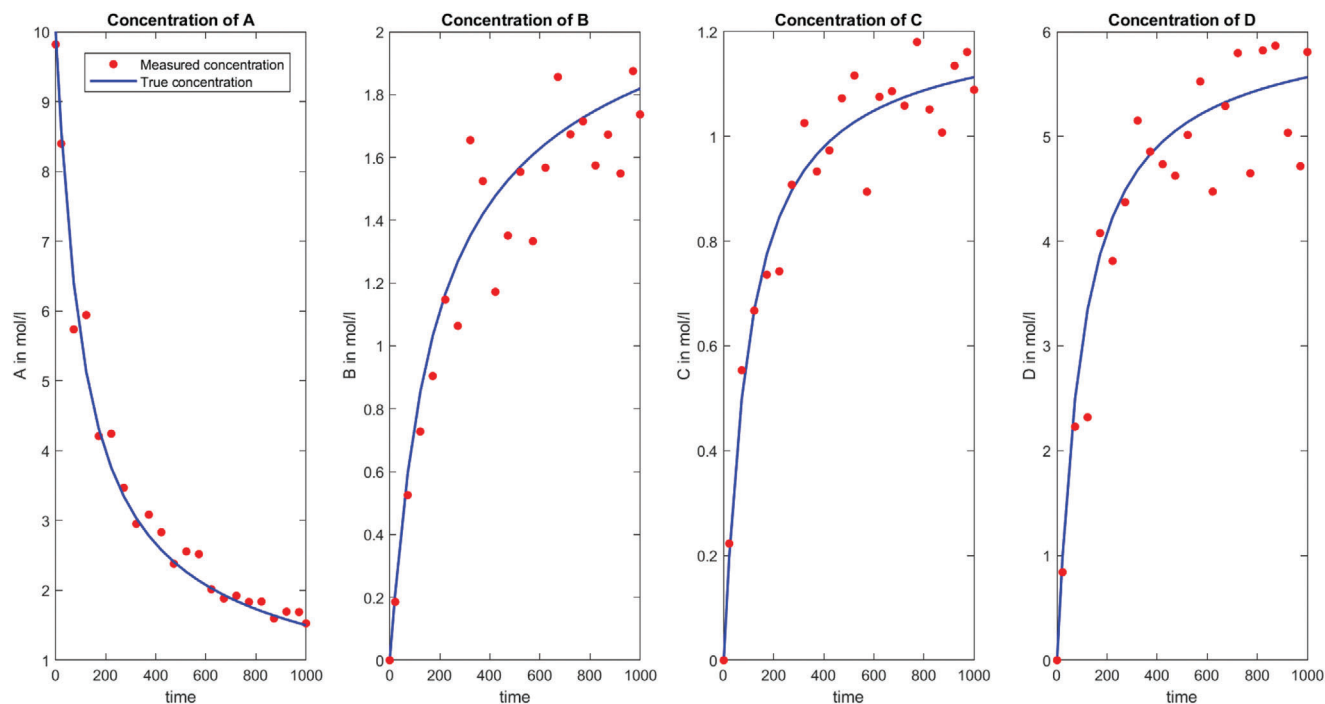


with initial concentrations  $A_0 = 10 \frac{\text{mol}}{\text{l}}$ ,  $B_0 = C_0 = D_0 = 0 \frac{\text{mol}}{\text{l}}$ . From here onwards, we omit the units of the parameters. The

reaction temperature is not important here. We simulate the reaction for a time span of 1000 s and observe concentrations of A, B, C, and D on 22 different time points. We then perturb each data point by a factor that is normally distributed with mean 1 and standard deviation  $\sigma = 0.1$  to simulate measurement errors (Figure 5).

#### 3.1.1. Estimating Parameters with Data about A

As a first parameter estimation step, we determine best-fitting parameters  $k_{ab}$ ,  $k_{ac}$  and  $k_{ad}$ , using only the data points corresponding to A. Using the Gauss–Newton method with essential directions with initial values given by  $(k_{ab})_0 = (k_{ac})_0 = (k_{ad})_0 =$



**Figure 5.** Exact and measured concentrations of A, B, C, and D. It might appear as if the data points of substance A were perturbed only by a much smaller amount compared to the other substances but this is only due to the scaling of the figures.

$10^{-5}$ , we obtain that the residual is minimized for  $k_{ab} = 0.89 \times 10^{-4}$ ,  $k_{ac} = k_{ad} = 2.85 \times 10^{-5}$ , yielding a residual of  $\approx 0.08$ . The estimate of  $k_{ab}$  shows an error of 11% while the estimates of  $k_{ac}$  and  $k_{ad}$  deviate strongly from their true values. The condition number of the Jacobian results to be  $\kappa \approx 10^8$ , a value that can be interpreted as sign of strong correlation of at least two parameters. Actually, the analysis of degrees of freedom shows only two out of three independent parameter directions here.

For this reason, we will now investigate the probability densities for both parameters. We deploy the prescaled MALA and make the correct assumption that the data are subject to normally distributed measurement errors with  $\sigma = 0.1$ . We bound the parameter space to the interval  $[10^{-6}, 10^{-3}] \times [10^{-6}, 10^{-4}] \times [10^{-6}, 10^{-4}]$  for  $k_{ab}$ ,  $k_{ac}$  and  $k_{ad}$ , resulting in a uniform distribution on this interval as the prior distribution  $p_{\Theta}$ . We specify as residual function the function introduced in Equation (6). We set up the prescaled MALA so that a step should on average have the length of one fiftieth of the length of the parameter space in each direction. The sampling sequence created has length 2500. As initial values, we choose the estimated parameters given above but discard the first 500 steps of the algorithm to minimize dependence on the initial values. The acceptance ratio lies at 75% which is slightly higher than the supposed optimal range of 50-70% mentioned in Section 2.3 but still suitable.

The result is shown in **Figure 6** (blue curves). As we can see, we get a broad range of probable values for all three parameters, especially  $k_{ac}$  and  $k_{ad}$ . It is important to observe here that in spite of the deviation of the estimated parameters from the true values, the true values are well inside the range of probable parameters. However, the data about A do not seem to be enough to determine

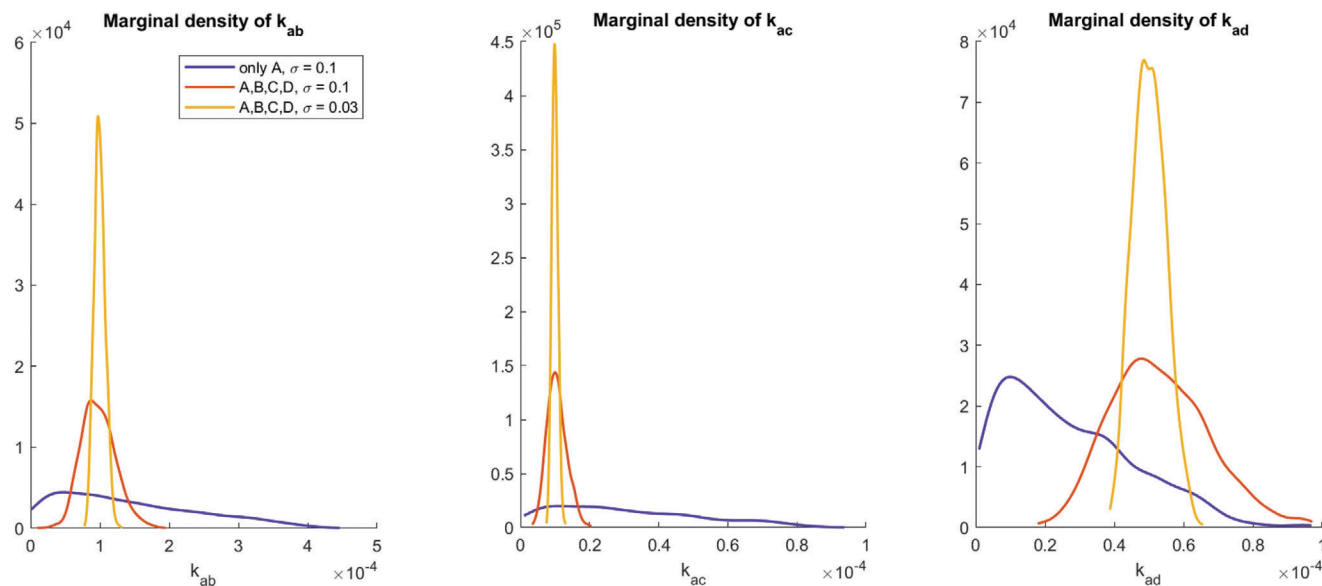
good estimates for  $k_{ac}$  and  $k_{ad}$ . This lack of data together with the assumption of a relatively high measurement uncertainty yields a high uncertainty of the parameters.

### 3.1.2. Including Data about B, C and D

We now include the observed values of B, C, and D into the data set, which also consist of 22 data points each. The parameter estimation using Gauss-Newton with essential directions gives  $k_{ab} = 0.89 \times 10^{-4}$ ,  $k_{ac} = 0.95 \times 10^{-5}$  and  $k_{ad} = 4.57 \times 10^{-5}$  with a condition number of  $\kappa \approx 2$ ; so, apparently, the inclusion of B, C, and D significantly improves the parameter estimation. The estimated values, however, still deviate from the true parameters, especially the estimate for  $k_{ab}$ . In order to check whether there still exists a large uncertainty around these values, we again sample parameters using prescaled MALA from the posterior distribution that comes from the inclusion of B, C, and D. As we can see in **Figure 6** (red curves), the densities for  $k_{ac}$  and  $k_{ad}$  become much slimmer, yielding a more reliable estimate for  $k_{ac}$ . Plus, the true values are still inside the range of probable values.

### 3.1.3. Effect of a Lower Measurement Uncertainty

Let us assume, we take the measurements now with a more precise apparatus. To this end, we perturb the original data again but with  $\sigma = 0.03$ . The result of the Gauss-Newton scheme now is  $k_{ab} = 1.0 \times 10^{-4}$ ,  $k_{ac} = 0.99 \times 10^{-5}$ , and  $k_{ad} = 4.96 \times 10^{-5}$ , with a condition number of  $\kappa \approx 1$ . This is much closer to the true



**Figure 6.** KDE visualization of the marginal densities of  $k_{ab}$ ,  $k_{ac}$ , and  $k_{ad}$  with data perturbed with  $\sigma = 0.1$  and  $\sigma = 0.03$ . The inclusion of B, C, and D is vital to generate reliable estimates for  $k_{ac}$  and  $k_{ad}$ . A lower data measurement error yields better certainty of the parameters.

values. Sampling from the ensuing posterior distribution (with  $\sigma = 0.03$ ) with prescaled MALA again generates densities which are much more precise than with  $\sigma = 0.1$ .

