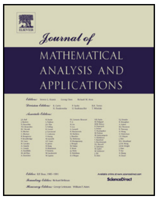
J. Math. Anal. Appl. ••• (••••) •••–•••

Contents lists available at ScienceDirect

Journal of Mathematical Analysis and Applications

www.elsevier.com/locate/jmaa



# Parameter estimation in SDEs via the Fokker–Planck equation: Likelihood function and adjoint based gradient computation

Barbara Kaltenbacher [a,*], Barbara Pedretscher [b]

[a] Department of Mathematics, Alpen-Adria-Universität Klagenfurt, Universitätsstraße 65-67, 9020 Klagenfurt, Austria
[b] KAI – Kompetenzzentrum Automobil- und Industrieelektronik GmbH, Europastraße 8, 9524 Villach, Austria

ARTICLE INFO

*Article history:*
Received 1 February 2018
Available online xxxx
Submitted by P. Sacks

*Keywords:*
Parameter identification
Stochastic differential equation
State space model
Likelihood function
Adjoint method

ABSTRACT

In this paper we consider the problem of identifying parameters in stochastic differential equations. For this purpose, we transform the originally stochastic and nonlinear state equation to a deterministic linear partial differential equation for the transition probability density. We provide an appropriate likelihood cost function for parameter fitting, and derive an adjoint based approach for the computation of its gradient.

## 1. Introduction

Many processes in applications ranging from science and technology via finance to biology can be modeled by stochastic processes. Our work is particularly motivated by fatigue degradation modeling, cf., e.g., [10,1,3,12]. These models often contain parameters that are not directly accessible to measurements and therefore have to be fitted from additional observations of the system, usually given at discrete time instances $0 < t_1 < \cdots < t_n < T$, within the time interval $[0, T]$, in which the stochastic evolution takes place. This leads to the formulation as a stochastic state space model (SSM):

$$\text{state equation:} \qquad \begin{cases} dX_t &= a^\theta(t,\, X_t)dt + b^\theta(t,\, X_t)dW_t\,, \\ X_0 &\sim u_0^\theta\,, \end{cases} \tag{1a}$$

$$\text{observation equation:} \qquad Y_{t_i} = h_i^\theta(X_{t_i}) + \eta_i, \quad \eta_i \sim \Phi^\theta. \tag{1b}$$

[*] Corresponding author.
*E-mail addresses:* barbara.kaltenbacher@aau.at (B. Kaltenbacher), barbara.pedretscher@k-ai.at (B. Pedretscher).

https://doi.org/10.1016/j.jmaa.2018.05.048

The state equation is a stochastic differential equation (SDE) with drift $a^\theta$, diffusion coefficient $b^\theta$ and Wiener process $W$ [12]. The superscript $\theta$ denotes the model parameters to be identified.

The stochastic process $X$ on the interval $[0,T]$ is a family of random variables

$$X_t : \Omega \to D \subseteq \mathbb{R}^m, \quad t \in [0,T], \tag{2}$$

on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. In the context of degradation modeling, the state vector $X$ may consist of, e.g., crack length or volumetric share of damaged material, local strains, and grain misorientation, to name just a few examples of relevant physical quantities evolving over time. Stochasticity of this evolution is triggered, e.g., by random initial void and grain distribution, and randomness of crack propagation directions.

Drift and diffusion are defined as possibly nonlinear functions on the time – state space cylinder

$$\begin{aligned} a^\theta : &\quad (0,T) \times D \to \mathbb{R}^m, \\ b^\theta : &\quad (0,T) \times D \to \mathbb{R}^{m\times m}, \end{aligned} \tag{3}$$

where the unknown parameters $\theta$ are contained in a subset $Q$ of $\mathbb{R}^d$.