In summary, as we can see in Figure 6 (orange curves), higher trust in the observed data—resulting in a smaller value for  $\sigma$ —and inclusion of the concentrations of B, C, and D in the parameter estimation procedure now enables us to determine all parameters with only a small amount of uncertainty. It is important to understand that the specified value for  $\sigma$  depends on the measurement precision of the experimentalist. The more precisely the measurements are taken, the more meaningful generally the result is. This is also illustrated by the fact that the parameter densities become slimmer, meaning that the range of likely values for a parameter is narrowed down with higher measurement precision. This means, receiving a small residual in the classical PE and a large range of parameters with a high value for  $\sigma$  need not be a contradiction: It means that if the data uncertainty is large, then the residual can be made small but at the same time the optimal parameters are not reliable.

Note that the overall measurement error may have different sources. In most cases one determines  $\sigma$  as a repetition error, that is, measuring the same sample  $n$  times for example by spectroscopic or chromatographic methods. Then  $\sigma$  is usually small, but we neglect all errors caused by sample preparation which usually needs several operations like dilution, filtration, neutralization, extraction, etc. Further errors arise from the evaluation of the raw data, that is, the transformation of the detector signal to the final measuring quantity. This may involve setting the correct baseline, choice of calibration curve, etc. The latter ones are usually higher than just the repetition error. By choosing different values for  $\sigma$ , we want to demonstrate the importance of the evaluation of correct  $\sigma$ .

If one has either, ideally, estimated the measurement precision to be high or at least is confident in the data measurements then  $\sigma$  should be chosen small. As a consequence, the parameter den-

sities will likely be very sharp and centered around the optimal value from classical PE.

In conclusion, this example illustrates how the reliability of parameter estimation depends on the data at hand and the amount of trust we can put into it.

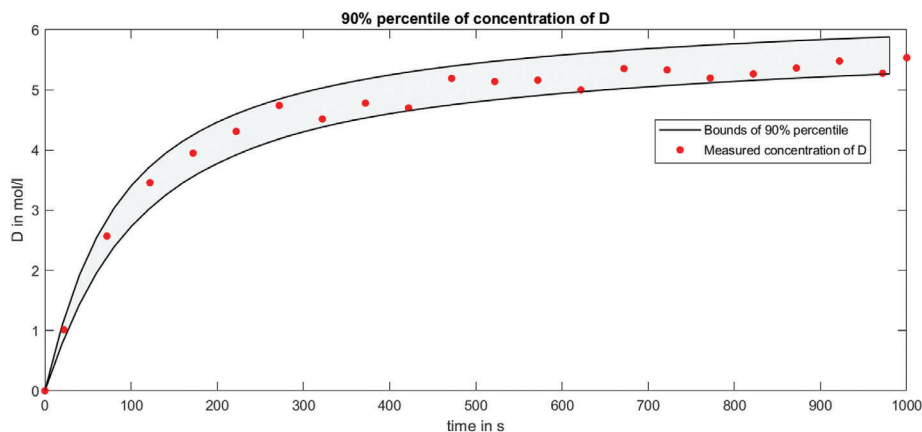
### 3.1.4. Uncertainty Propagation

We can now substitute the sampled parameters into the model itself to see the range of probable outcomes of a simulation. To this end, we have simply created one trajectory for each sampled parameter vector, and at each point in time discarded the upper and lower 5% of values, leaving us with the 90% percentile. In Figure 7, we can see that all data points for D lie inside the 90% percentile of simulations. This indicates that our model can explain the data quite well, taking into account the data uncertainty (of additional interest could be to perform forward simulations, that is, beyond the time span for which the data were taken). This should be no surprise since the model we used to estimate the dynamics is precisely the one we used to create to data. This does not always have to be the case, for example, if the data are generated by a complex chemical experiment for which we only have a much simpler model, as the next example shows.

*Remark 2.* In the above example, one might argue that trying to fit all parameters only using the measurements for substance A is not reasonable. However, in more complex examples, this is the typical situation: having a set of measurements and a given model with some unknown parameters where it is not clear whether it is possible to get a unique or narrowly-distributed parameter fit.

## 3.2. Example 2: Radical Polymerization

For the next example, we will address a situation which one often encounters in polymer research. Suppose there is a new



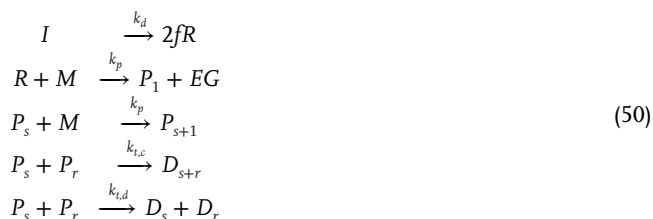
**Figure 7.** Illustration of uncertainty propagation: 90% percentile of concentrations of  $D$  for forward simulations with sampled parameters. We can see that the model is capable to explain the data within its uncertainty.

monomer  $M$  in whose basic kinetics in radical polymerization we are interested. First, you have to decide on the kinetic model, that is, what are the elementary steps which happen during this process. Radical polymerization is a chain process, that means the kinetic chain must be initiated, it must propagate, and finally terminate. Let us start with a rather simple model which one can easily pick off from text books of polymer science.

### 3.2.1. The Chemical Reaction Models

The initiator  $I$  decomposes to two radicals  $R$  with a rate coefficient  $k_d$  and an efficiency  $f$ , which considers that not all radicals  $R$  may start a kinetic chain because they are destroyed by some (unknown) side reactions before they can add the first monomer unit.

$R$  starts chains by adding the first monomer  $M$  giving a polymerization-active, “living” chain  $P_1$  of length 1. The active chains grow by adding one monomer after the other to chains  $P_s$  of length  $s$  (often named macroradical) until they terminate with each other yielding dead polymer  $D$ , the final product which you can isolate and sell. Termination can either be by combination of two active chains  $P_s$  and  $P_r$  yielding dead polymer  $D_{s+r}$  of length  $s+r$  or by disproportionation, when the chain length of the active chains are preserved yielding two dead chains  $D_s$  and  $D_r$ . When the initiator radical  $R$  starts a growing chain, the group  $R$  is incorporated as end group (EG) in the polymer chain, and to account for the concentration of these incorporated groups, a massless reaction product, a counter  $EG$ , is introduced as auxiliary quantity.



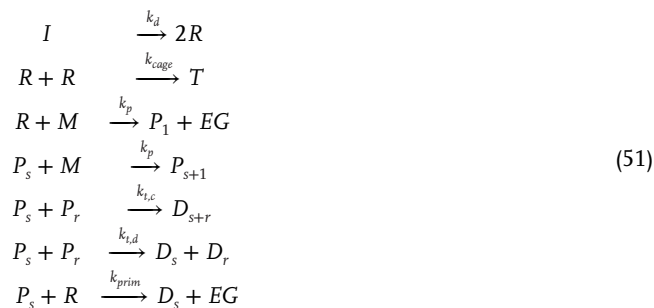
Reaction scheme for the assumed polymerization mechanism for which rate coefficients will be estimated using “experimental” data generated with the “real” model given in (49).

One has to keep in mind that the life-time of an active chain is in the order of seconds or less, so  $R$  must be continuously generated by decomposition of  $I$  to generate new active macroradicals which grow to chain lengths of some hundreds or thousands until they terminate.

It is obvious that the mode of termination has a strong influence on the chain length of the final polymer  $D$ , that is, of its molar mass. The same holds for the efficiency  $f$ , which determines how many chains will be effectively started, that is, on how many chains the polymerized monomer will be finally distributed. Thus, it is important to know the values of  $f$ ,  $k_{t,c}$ , and  $k_{t,d}$  to be able to design a new polymer grade with a desired chain length or molar mass.

We assume that we know the exact value of  $k_d$ , usually provided by the supplier of the initiator, and the value of the propagation rate coefficient  $k_p$  from some independent measurement.<sup>[46]</sup> The task now is to estimate values of the unknown coefficients  $f$ ,  $k_{t,c}$ , and  $k_{t,d}$ .

At this stage, one has to perform a couple of experiments and measurements. Here, we replace the experiment by simulation results, but with a somewhat more complex, “real” model including reactions which are responsible for the efficiency  $f$  of the initiating radicals  $R$ .



Reaction scheme for the “real” polymerization mechanism to generate “experimental” data together with rate coefficients and recipe from **Table 1**.

Typical initiators are so-called azo-initiators which upon heating give two small radicals  $R$  and one nitrogen molecule  $N_2$  which



**Table 1.** Upper: rate coefficients used in model (51) to produce the data. Below: recipe for generation of data.

Parameter	Value		
$k_d$	$7 \times 10^{-4} \text{ s}^{-1}$		
$k_{cage}$	$3 \times 10^{10} \text{ l mol}^{-1} \text{ s}^{-1}$		
$k_{prim}$	$5 \times 10^8 \text{ l mol}^{-1} \text{ s}^{-1}$		
$k_p$	$2 \times 10^3 \text{ l mol}^{-1} \text{ s}^{-1}$		
$k_{t,c}$	$2 \times 10^7 \text{ l mol}^{-1} \text{ s}^{-1}$		
$k_{t,d}$	$0 \text{ l mol}^{-1} \text{ s}^{-1}$		
Compound	$M \text{ [kg mol}^{-1}\text{]}$	$m \text{ [kg]}$	Conc. $[\text{mol L}^{-1}]$
M	0.1	0.2	2.0
I	0.2	0.002	0.01
S	0.1	0.8	7.98

are captured in a so-called solvent cage, that is, they are surrounded by solvent molecules. The reaction of such small radicals with each other occur near diffusion control, and so these two  $R$  may react with each other to give a non-reactive compound  $T$  before they diffuse out of this solvent cage to meet a monomer and start a growing chain. This is called the cage effect. Once they have left the solvent cage, these primary initiator radicals may also react with growing chains to dead polymer  $D$ , a reaction which is called primary radical termination. Again, EG gives the concentration of end groups in polymer chains. Contrary to the assumed model, end groups from the initiator radical  $R$  are generated by chain initiation as well as by primary radical termination. These two reactions add an inherent “initiator efficiency” to the reaction system, and depending on the value of  $k_{cage}$  and  $k_{prim}$  the amount of radicals available for chain initiation will be

reduced. Note that this efficiency is not constant throughout the reaction, in contrast to the simple model (50).