The SSM (1) is a hybrid system in the sense that observations are only available at discrete time instances $t_i$, whereas the evolution is continuous in time. Generally, the observation equation (1b) contains (possibly time dependent) model functions

$$h_i^\theta : D \to \mathbb{R}^k,$$

with random noise to model measurement errors

$$\eta_i : \tilde{\Omega} \to \mathbb{R}^k, \quad t_i \in \{t_1, \ldots, t_n\}, \tag{4}$$

on another probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$. For simplicity of notation and since distinction from $(\Omega, \mathcal{F}, P)$ will always be clear from the context, we will skip the tilde in the following.

Our aim is to identify the true parameter vector $\theta$ in the SSM (1) by means of available indirect measurements. In the context of degradation modeling, the measurements, which can be used for this purpose, are, e.g., resistance and orientation measurements.

Parameter identification will here be performed by a Maximum Likelihood approach, which is based on maximizing the probability density of the observations by considering

- the stochastic differential state equation,
- the initial conditions,
- the measurement noise, and
- the (physical) parameter constraints.

The problem of parameter identification in SDEs has been studied by many authors, see, e.g., [13] and the references therein. Note however that we here deal with the difficulty of only indirect observations (1b), which are given at possibly only few time instances. This setting is on one hand relevant for real applications, on the other hand, it rules out the use of standard parameter estimation approaches.

The history dependence of the evolution, as well as the fact that two different kinds of stochastic processes and random variables, namely the state process $X$ and noise process $\eta$, are involved, considerably complicates the formulation of a stochastically consistent likelihood function for the general SSM (1). To overcome this problem, we transform SDE (1a) to a deterministic model. For this purpose, the SDE's corresponding Kolmogorov forward or Fokker–Planck (FP) equation, which basically describes the evolution of the state

Please cite this article in press as: B. Kaltenbacher, B. Pedretscher, Parameter estimation in SDEs via the Fokker–Planck equation: Likelihood function and adjoint based gradient computation, J. Math. Anal. Appl. (2018), https://doi.org/10.1016/j.jmaa.2018.05.048

density, will be used as state equation instead of the SDE, cf., e.g., [7]. The FP equation will be set up in terms of the transition probability density $u$ of the stochastic process $X$, which is defined as, [15]:

$$u(t,x)\,dx = u(t,x;x_0)\,dx \;=\; \mathbb{P}\left(x < X_t \leq x + dx \,|\, X_0 = x_0\right), \tag{5}$$

where $X_0$ is assumed to be distributed according to the density $u_0^\theta$,

$$u(t,x) = u(t,x\,|\,X_0 \sim u_0^\theta), \qquad t > 0,\; x \in D,$$

cf. (1a). Therewith, the FP equation of (1a) reads as follows:

$$\frac{\partial}{\partial t} u(t,x) = -\nabla \cdot J^\theta(t,x), \quad x \in D, \tag{6}$$

for $t \in (0,T)$ with probability flux

$$J^\theta = a^\theta u - \frac{1}{2}\nabla \cdot \left(b^\theta {b^\theta}^T\right) u, \tag{7}$$

and initial conditions

$$u(0,x) = u_0^\theta(x), \quad x \in D, \tag{8}$$

where $D$ is a domain comprising the state space. Note that (6) is always a linear PDE, even if (1a) is a nonlinear SDE. Equation (6) can be written in divergence form [14]

$$\frac{\partial}{\partial t} u(t,x) = \nabla \cdot \left(B^\theta \nabla u - A^\theta u\right)(t,x), \tag{9}$$

where

$$A^\theta = a^\theta - \frac{1}{2}\nabla \cdot \left(b^\theta {b^\theta}^T\right), \tag{10}$$

$$B^\theta = \frac{1}{2}\left(b^\theta {b^\theta}^T\right), \tag{11}$$

with the matrix divergence $\nabla \cdot C = \left(\sum_{j=1}^m \frac{\partial}{\partial x_j} C_{ij}\right)^T_{i=1,\,\ldots,\,m}$. Thus, if $B^\theta$ is uniformly positive definite on $(0,T) \times D$, then (9), i.e., (6) is parabolic.

As, for fixed time $t$, $u(t,\cdot)$ represents a probability density, it has to fulfill mass conservation and positivity:

$$\int_D u(t,x)\,dx = 1, \qquad t \in [0,T], \tag{12}$$

$$u(t,x) \geq 0, \qquad (t,x) \in [0,T] \times D. \tag{13}$$

To guarantee mass conservation, we impose the no-flux boundary condition

$$J^\theta(t,x) \cdot n_D(x) \;=\; 0, \qquad x \in \partial D \tag{14}$$

for $t \in (0,T)$. Indeed, using the Divergence Theorem, it is easy to see that this implies (12), provided

$$\int_D u_0^\theta(x)\,dx = 1$$

Please cite this article in press as: B. Kaltenbacher, B. Pedretscher, Parameter estimation in SDEs via the Fokker–Planck equation: Likelihood function and adjoint based gradient computation, J. Math. Anal. Appl. (2018), https://doi.org/10.1016/j.jmaa.2018.05.048

holds: Upon exchangeability of time differentiation and spatial integration, we have

$$\frac{d}{dt}\int\limits_D u(t,x)\,dx = \int\limits_D \frac{\partial}{\partial t}u(t,x)\,dx = -\int\limits_D \nabla\cdot J^\theta(t,x)\,dx = -\int\limits_{\partial D} J^\theta\cdot n_D\,ds = 0$$

by (14). Altogether we end up with the weak formulation

$$\int\limits_D \left(\frac{\partial u}{\partial t}(t,x)v(x) - J^\theta(t,x)\cdot\nabla v(x)\right)\,dx = 0 \quad \forall v\in H_0^1(D)\,. \tag{15}$$

Well-posedness of the initial value problem (15), (8) on a finite time interval $(0,T)$ follows from standard analysis of linear parabolic PDEs, cf. [8, Theorems 2, 3, 4, Section 7.1.2], provided all coefficients are in $L^\infty((0,T)\times D)$, $u_0^\theta\in L^2(D)$, and $B^\theta$ is uniformly positive definite on $(0,T)\times D$. To guarantee global in time well-posedness and convergence to a stationary solution of the FP equation as $t\to\infty$, a condition like

$$A^\theta\cdot n_D < 0 \qquad \text{on } \partial D \tag{16}$$

is needed, see [6] for the case of scalar valued diffusion $b^\theta$. In view of (10), condition (16) means that at the isolating (cf. (14)) boundary, the diffusion has to dominate the drift to prevent emergence of singularities. Moreover, mass conservation (12) and positivity (13) are made rigorous in [6] in this scalar diffusion case. In the general case of anisotropic diffusion, these questions, and in particular also large time behavior have been studied in [2] for $D=\mathbb{R}^n$, i.e., without boundary conditions. We expect that a combination of the techniques from [2,6] allows to prove mass conservation, positivity, global in time well-posedness, and convergence to a stationary state also in the anisotropic diffusion setting on bounded domains, as relevant here.

**Remark 1.** Concerning initial data and initial observations, we always assume to have a possibly parameter dependent Ansatz $u_0^\theta$ for the initial data; in case $u_0$ is known, parameter dependence may be skipped in the notation $u_0^\theta = u_0$; also the case of $u_0^\theta$ being an arbitrary function can be included by regarding the initial data itself as (infinite dimensional) part of the parameter $\theta$.

Initial observations $h_0^\theta(X_{t_0})+\eta_0$ at $t_0 := 0$ might or might not be available. For simplicity of exposition we only consider the case without initial observations. The case with observations at time $t=0$ can be covered by considering the limit $t_1\to 0$.