To generate “experimental” data, we use the “real” kinetic model (51), together with the rate coefficients given in Table 1, left. The kinetic model is translated to a system of ODEs assuming the mass balance for a batch reactor together with the mass action law of reaction rates and constant densities of  $1 \frac{\text{kg}}{\text{l}}$  for all compounds together with the initial reactor load and molar masses  $M$  for monomer  $M$ , initiator  $I$ , and solvent  $S$  given in Table 1, right. In the following, we denote the complete reactor setup as the recipe.

The data are shown in Figure 8 and the exact values in Table B1 in Appendix B1.

### 3.2.2. Estimation of Parameters

We will now estimate the parameters  $f$ ,  $k_{t,c}$ , and, later,  $k_{t,d}$ , using model (50) based on the data generated by model (51).

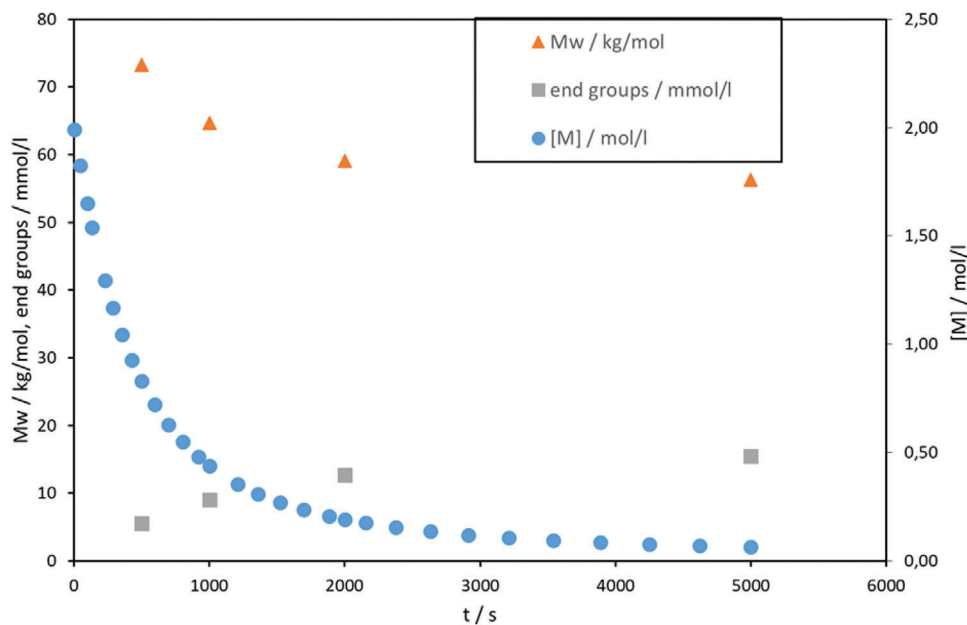
*Test 1: Fitting  $f$  and  $k_{t,c}$  with Monomer Concentration  $[M]$ :* At first, we estimate  $f$  and  $k_{t,c}$  only with the data of the monomer concentration  $[M]$ . We select as initial values  $f_0 = 1$  and  $(k_{t,c})_0 = 10^6$ .  $k_{t,d}$  is assumed to be 0 as it is in the real model (note that since both models differ this does not have to be a good value in the simple model). Using the Gauss–Newton with essential directions algorithm in Predici, we find as optimal results

$$f = 0.17, \quad k_{t,c} = 3.6 \times 10^6 \quad (52)$$

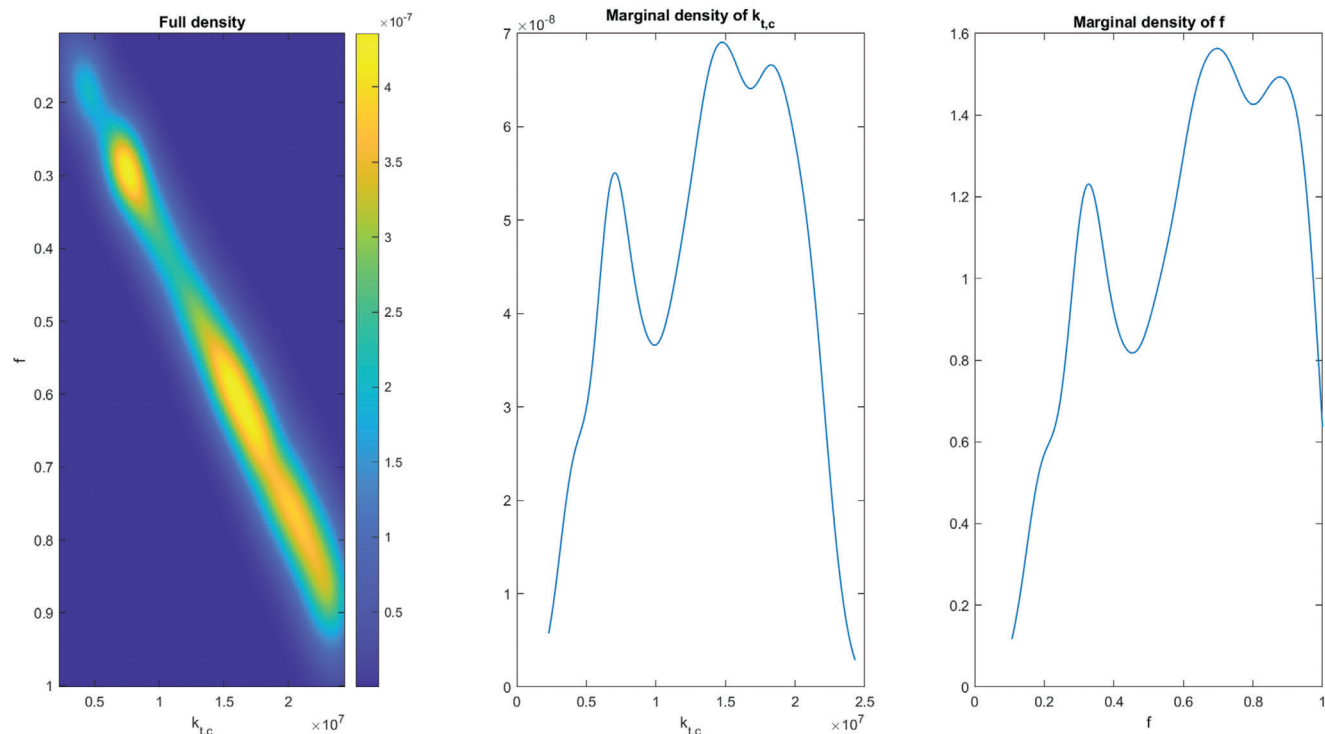
while

$$R_X = 0.130, \quad \text{condition } \kappa \approx 1000 \quad (53)$$

The high condition number should be a worrying sign. It indicates that parameters might be correlated so that there



**Figure 8.** Data from the “real” model (51) for monomer concentration  $[M]$ , weight average molar mass  $M_w$  and concentration of end groups of dead chains coming from the small radical  $R$  from reactions 3 and 7 of (51). The exact values can be found in Table B1.



**Figure 9.** Test 1: full approximated probability density of  $f$  and  $k_{t,c}$  (left), and marginal densities (middle and right). The images suggest a large range of probable choices for the parameters.

exist significantly different good parameters. For example, if we choose as initial parameters  $f_0 = 0.5$  and  $(k_{t,c})_0 = 10^7$ , the result is

$$f = 0.48, \quad k_{t,c} = 1.04 \times 10^7, \quad R_X = 0.132, \quad \kappa \approx 300 \quad (54)$$

Apparently, there exist very different parameters which give very similar residuals. If we now assume the data to be subject to measurement errors, it becomes immediately unclear, which of these two parameters is better suited in our model because, if the true data are slightly different from the observed data, then the parameters in (54) could easily yield a better residual than the ones in (52).

It is thus worthwhile to view the PE here from the Bayesian perspective and approximate the probability distribution. To this end, we assume to have made a measurement error of  $\sigma = 0.05$  and use the prescaled MALA. Note that here one could also use the grid-based approach since we only consider two different parameters. However, we also want to showcase the applicability of the more complex MALA approach which is better suited for high-dimensional parameters. Details on the prescaled MALA in this example are found below.

The ensuing distribution is visualized using KDE in **Figure 9**. As we can see, we cannot reliably estimate the parameters since the region of parameters with high probabilities is vast for both parameters, especially for  $f$ . Considering the structure of the chemical reaction, this should not come as a surprise. While a high efficiency  $f$  denotes a high fraction of effectively initiating radicals and thus a fast conversion of the monomer,  $k_{t,c}$  is the rate of the termination by combination during which radicals are destroyed. As such,  $f$  and  $k_{t,c}$  work against each other, yielding

multiple possible choices for them with similarly small residual. Together with the assumption of possible data measurement errors, this yields a large range of similarly probable parameters.

Please see Appendix B.2 for technical details of the application of the PE methods in this section.