For later use, we finally state the deterministically transformed SSM on the $i$th sub-time interval

$$\Sigma_i(\theta,u):\qquad \begin{cases} \int\limits_D \left(\frac{\partial u}{\partial t}(t,x)v(x) - J^\theta(t,x)\cdot\nabla v(x)\right)\,dx = 0 \quad \forall v\in H_0^1(D)\,, \\ Y_i = h_i^\theta(X_{t_i})+\eta_i, \quad i\in\{1,\dots,n\}\,, \end{cases} \tag{17}$$

where $J^\theta$ is defined as in (7).

Parameter identification in the transformed SSM (17) by means of Maximum Likelihood estimation requires to solve the following optimization problem:

$$\begin{array}{ll} \max\limits_{\theta,\,u} & \Psi(\theta,u;y) \\ \text{s.t.} & \Sigma(\theta,u) \end{array} \tag{18}$$

where $\Sigma$ represents the SSM on the entire time interval $[0,T]$ and $\Psi$ is defined by the likelihood function $\psi$, i.e., the probability density function of the data $y$ for fixed parameters $\theta$. To this end, in Section 2, we will derive a stochastically consistent formulation of the likelihood function $\psi$. In order to apply gradient based optimization methods to (18), we will derive an adjoint based approach that takes into account the special structure of the cost function in Section 3, which is the core of this paper. Finally, in Section 4, we will draw some conclusions and provide an outlook on future research in this context.

## 2. The likelihood function

To define the likelihood function $\psi$ for the optimization problem (18), recall that

$$\mathbb{P}\left(h^\theta(X)+\eta\in B\right)=\int\limits_B \psi(y)\,dy, \tag{19}$$

where $B$ is an arbitrary element of the $k\cdot n$ dimensional Borel $\sigma$ Algebra $\mathcal{B}^{k\cdot n}$, in view of the fact that observations are available at $\{t_1,\dots,t_n\}$.[1] The value $\psi(y)$ represents the likelihood of the data $y=(y_1,\dots,y_n)$. To obtain an explicit expression for $\psi$, we impose the following conditions.

**Assumption 1.**

(i) The stochastic process $X$ is a Markov process, i.e.

$$\mathbb{P}\left(X_{t+\Delta t}=x\,|\,X(s),0\le s\le t\right)=\mathbb{P}\left(X_{t+\Delta t}=x\,|\,X(t)\right), \quad \forall t,\,\Delta t>0,$$

with a probability density $u$ as in (5).

(ii) For all $i\in\{1,\dots,n\}$, the random variables $X_{t_i}$ and $\eta_i$, and therewith $h_i^\theta(X_{t_i})$ and $\eta_i$ are stochastically independent.

(iii) The components of the measurement noise $\underline{\eta}=(\eta_1,\,\dots,\,\eta_n)$ corresponding to different time instances are mutually independent and we assume a joint density to exist, which then has to be of the form

$$\phi^\theta(\eta_1,\,\dots,\,\eta_n)=\prod_{i=1}^n\phi_i^\theta(\eta_i).$$

**Proposition 2.** *Under Assumption 1, the likelihood function has the two representations*

$$\psi(y)=\int\limits_{D^n}u_1(x_1)\prod_{j=2}^n u(x_j\,|\,X_{j-1}=x_{j-1})\prod_{i=1}^n\phi_i^\theta\left(y_i-h_i^\theta(x_i)\right)\,d(x_1,\dots,x_n), \tag{20}$$

*where* $u_1(x_1)=\tilde{u}_1(t_1,x_1)$ *(cf. (23)) and for all* $j\in\{2,\,\dots,\,n\}$, $u(x_j\,|\,X_{j-1}=x_{j-1})=\hat{u}_j(t_j,x_j;x_{j-1})$, *where*

$$\begin{aligned}&\hat{u}_j(\cdot,\cdot;x_{j-1})\ \textit{solves the FP equation (15) on}\ (t_{j-1},\,t_j),\ \textit{with initial conditions}\\ &\hat{u}_j(t_{j-1},x;x_{j-1})=\delta_{x_{j-1}}(x)\end{aligned} \tag{21}$$

*and*

$$\psi(y)=\int\limits_D\rho_n^\theta\,(x_n)\,\tilde{u}_n\,(t_n,\,x_n)\,dx_n\,, \tag{22}$$

[1] Extension to the more general case of having $k_i$ dimensional observations at time instance $t_i$ is straightforward.