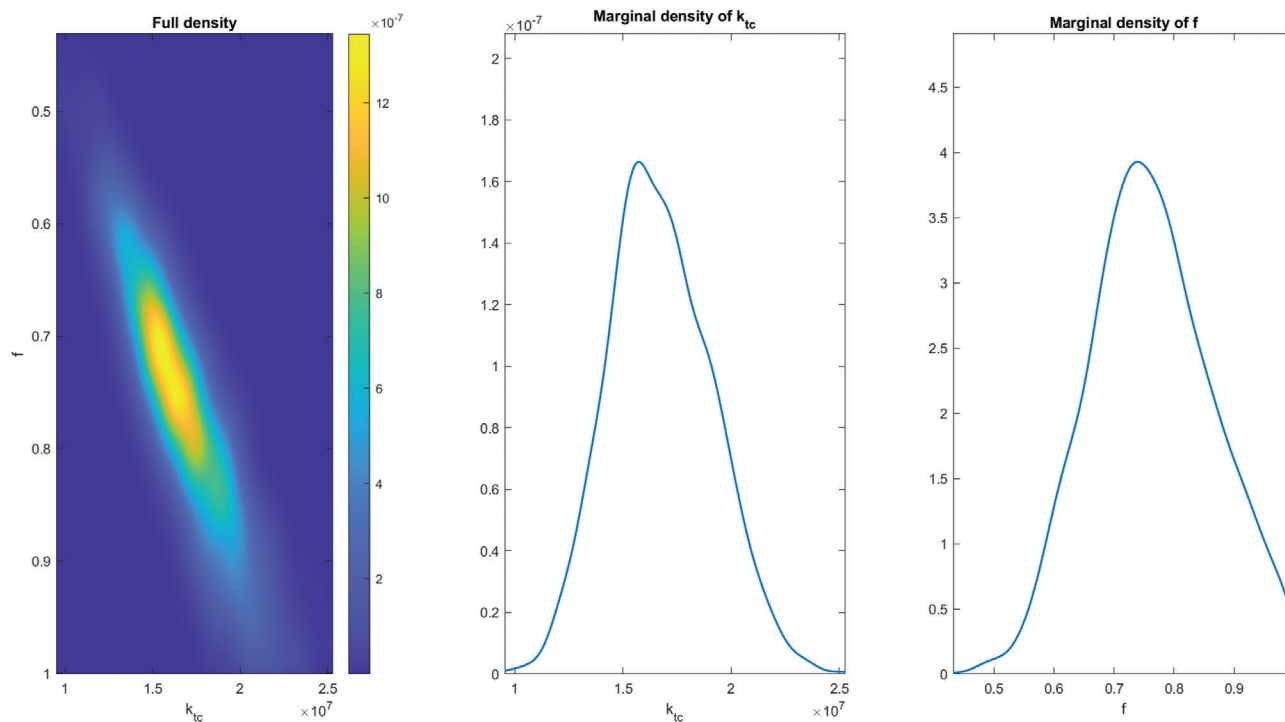
*Remark 3.* Note that, although  $f = 0.17$  together with  $k_{t,c} = 3.6 \times 10^6$  also gives a small residual, apparently parameters in its vicinity do not since only few samples were found in this region. This is, in fact, an artifact of the MALA algorithm: proposals which are close to this parameter are still rejected because there the residual is high. As a consequence, the small region around  $f = 0.17, k_{t,c} = 3.6 \times 10^6$  is slightly undervalued by the created sequence of samples.

*Test 2:  $f$  and  $k_{t,c}$  with  $[M]$  and  $M_w$ :* We include the four data points of the average molar mass  $M_w$  into the data and again perform parameter estimation with the essential directions approach in Predici. The result is

$$f = 0.72, \quad k_{t,c} = 1.6 \times 10^7, \quad R_X = 0.123, \quad \kappa \approx 3 \quad (55)$$

Now, with a very small condition number, we are able to determine  $f$  and  $k_{t,c}$ . Their values are also in accordance with the result of the Bayesian PE in the previous setting: 0.72 for  $f$  and  $1.6 \times 10^7$  for  $k_{t,c}$  were inside the set of probable values.

We again perform the Bayesian PE with prescaled MALA for this setting and find a much smaller range of probable parameters, which also includes the optimal parameters stated above (**Figure 10**). The optimal parameters are close to the maxima of the determined distribution. Apparently, the inclusion of the



**Figure 10.** Test 2: Full approximated probability density of  $f$  and  $k_{t,c}$  (left), and marginal densities (middle and right). With the inclusion of  $M_w$ , the range of probable parameters becomes much slimmer.

average molar mass  $M_w$  into the data has a strong effect on the reliability of the parameter estimation: the residual could be made small both with and without  $M_w$ , but the parameter region in which this is possible is significantly narrowed down by inclusion of  $M_w$ .

**Test 3:  $f, k_{t,c}$  and  $k_{t,d}$  with  $[M]$  and  $M_w$ :** We now assume the parameter  $k_{t,d}$  to be unknown, too. Note that we do not intend the different parts of this example to build on each other, but rather aim to show what happens when the knowledge we have about the model and the provided data varies. With  $[M]$  and  $M_w$  as data and initial values given by  $f_0 = 0.5, (k_{t,c})_0 = (k_{t,d})_0 = 10^7$ , we get similarly ambiguous results as in Test 1. For initial values  $f_0 = 0.5, (k_{t,c})_0 = 10^7, (k_{t,d})_0 = 1$  we find

$$f = 0.72, \quad k_{t,c} = 1.59 \cdot 10^7, \quad k_{t,d} = 1, \quad R_X = 0.125, \\ \kappa \approx 1400 \quad (56)$$

while for initial values  $f_0 = 0.5, (k_{t,c})_0 = 1, (k_{t,d})_0 = 10^6$  we get

$$f = 0.71, \quad k_{t,c} = 1.45 \cdot 10^7, \quad k_{t,d} = 1.06 \cdot 10^6, \quad R_X = 0.125, \\ \kappa \approx 8900 \quad (57)$$

Apparently, while  $f$  is nearly at the same value, there are multiple possibilities for combinations between  $k_{t,c}$  and  $k_{t,d}$ .

The implication of this needs to be stressed. Although these parameters are minima of the residual function, they need not be predictive: if we use them to simulate other properties of the reaction, for example, simply simulate forward in time, they might yield significantly different results than other optimal parameters. Hence, one would have to ask, which prediction would then

have to be expected? Taking the probabilities for the different parameters into account, we will see a range of probable scenarios. This will be illustrated in Test 4.

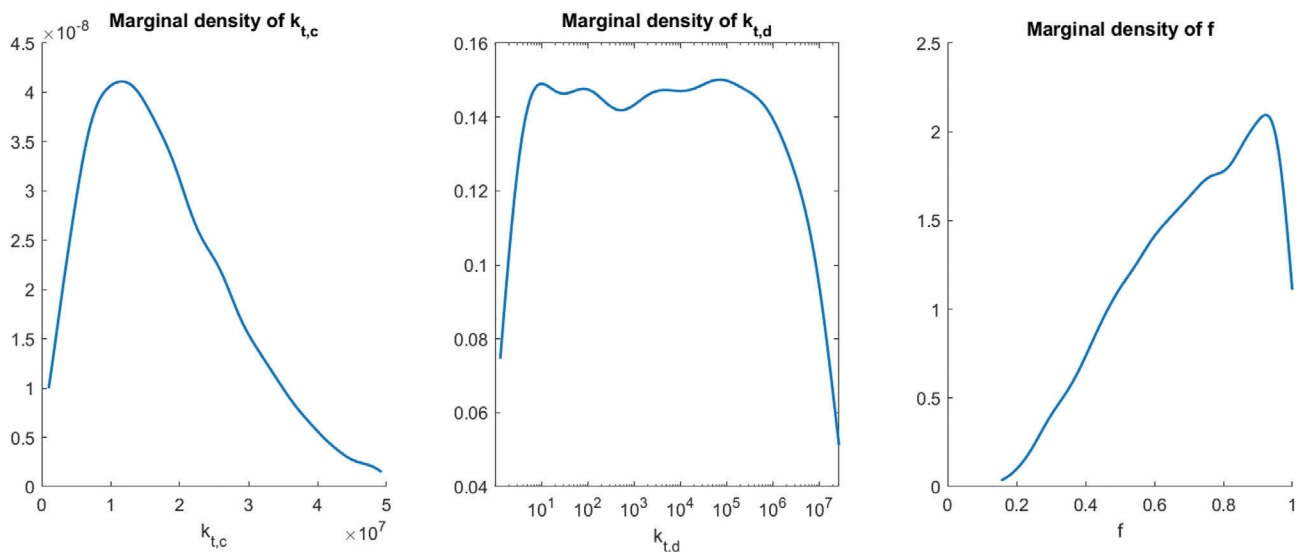
The Bayesian PE with  $\sigma = 0.05$  again sheds light on which parameters are how probable through the posterior distribution. As it turns out, we find a broad range of values with high probabilities (**Figure 11**). It implies that with  $k_{t,d}$  unknown and the data at hand, we cannot find a reliable estimate for the three parameters.

**Test 4:  $f, k_{t,c}$  and  $k_{t,d}$  with  $[M], M_w$ , End Groups and The Molar Mass Distribution:** We now include the end groups and full molar mass distribution into the data in the hope that it yields a more precise estimate of the parameters. With initial values given equally as in Test 3, we get as optimal parameters

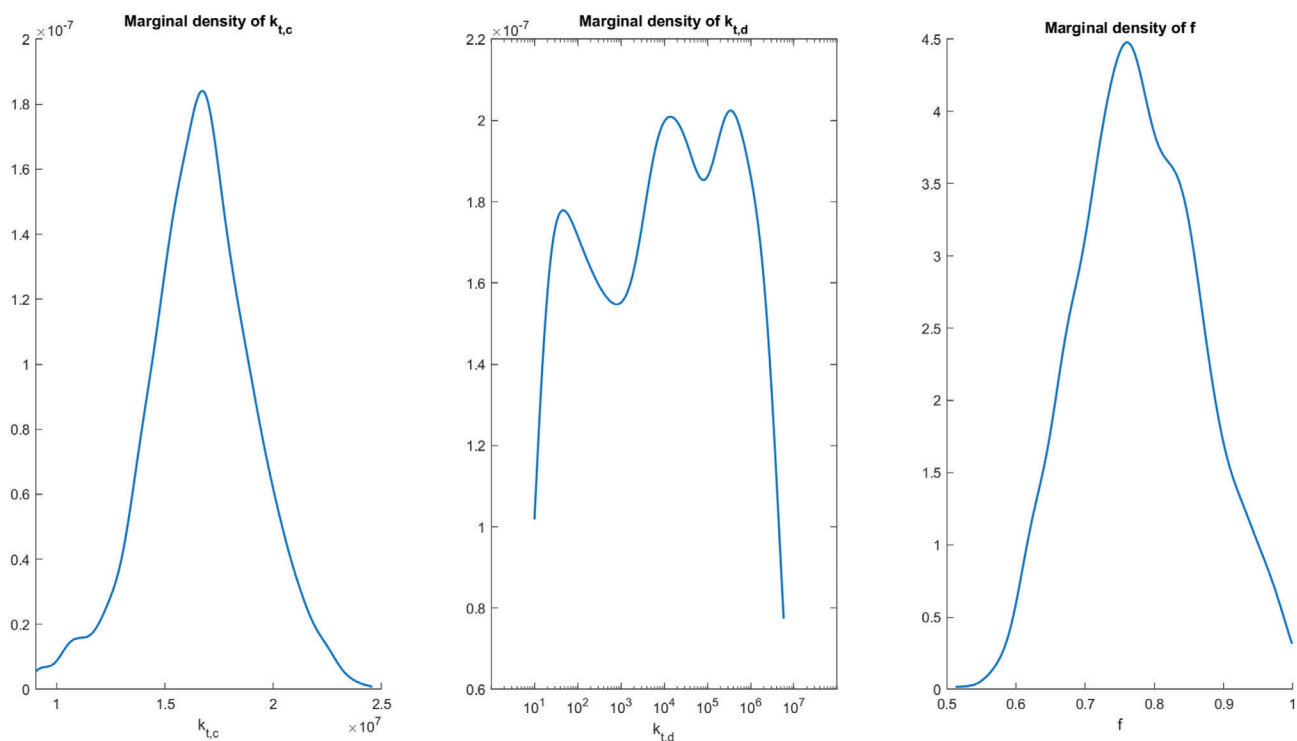
$$f = 0.76, \quad k_{t,c} = 1.66 \cdot 10^7, \quad k_{t,d} = 1.09 \cdot 10^5, \quad R_X = 0.124, \\ \kappa \approx 170 \quad (58)$$

These parameters are similar to the ones in Test 3, but they come with a much lower condition number. We see, however, that  $k_{t,d}$  is quite close to its initial value. This should make us suspicious since it could indicate that there exist many local minima with different values of  $k_{t,d}$  and simply the one closest to the initial value was found.

To reveal whether this is true or whether  $k_{t,d} = 1.09 \times 10^5$  is in fact the unique best choice, we again start the Bayesian PE at these optimal parameters. Compared to Test 3, the distributions of  $k_{t,c}$  and  $f$  become much slimmer and resemble normal distributions centered around the optimal parameters determined above (**Figure 12**).  $k_{t,d}$ , however, is distributed across a vast range



**Figure 11.** Test 3: Since  $k_{t,c}$  and  $k_{t,d}$  both impact the destruction of radicals and  $f$  counteracts these values being responsible for the number of initiating radicals, their values are highly non-unique.



**Figure 12.** Test 4: With data about the molar mass distribution and end groups included, the parameters can be much more reliably inferred since the distributions are much slimmer than in Test 3.  $k_{t,d}$  seems not to impact the outcome of the model much since many very different values seem probable.

of values from 10 to  $10^7$ . This is because there exist multiple different combinations of values for  $k_{t,d}$ ,  $k_{t,c}$ , and  $f$  which yield a similarly small residual. In other words, the parameters are correlated. This is also expressed by the fact that in the Gauss–Newton scheme, there existed two essential directions and not three. The range for the combinations, however, is much smaller for the latter two than for  $k_{t,d}$ .

With all data points included, we were able to learn important properties of our model. We could find optimal values for  $k_{t,c}$  and  $f$  by minimizing the residual function. Taking into account the possibility of measurement errors that perturb the data, we could quantify how this uncertainty affects the parameter estimation. We could also see that for  $k_{t,d}$  many different values, between 10 and  $10^7$ , are permitted.

**Table 2.** Second recipe used to generate another data set.

Compound	$M$ [kg mol <sup>-1</sup> ]	$m$ [kg]	Conc. [mol L <sup>-1</sup> ]
M	0.1	0.5	3.78
I	0.2	0.02	0.075
S	0.1	0.8	6.06

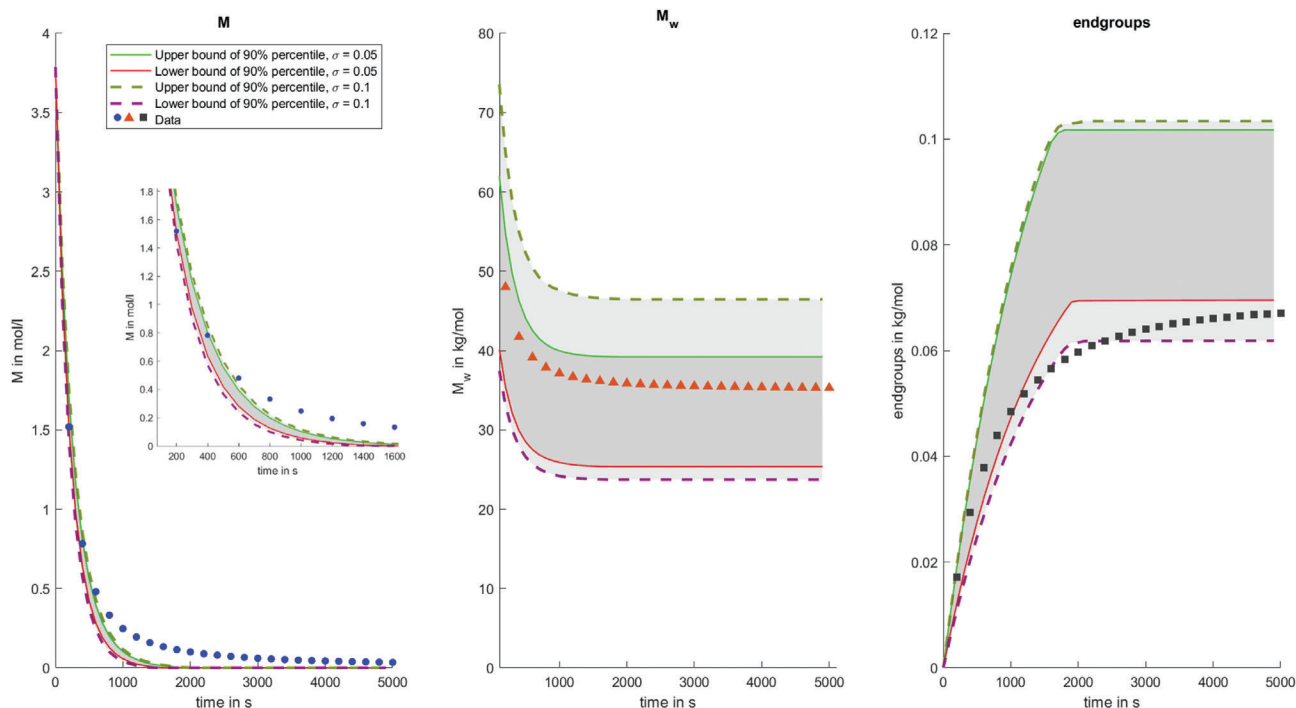
As the last step, we will now investigate the possible outplays of the model with other recipes based on their probabilities. In doing so, we also show how to verify whether the simple model is generally a good approximation to the true model, opposed to only for the recipe used in generation of the data depicted in Figure 8.

We generate data points with the true model (51) with a different recipe, given in Table 2.

Afterward, we make forward simulations with parameters with the simple model (50) which were sampled from the 3D probability distribution for the parameters and visualize the 90% percentile in each time step for these forward simulations (i.e., in each time step we show the boundaries above and beneath of which only 5% of the forward simulations lie). We compare these forward simulations with the data and want to check if the data lie inside the range of possible forward simulations. The aim is to find out whether the simple model can accurately approximate the true model for different recipes, taking into account that the parameter estimation is subject to uncertainty. If this is the case, it means that the parameters estimated on the basis of the data in Table B1 are applicable to other recipes, too. We find (Figure 13)

that while the evolutions of  $M_w$  and the end groups are reasonably well captured, the concentration of the monomer takes a quite different path, indicating that the simple model is not able to predict its concentration well. One could argue that maybe the data measurement error was actually higher than initially assumed so that we are led to the wrong parameters. In order to take this into account, we set  $\sigma = 0.1$ , again sample from the probability distribution corresponding to this value for  $\sigma$ —the probability distributions are not visualized here but naturally become broader—and generate predictions. As we can see, even assuming a high measurement error, that is, allowing the idea that many different parameters are candidates to be optimal, the simple model is not capable to accurately predict the evolution of the monomer concentration. This gives us an important message that the classical PE could not have sent: there exists a modeling error in the simple model (50) in regard to the true model (51); that is, there are components of the true, unknown model that are badly or not at all captured by the simple model. Hence, we must modify our simple model if we strive to forecast the monomer concentration. Of course, if we are only interested in  $M_w$  and end groups, then the simple model seems to be sufficient. For this, the Bayesian PE sheds light on where the outcomes of the true model lie with different recipes.

In order to improve the simple model, there does not exist a blueprint in the sense of a clear algorithmic way how to do that. It requires expertise in regard to the chemical model to find the missing parts of the simple model. This is also the reason why we do not take the discussion of this example further since we want to showcase specific generally recommendable steps one should take in order to gain maximal insight into the uncertainty of the



**Figure 13.** Test 4: 90% percentiles of forward simulation of model (50) with a different recipe for samples of the posterior distribution generated with  $\sigma = 0.05$  (solid lines; dark grey) and  $\sigma = 0.1$  (dashed lines; light grey) and data points generated with the true model (51). We see that even if we assume to have made a higher measurement error ( $\sigma = 0.1$  instead of 0.05), the simple model is still not enough to explain the evolution of the concentration of  $M$ .





estimated parameters and how it translates to forward simulation. For completion, for this model, it is the step of termination with  $R$  with parameter  $k_{prim}$  that should be included into the simple model to give a much better prediction for the monomer.

#### 4. Conclusion

In this article, we have illustrated, compared, and combined two different approaches to parameter estimation: (1) the classical approach that focuses on minimizing the residual function which measures the distance between the outcome of a model and the observed data, and (2) the Bayesian approach which quantifies the uncertainty that the parameters underlie. Both play an important role in estimation of parameters in chemical processes such as polymerization. In this article, we argued that only the interplay between both approaches allows reliable parameter estimation in typical real-world scenarios.