Please cite this article in press as: B. Kaltenbacher, B. Pedretscher, Parameter estimation in SDEs via the Fokker–Planck equation: Likelihood function and adjoint based gradient computation, J. Math. Anal. Appl. (2018), https://doi.org/10.1016/j.jmaa.2018.05.048

*where for all* $j \in \{1, \dots, n\}$,

$$\begin{gathered}\tilde{u}_j \textit{ solves the FP equation } (15) \textit{ on } (t_{j-1}, t_j), \textit{ with initial conditions}\\ \tilde{u}_j(t_{j-1}, x) = \begin{cases} u_0^\theta(x), & \textit{if } j = 1, \\ \lim\limits_{t \to t_{j-1}^-} \tilde{u}_{j-1}(t, x)\, \rho_{j-1}^\theta(x), & \textit{if } j \in \{2, \dots, n\}, \end{cases}\end{gathered} \tag{23}$$

*and*

$$\rho_j^\theta(x_j) = \phi_j^\theta(y_j - h_j^\theta(x_j)) \,. \tag{24}$$

**Proof.** The joint probability density of the stochastic process $X$ evaluated at the discrete time instances $\{t_1, \dots, t_n\}$, i.e., of $\underline{X} = (X_1, \dots, X_n) = (X_{t_1}, \dots, X_{t_n})$, will be denoted by $\underline{u}$, i.e.

$$\mathbb{P}_X(A) = \int\limits_A \underline{u}(x)\, d\zeta(x) \qquad \forall A \in \mathcal{B}^{m \cdot n}.$$

To formulate the likelihood function, recall that $\psi$ is the probability density of $\underline{Y} = h^\theta(\underline{X}) + \underline{\eta}$, evaluated at the time points $\{t_1, \dots, t_n\}$. Here, the observation operator is of the form

$$h^\theta : D^n \subseteq \mathbb{R}^{m \cdot n} \to \mathbb{R}^{k \cdot n}, \tag{25}$$

and $f_{h^\theta(\underline{X})}$ denotes the density of the random variable $h^\theta(\underline{X})$.

Then, the probability of an arbitrary Borel set $B$ is given by

$$\begin{aligned} \mathbb{P}\left(h^\theta(\underline{X}) + \underline{\eta} \in B\right) &= \int\limits_B \psi(y)\, dy \\ &= \int\limits_B \int\limits_{\mathbb{R}^{k \cdot n}} f_{h^\theta(\underline{X})}(y - s)\, \phi^\theta(s)\, ds\, dy \\ &= \int\limits_{\mathbb{R}^{k \cdot n}} \int\limits_B f_{h^\theta(\underline{X})}(y - s)\, \phi^\theta(s)\, dy\, ds \\ &= \int\limits_{\mathbb{R}^{k \cdot n}} \int\limits_{B - \{s\}} f_{h^\theta(\underline{X})}(z)\, dz\, \phi^\theta(s)\, ds \\ &= \int\limits_{\mathbb{R}^{k \cdot n}} \int\limits_{{h^\theta}^{-1}(B - \{s\})} \underline{u}(x)\, dx\, \phi^\theta(s)\, ds \\ &= \int\limits_{\{(x,s) \in D^n \times \mathbb{R}^{k \cdot n} :\, h^\theta(x) + s \in B\}} \underline{u}(x)\, \phi^\theta(s)\, d(x, s)\,, \end{aligned} \tag{26}$$

where we have used the Convolution Theorem cf., e.g., [11, Appendix B4] and Assumption 1 (ii) in the second, Fubini's Theorem in the third, the substitution $z := y - s$ in the fourth, and the fact that

$$\int\limits_{B - \{s\}} f_{h^\theta(\underline{X})}(z)\, dz = \mathbb{P}\left(h^\theta(\underline{X}) \in B - \{s\}\right) = \mathbb{P}\left(\underline{X} \in {h^\theta}^{-1}(B - \{s\})\right)$$

in the fifth equality.