In this light, it is vital to understand that both approaches do not contradict but rather complement each other. Classical PE gives the answer to the more intuitive question: it tells us what the optimal parameter is that brings the model outcome closest to the observed data. Bayesian PE, on the other hand, assumes a probability distribution on the data which translates to a probability distribution of the parameters. It complements the classical PE in the sense that it tells us the degree of reliability with which we can assume to have found the optimal parameters. If one finds that the distribution allows a large region of probable parameters, it indicates that the data at hand are either not expressive enough or their measurement error is too high. Additionally, the parameter that was determined in classical PE can be used to make forecasts for future time points of the model simulation or entirely different recipes as done in Test 4 of Example 2. Plus, one has to take into account that the uncertainty in the parameters naturally translates to these forecasts. On the basis of the probability distribution of the parameters from Bayesian PE, we can estimate the probabilities for the outcomes of these new simulations. Further, this allows us to infer whether the model we use actually offers an accurate description of the real experiment that generated the data.

The uncertainty of the PE arises if one does not have perfect reliability of the data. It must be stressed that this data measurement error must be quantified by the experimentalist first (in the form of a value for  $\sigma$ ). Afterward, the data uncertainty translates to parameter uncertainty which itself translates to a range of possible model outcomes for other recipes or future points in time. If the data can be measured almost perfectly, the Bayesian PE will give overwhelming probability to the parameters that were determined as optimal in the classical PE and the probable model outcomes will lie in a very small range around the outcome that corresponds to the optimal parameters. On the contrary, of course, if the data underlie large measurement errors, then the result from classical PE does not deserve a large amount of trust and consequently the parameter distributions from Bayesian PE will be broad.

In order to communicate these characteristics of the approaches as well as their interplay, we explained their motivation, theory, and applicability in detail. We further introduced practical methods for both—the Gauss–Newton method with essen-

tial directions for classical PE and the novel sampling technique prescaled MALA for Bayesian PE. Last, we showcased how to apply both approaches and how to interpret the results in two examples of different complexity.

In total, the whole work flow to perform Parameter Estimation with Uncertainty Quantification we suggest in this article can be summarized as follows:

1. Generate data with an experiment.
2. Define a model, for example, of the form in Equation (1).
3. Estimate optimal parameters—that is, the global minimum of the residual function, defined, for example, as in Equation (6)—with Gauss–Newton with essential directions (see Section 2.2), using multiple different initial conditions since many optimization methods are prone to get stuck in local minima.
4. Quantify the data measurement error  $\sigma$ .
5. Define the prior distribution in the Bayesian framework as the data-independent intuition of the parameters and the likelihood function (see Section 2.3).
6. Sample from the ensuing posterior distribution, for example, using prescaled MALA (see Section 2.3 and Appendix A.2) with initial values chosen as the optimal values from classical PE.
7. Visualize the parameter distribution (for parameters with dimension bigger than 2 the marginal distribution), for example, using KDE (Equation (48)).
8. Use sampled parameters to investigate probable outcomes of the model for future points in time or different recipes.
9. If the experiment (respectively its model formulation) with which the data were generated is unknown and it is unclear whether the used model is a good description, take the real experiment again with a different recipe and compare the range of predicted model outcomes with the newly generated data. If the data lie far outside the range of probable model outcomes, this indicates that the used model is insufficient to describe the real experiment. In this case, problem-specific expert knowledge is required to find an improvement.

All these steps are implemented and were executed in Predici.

While in this article we focused on the theoretical foundations of classical and Bayesian PE, in part 2 of this article series, we will apply both approaches to a more comprehensive example and illustrate all the details and specifics one has to take into account to infer the maximal amount of information about parameters, predictions, and model quality from the data.

### Appendix A: Additional Information on Metropolis-Adjusted Langevin Algorithm

#### A.1. Fundamentals of Metropolis Sampling

##### A.1.1. The Metropolis-Hastings Algorithm

The Metropolis-adjusted Langevin Algorithm (MALA) is a special case of the Metropolis-Hastings algorithm (MH).<sup>[41]</sup> MH generates a sequence of states that are distributed by a desired probability distribution  $\pi$  by taking subsequent steps according to a

probability density  $T(\theta_{k+1}, \theta_k)$ , that is, if  $\theta_k = \theta$ , then  $\theta_{k+1}$  is distributed by  $T(\cdot, \theta)$ . It is required that  $\pi$  be the *invariant distribution* of this sequence. As the name suggests, this means that the density of parameters does not change over time, that is, the probability density for  $\theta_k$  is identical to the one of  $\theta_{k+1}$ . Formally, the invariant distribution fulfills

$$\pi(\theta) = \int \pi(\theta') T(\theta, \theta') d\theta' \quad (\text{A1})$$

The sequence is generated by drawing proposals

$$\tilde{\theta}_{k+1} \sim q(\cdot, \theta_k) \quad (\text{A2})$$

from a *proposal distribution*  $q$ . The choice for  $q$  is almost arbitrary. The only restrictions on  $q$  are that it be non-negative and that it holds  $q(\theta', \theta) > 0 \Leftrightarrow q(\theta, \theta') > 0$ .

This proposal is then accepted with probability

$$\alpha(\tilde{\theta}_{k+1}, \theta_k) = \min\left\{1, \frac{\pi(\tilde{\theta}_{k+1}) q(\theta_k, \tilde{\theta}_{k+1})}{\pi(\theta_k) q(\tilde{\theta}_{k+1}, \theta_k)}\right\} \quad (\text{A3})$$

If  $q$  is symmetric, for example, a Gaussian distribution, the  $q$ -terms cancel out and can be omitted.

With this, we can make the following observation: let  $T(\theta', \theta)$  be the density of making a transition from  $\theta$  to  $\theta'$  in the sequence generated by MH. For  $T$  it holds

$$T(\theta', \theta) = q(\theta', \theta) \alpha(\theta', \theta) \quad (\text{A4})$$

since in order to reach  $\theta'$  from  $\theta$ ,  $\theta'$  has to be drawn as the proposal (density  $q(\cdot, \theta)$ ) and additionally has to be accepted (with probability  $\alpha(\theta', \theta)$ ). This yields

$$\begin{aligned} T(\theta', \theta) &= q(\theta', \theta) \alpha(\theta', \theta) \\ &= q(\theta', \theta) \min\left\{1, \frac{\pi(\theta') q(\theta, \theta')}{\pi(\theta) q(\theta', \theta)}\right\} \\ \Rightarrow \pi(\theta) T(\theta', \theta) &= \pi(\theta) q(\theta', \theta) \min\left\{1, \frac{\pi(\theta') q(\theta, \theta')}{\pi(\theta) q(\theta', \theta)}\right\} \\ &= \min\{\pi(\theta) q(\theta', \theta), \pi(\theta') q(\theta, \theta')\} \end{aligned} \quad (\text{A5})$$

Analogously, we obtain

$$\pi(\theta') T(\theta, \theta') = \min\{\pi(\theta') q(\theta, \theta'), \pi(\theta) q(\theta', \theta)\} \quad (\text{A6})$$

which yields

$$\pi(\theta') T(\theta, \theta') = \pi(\theta) T(\theta', \theta) \quad (\text{A7})$$

This is the detailed balance property which ensures

$$\begin{aligned} \int \pi(\theta') T(\theta, \theta') d\theta' &= \int \pi(\theta) T(\theta', \theta) d\theta' = \pi(\theta) \underbrace{\int T(\theta', \theta) d\theta'}_{=1} \\ &= \pi(\theta). \end{aligned} \quad (\text{A8})$$

As a direct consequence, MH creates a sequence of states for which  $\pi$  is the invariant distribution. Thus, we can draw states from this sequence whose distribution in fact converges to  $\pi$ .

### A.1.2. The Langevin Sampler

We can address the Metropolis-adjusted Langevin algorithm from a different perspective. For a stochastic differential equation (SDE) of the form

$$d\theta_t = -\text{grad} V(\theta_t) dt + \sqrt{2} dB_t \quad (\text{A9})$$

where  $B_t$  is the Standard Brownian motion, it holds that the invariant distribution of realizations of this SDE is given by

$$\pi(\theta) = \frac{1}{Z} \exp(-V(\theta)) \quad (\text{A10})$$

Since the solution of an SDE is not a discrete sequence of states but rather a time-continuous function in time, invariant distribution here means that the density of  $\theta_t$  having a certain value is independent of  $t$ . Over time, the density of states converges to this invariant distribution. As a consequence, by setting  $V = -\log(\pi)$ , we find that the SDE has invariant distribution

$$\exp(\log(\pi)) = \pi \quad (\text{A11})$$

We should therefore be interested in creating realizations of the SDE

$$d\theta_t = \text{grad} \log(\pi) + \sqrt{2} dB_t. \quad (\text{A12})$$

In Section 2.3, we introduced  $\pi$  as equal to the function  $\frac{1}{Z} \exp(-S_X)$ . Thus, the SDE becomes

$$d\theta_t = -\frac{1}{Z} \text{grad} S_X(\theta_t) + \sqrt{2} dB_t \quad (\text{A13})$$

A common method to create realizations of SDEs is the Euler-Maruyama method.<sup>[47]</sup> It approximates the SDE by taking discretized steps of the form

$$\theta_{k+1} = \theta_k - \Delta t \text{grad} S_X(\theta_k) + \sqrt{2\Delta t} r_k \quad (\text{A14})$$

where  $r_j$  is a normally distributed random variable with mean 0 and variance 1. The normalization constant  $\frac{1}{Z}$  simply is accounted for by the step size  $\Delta t$ .

While in the original SDE, after some time, the  $\theta_k$  should be distributed by  $\exp(-S_X)$ , this property might be violated for the states of this sequence due to approximation errors induced by the step size  $\Delta t$ . By choosing a small step size, one can install convergence of the distribution of the  $\theta_k$  to a distribution that is closer to  $\exp(-S_X)$ .

Until here, we have not used the acceptance probability  $\alpha$ . In fact, creating a sequence of states from the approximated SDE is referred to as the Unadjusted Langevin Algorithm (ULA). In order to achieve that the detailed balance property is fulfilled and that the sequence thus has the desired distribution as its invariant

distribution, in MALA one extends ULA by the Metropolis-step of only accepting a candidate  $\tilde{\theta}_{k+1}$  with probability

$$\alpha = \min \left\{ 1, \frac{e^{-S_X(\tilde{\theta}_{k+1})} q(\theta_k, \tilde{\theta}_{k+1})}{e^{-S_X(\theta_k)} q(\tilde{\theta}_{k+1}, \theta_k)} \right\} \quad (\text{A15})$$

where

$$q(\theta', \theta) = \exp \left( -\frac{1}{4\Delta t} \|\theta' - \theta + \Delta t \text{grad} S_X(\theta)\|^2 \right) \quad (\text{A16})$$

The proposal distribution  $q$  has this form because in order to draw  $\tilde{\theta}_{k+1}$ , we move to  $\theta_k - \Delta t \text{grad} S_X(\theta_k)$  and then add a number that is normally distributed with mean 0 and variance  $2\Delta t$ . Thus, the density of  $\tilde{\theta}_{k+1}$  is a normal distribution around  $\theta_k - \Delta t \text{grad} S_X(\theta_k)$  with variance  $2\Delta t$ , which is precisely reflected in  $q$ .

In summary, MALA combines two different algorithms that have the same goal: the Metropolis–Hastings algorithm and the approximation of a solution of an appropriate SDE. In both algorithms, a sequence is generated whose states are distributed by a chosen distribution. Technically, the use of MH would suffice for that. In MH, a bad choice of  $q$  can force the candidates to mainly come from low regions of  $\pi$  so that few candidates are accepted and it would take long for the states of the sequence to be distributed by  $\pi$ . Through the use of the gradient step in MALA, candidates are usually chosen in regions of interest, yielding a faster convergence.

## A.2. Prescaled MALA

We present here the exact derivation of the step sizes in the prescaled MALA. In the algorithm, a step has the form

$$\tilde{\theta}_{k+1} = \theta_k - P \text{grad} S_X(\theta_k) + \sqrt{2P^{\frac{1}{2}}} r_k, \quad r_k \sim \mathcal{N}(0, Id) \quad (\text{A17})$$

$Id$  is the  $d \times d$  identity matrix.

The proposal distribution  $q$  is then given by

$$q(\theta', \theta) = \exp \left( -\frac{1}{4} (\theta' - \theta + P \text{grad} S_X(\theta))^T \right. \\ \left. \times P^{-1} (\theta' - \theta + P \text{grad} S_X(\theta)) \right) \quad (\text{A18})$$

The choice of  $P$  is done in the following way:

1. Specify the desired average length of a step in each of the  $d$  parameter directions, denoted by  $\tau_1, \dots, \tau_d$ . The term “average” is used with respect to the posterior distribution from which the algorithm samples. Over all samples of the resulting time series, the step taken toward the proposal should have average length  $\tau_i$  in direction  $i$ . A step consists of a deterministic part  $P \text{grad} S_X(\theta_k)$  and a stochastic part  $\sqrt{2P^{\frac{1}{2}}} r_k$ , which both contribute to the length of the step. The average, or expected, step length in direction  $i$ , dependent on  $P_{ii}$ , is given by

$$\mathbb{E}(|-P_{ii} \text{grad} S_X(\theta_k) + \sqrt{2P_{ii}} (r_k)_i|) \quad (\text{A19})$$

It is difficult to find an analytical expression to make this property be equal to a chosen  $\tau_i$ . Still, we can find an upper bound by

$$\begin{aligned} \mathbb{E}(|-P_{ii} \text{grad} S_X(\theta_j) + \sqrt{2P_{ii}} (r_k)_i|) \\ \leq \mathbb{E}(|P_{ii} \text{grad} S_X(\theta_j)|) + \mathbb{E}(|\sqrt{2P_{ii}} (r_k)_i|) \\ = \mathbb{E}(|P_{ii} \text{grad} S_X(\theta_j)|) + 2\sqrt{\frac{P_{ii}}{\pi}} \end{aligned} \quad (\text{A20})$$

because of the triangle inequality and the fact that

$$\mathbb{E}(|\sqrt{2P_{ii}} (r_k)_i|) = \sqrt{2P_{ii}} \sqrt{\frac{2}{\pi}} = 2\sqrt{\frac{P_{ii}}{\pi}} \quad (\text{A21})$$

This holds because for a normally distributed random variable  $r \sim \mathcal{N}(0, 1)$ , it holds that  $\mathbb{E}(|cr|) = c\sqrt{\frac{2}{\pi}}$ .

2. Draw  $n$  points  $\theta_1, \dots, \theta_n$  randomly from the parameter space and compute the gradients  $\text{grad} S_X(\theta_1), \dots, \text{grad} S_X(\theta_n)$ . In this article, we always used  $n = 100$ .
3. In each direction  $1, \dots, d$ , compute the average length of the gradient

$$\mathbb{E}((\text{grad} S_X)_i) \approx \overline{(\text{grad} S_X)_i} = \frac{\sum_{k=1}^n \exp(-S_X(\theta_k)) |\text{grad} S_X(\theta_k)_i|}{\sum_{s=1}^n \exp(-S_X(\theta_s))} \quad (\text{A22})$$

4. We can then solve

$$P_{ii} \overline{(\text{grad} S_X)_i} + 2\sqrt{\frac{P_{ii}}{\pi}} = \tau_i \quad (\text{A23})$$

which can be transformed to

$$P_{ii} + 2\frac{\sqrt{P_{ii}}}{\sqrt{\pi}(\text{grad} S_X)_i} - \frac{\tau_i}{(\text{grad} S_X)_i} = 0 \quad (\text{A24})$$

Solving the quadratic equation for  $Q_i := \sqrt{P_{ii}}$  we get

$$Q_i = \frac{-1}{\sqrt{\pi}(\text{grad} S_X)_i} + \sqrt{\frac{1}{\pi(\text{grad} S_X)_i^2} + \frac{\tau_i}{(\text{grad} S_X)_i}} \quad (\text{A25})$$

so that

$$P_{ii} = \left( \frac{-1}{\sqrt{\pi}(\text{grad} S_X)_i} + \sqrt{\frac{1}{\pi(\text{grad} S_X)_i^2} + \frac{\tau_i}{(\text{grad} S_X)_i}} \right)^2 \quad (\text{A26})$$

If desired, one can easily replace the average gradient length by a maximal or minimal gradient length in step 3. Note that the time step is not adaptively chosen but still determined in advance. This yields that the property of the algorithm that we exploit, namely

that it produces the desired distribution of samples, is not damaged.

### A.3. Details on Kernel Density Estimation

The Kernel function  $K$  should be nonnegative, monotonically increasing as  $\theta$  approaches 0, and its integral over  $\mathbb{R}^d$  should be equal to 1. A typical choice is

$$K(\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\|\theta\|_2^2\right) \quad (\text{A27})$$

so that  $K(H^{-1}(\theta_k - \theta)) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\|H^{-1}(\theta_k - \theta)\|_2^2\right)$

With the bandwidths, one can regulate the influence of samples on  $\hat{p}_{KDE}(\theta_k)$  depending on the distance between  $\theta_k$  and  $\theta$ . In case that  $p$  is a Gaussian function, a recommended choice for  $h_i$  is given by [48]

$$h_i^* = 1.06v_i n^{-1/5} \quad (\text{A28})$$

where  $v_i$  is the standard deviation of the  $i$ th entries of the samples. It indicates that the more samples one has at hand, the smaller the bandwidths should be chosen.

With KDE, we can then evaluate  $\hat{p}_{KDE}$  on a fine grid through Equation (48) and obtain a fine visualization of an approximation of  $p$ . Note that in general  $\hat{p}_{KDE}(\theta_k)$  is unequal to  $p(\theta_k)$  but can be brought closer by decreasing the  $h_i$ . This in turn is already advised by Equation (A28) but needs a sufficient number of samples.

## Appendix B: Details on The Radical Polymerization Example

### B.1. Data in Section 3.2

#### B.2. Technical Details on The Application of PE Methods in Section 3.2

##### B.2.1. Test 1: Fitting $f$ and $k_{t,c}$ with Monomer Concentration $[M]$

We used the prescaled MALA with 12 000 steps, out of which we discarded the first 2000 to give the sequence of points time to lose dependence on the initial values. This number of steps is high for only two parameters compared to Test 2. The explanation is natural: Such a high number becomes necessary if the region of parameters with high probabilities is large because it takes more time to comb this entire region compared to a small region. The initial values were set to the optimal parameters determined above. We assume a measurement error of  $\sigma = 0.05$  and a uniform prior with bounds given by  $[0.1, 1]$  for  $f$  and  $[10^4, 10^9]$  for  $k_{t,c}$ . In the pre-scaled MALA, we set  $\tau_{1,2}$  so that in each direction a step has the average length of one 200th of the length of its parameter domain. The acceptance ratio was 72%. In the Gauss–Newton scheme with essential directions, the number of essential directions was 1 for both sets of initial values.

**Table B1.** Data set produced with model (51).

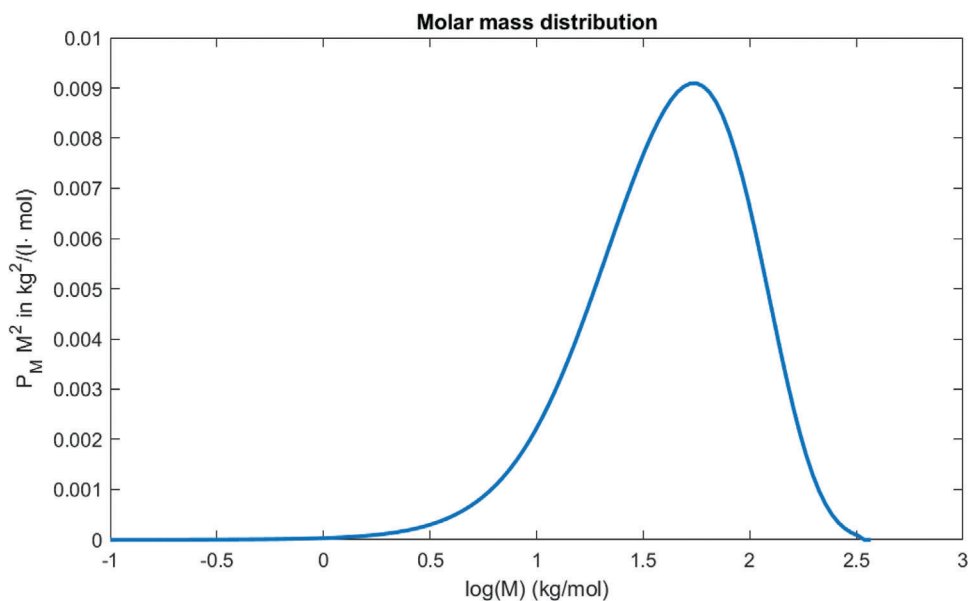
Time in s	$[M]$ / mol/l	$M_w$ / kg/mol	end groups / mmol/l
1	1.99		
44	1.83		
96	1.65		
132	1.54		
227	1.3		
286	1.17		
353	1.04		
429	0.93		
500	0.83	73	5.5
597	0.72		
698	0.63		
805	0.55		
921	0.48		
1000	0.44	65	9.0
1210	0.35		
1360	0.31		
1520	0.27		
1696	0.24		
1887	0.21		
2000	0.19	59	12.7
2153	0.17		
2379	0.15		
2634	0.14		
2912	0.12		
3214	0.11		
3539	0.09		
3886	0.08		
4249	0.08		
4624	0.07		
5000	0.06	56	15.5

##### B.2.2. Test 2: $f$ and $k_{t,c}$ with $[M]$ and $M_w$

In the Gauss–Newton with essential directions algorithm, the number of essential directions was 2. We used prescaled MALA with 6000 steps, out of which we discarded the first 1000, assuming a measurement uncertainty of  $\sigma = 0.05$ . The initial values were the same in Test 1. The acceptance ratio was 57%.

##### B.2.3. Test 3: $f$ , $k_{t,c}$ and $k_{t,d}$ with $[M]$ and $M_w$

In the Gauss–Newton algorithm, the number of essential directions was 2. As prior distribution, we again used a uniform distribution over the interval used so far for  $f$  and  $k_{t,c}$ , while for  $k_{t,d}$  we used the interval  $[1, 10^8]$ . As initial values, we chose  $f_0 = 0.5$ ,  $(k_{t,c})_0 = 10^7$ ,  $(k_{t,d})_0 = 1$ . We performed 6000 steps of which we discarded the first 1000 to reduce dependence on the initial values. The acceptance ratio was 72%. Since the range for  $k_{t,d}$  was unclear a priori, we replaced it by its logarithm to the base 10 and transformed it back for the visualization.



**Figure B1.** Molar mass distribution (MMD) generated with the “real” model in equation (51) and rate coefficients and recipe of Table . MMD is given as the outcome of gel permeation chromatography (GPC), the state of the art method to determine MMD. For various representations of MMD, see for example ref. [49].

#### B.2.4. Test 4: $f$ , $k_{t,c}$ and $k_{t,d}$ with $[M]$ , $M_w$ , End Groups and The Molar Mass Distribution

In the Gauss–Newton algorithm, the number of essential directions was 2. As prior distribution, we again used a uniform distribution over the interval used so far for  $f$  and  $k_{t,c}$ , while for  $k_{t,d}$  we used the interval  $[10, 10^8]$ . As initial values, we chose  $f_0 = 0.5$ ,  $(k_{t,c})_0 = 1e7$ ,  $(k_{t,d})_0 = 1$ . Again, we performed 6000 steps of which we discarded the first 1000. The acceptance ratio was 71%.

### Acknowledgements

This research has been funded by Deutsche Forschungsgemeinschaft (DFG) through grant CRC 1114 “Scaling Cascades in Complex Systems” Project Number 235221301, as well as through project EF4-5 of the DFG Cluster of Excellence MATH+.

Open access funding enabled and organized by Projekt DEAL.

### Conflict of Interest

The authors declare no conflict of interest.

### Data Availability Statement

Research data are not shared.

### Keywords

deterministic and Bayesian parameter estimation, experimental data, polymerization kinetics, uncertainty quantification

Received: March 12, 2021  
Revised: May 31, 2021  
Published online: June 13, 2021

- [1] K.-D. Hungenberg, M. Wulkow, *Modeling and Simulation in Polymer Reaction Engineering: A Modular Approach*, Wiley VCH, New York Weinham **2018**.
- [2] N. Kazemi, T. Duever, A. Penlidis, *Macromol. React. Eng.* **2011**, *5*, 385.
- [3] R. Bindlish, J. Rawlings, R. Young, *AIChE J.* **2003**, *49*, 2071.
- [4] S. Roebnitz, P. Deuffhard, *A Guide to Numerical Modelling in Systems Biology*, Springer, New York **2015**.
- [5] N. Galagali, Y. Marzouk, *Chem. Eng. Sci.* **2015**, *123*, 170.
- [6] M. A. Katsoulakis, P. Vilanova, *J. Comput. Phys.* **2020**, *401*, 108997.
- [7] P. Loskot, K. Atitey, L. Mihaylova, *Front. Genet.* **2019**, *10*, 549.
- [8] A. Overstall, D. Woods, K. Martin, *Comput. Stat. Data Anal.* **2019**, *132*, 126.
- [9] S. Masoumi, T. A. Duever, A. Penlidis, R. Azimi, P. López-Domínguez, E. Vivaldo-Lima, *Macromol. Theory Simul.* **2018**, *27*, 1800016.
- [10] S. Matera, W. F. Schneider, A. Heyden, A. Savara, *ACS Catal.* **2019**, *9*, 6624.
- [11] L. Hakim, G. Lacaze, M. Khalil, K. Sargsyan, H. Najm, J. Oefelein, *Combust. Theory Modell.* **2018**, *22*, 446.
- [12] C. Schillings, B. Sprungk, P. Wacker, *Numer. Math.* **2020**, *145*, 915.
- [13] J. Bell, M. Day, J. Goodman, R. Grout, M. Morzfeld, *Combust. Flame* **2019**, *205*, 305.
- [14] C. Schillings, M. Sunnaker, J. Stelling, C. Schwab, *PLoS Comp. Biol.* **2015**, *11*, e1004457.
- [15] P. Naik, P. Pandita, S. Aramideh, I. Bilonis, A. M. Ardekani, **2019**, *23*, 981.
- [16] T. Minami, M. Kawata, T. Fujita, K. Murofushi, H. Uchida, K. Otori, Y. Okuno, *MRS Adv.* **2019**, *4*, 1125.
- [17] A. Nabifar, N. T. McManus, E. Vivaldo-Lima, A. Penlidis, *Macromol. Symp.* **2011**, *302*, 90.
- [18] A. J. Scott, A. Nabifar, C. M. R. Madhuranthakam, A. Penlidis, *Macromol. Theory Simul.* **2015**, *24*, 13.
- [19] M. Wulkow, *Macromol. React. Eng.* **2008**, *2*, 461.
- [20] E. Vafa, M. Shahrokhi, H. Abedini, *Chem. Eng. Commun.* **2013**, *200*, 20.





- [21] C. Shalizi, in *Advanced Data Analysis from an Elementary Point of View*, Cambridge University Press, England **2013**.
- [22] Y. Yang, F. Ye, *Front. Math. China* **2013**, *8*, 695.
- [23] A. van den Bos, *Numerical Methods for Parameter Estimation*, John Wiley & Sons, New York **2007**, pp. 163–210.
- [24] L. Xu, *Adv. Mech. Eng.* **2017**, *9*, 1.
- [25] R. Aster, B. Borchers, C. Thurber, *Parameter Estimation and Inverse Problems*, vol. 90, Academic Press, San Diego, CA **2005**.
- [26] S. M. Safdarnejad, J. Gallacher, J. Hedengren, *Comput. Chem. Eng.* **2015**, *86*.
- [27] L. T. Biegler, *Nonlinear Programming*, Society for Industrial and Applied Mathematics, Philadelphia, PA **2010**.
- [28] D. Constales, G. S. Yablonsky, D. R. D'hooge, J. W. Thybaut, G. B. Marin, *Advanced Data Analysis and Modelling in Chemical Engineering*, 1st Edition, Elsevier, Netherlands **2016**, pp. 285–306.
- [29] R. Penrose, *Math. Proc. Cambridge Philos. Soc.* **1955**, *51*, 406.
- [30] G. H. Golub, C. F. Van Loan, *Matrix Computations*, 4th ed., The Johns Hopkins University Press, Baltimore, MD **2012**.
- [31] P. Deuffhard, A. Hohmann, *Numerical Analysis in Modern Scientific Computing*, 2nd ed., Springer-Verlag, New York **2003**.
- [32] R. Telgmann, Ph.D. Thesis, FU Berlin **2008**.
- [33] A. M. Stuart, *Acta Numer.* **2010**, *19*, 451.
- [34] C. Schillings, C. Schwab, *Inverse Probl.* **2014**, *30*, 065007.
- [35] P. Saracco, M. Pia, *J. Phys.: Conf. Ser.* **2013**, *513*, 022033.
- [36] T. Dodwell, C. Ketelsen, R. Scheichl, A. Teckentrup, *SIAM Rev.* **2019**, *61*, 509.
- [37] G. Roberts, J. Rosenthal, *J. R. Stat. Soc. B* **1998**, *60*, 255.
- [38] M. Dashti, A. Stuart, in *Handbook of Uncertainty Quantification* (Eds: R. Ghanem, D. Higdon, H. Owhadi), Springer, New York **2017**.
- [39] H. AlRachid, L. Mones, C. Ortner, *SMAI J. Comput. Math.* **2018**, *4*, 57.
- [40] A. B. Duncan, N. Nuesken, G. Pavliotis, *J. Stat. Phys.* **2017**, *169*, 1098.
- [41] S. Chib, E. Greenberg, *Am. Stat.* **1995**, *49*, 327.
- [42] I. P. Cornfeld, S. V. Fomin, Y. G. Sinai, *Ergodic Theory*, Springer, **1982**.
- [43] A. Durmus, G. Roberts, G. Vilmart, K. Zygalakis, *Ann. Appl. Probab.* **2015**, *27*, 2195.
- [44] E. Parzen, *Ann. Math. Statist.* **1962**, *33*, 1065.
- [45] Y.-C. Chen, *Biostat. Epidemiol.* **2017**, *1*, 161.
- [46] O. Olaj, I. Bitai, F. Hinkelmann, *Makromol. Chem.* **2003**, *188*, 1689.
- [47] P. Kloeden, E. Platen, *The Numerical Solution of Stochastic Differential Equations*, vol. 23, Springer, New York **2011**.
- [48] B. U. Park, J. S. Marron, *J. Am. Stat. Assoc.* **1990**, *85*, 66.
- [49] R. Hutchinson, M. Aronson, J. Richards, *Macromolecules* **1993**, *26*, 6410.
- [50] D. Constales, G. S. Yablonsky, D. R. D'hooge, J. W. Thybaut, G. B. Marin, *Advanced Data Analysis & Modelling in Chemical Engineering*, Elsevier, Netherlands **2016**, pp. 285–306.