We now change variables in (26) by introducing the mapping

$$\varphi : (x, s) \mapsto (x, h^\theta(x) + s) := (x, y). \tag{27}$$

The functional determinant of the mapping $\varphi$ is given as

$$\det D\varphi = \det \begin{pmatrix} \frac{\partial\varphi_1}{\partial x} & \frac{\partial\varphi_1}{\partial s} \\ \frac{\partial\varphi_2}{\partial x} & \frac{\partial\varphi_2}{\partial s} \end{pmatrix} = \det \begin{pmatrix} I \in \mathbb{R}^{m\times m} & 0 \in \mathbb{R}^{m\times k} \\ Dh^\theta(x) \in \mathbb{R}^{k\times m} & I \in \mathbb{R}^{k\times k} \end{pmatrix} = 1, \tag{28}$$

and the image of the set over which we integrate in (26) is

$$\varphi\left(\left\{(x,\, s) \in D^n \times \mathbb{R}^{k\cdot n} : h^\theta(x) \,+\, s \,\in\, B\right\}\right) = \left\{(x, h^\theta(x) + s) : x \in D^n,\, s \in \mathbb{R}^{k\cdot n},\, h^\theta(x) + s \in B\right\} = D^n \times B.$$

Therewith,

$$\mathbb{P}\left(h^\theta(X) + \eta \in B\right) = \int\limits_B \int\limits_{D^n} \underline{u}(x)\, \phi^\theta(y - h^\theta(x))\, d(x, y)\,,$$

enables to rewrite the likelihood function as

$$\psi(y) = \int\limits_{D^n} \underline{u}(x)\, \phi^\theta(y - h^\theta(x))\, dx. \tag{29}$$

The Markov property (cf. Assumption 1 (i)) allows to split the joint probability density $\underline{u}$ in (29) as follows:

$$\underline{u}(x_1, \dots, x_n) = u_1(x_1) \prod_{j=2}^{n} u(x_j \,|\, X_{j-1} = x_{j-1}), \tag{30}$$

where for each $j \in \{2, \dots, n\}$, $u(x_j \,|\, X_{j-1} = x_{j-1}) = \hat{u}_j(t_j, x_j; x_{j-1})$ denotes the transition density obtained by solving the FP equation on the subinterval $(t_{j-1}, t_j)$ with Dirac delta initial condition $\delta_{x_{j-1}}$, cf. (21), and

$$u_1(x_1) = \int\limits_D \underline{u}(x_0, x_1)\, dx_0 = \int\limits_D u_0^\theta(x_0)\, u(x_1 \,|\, X_0 = x_0)\, dx_0\,,$$

i.e., $u_1(x_1)$ is given by the solution at time $t_1$ of the FP equation on $(0, t_1)$ with initial condition $u(0, x) = u_0^\theta(x)$, which follows from the linearity of the FP equation and Lemma 5 in the Appendix. This proves (20).

To show the second representation (22), we set

$$u^{(j)}(x_j, x_{j-1}) := \hat{u}_j(t_j, x_j; x_{j-1}),$$

with $\hat{u}(\cdot, \cdot; x_{j-1})$ as in (21). Assume first of all the case of observations for just two time points $\{t_1, t_2\}$, then the likelihood function is of the form:

$$\begin{aligned} \psi(y) &= \int\limits_D \left[ \int\limits_D \phi_1^\theta(y_1 - h_1(x_1)) u_1(x_1) u^{(2)}(x_2, x_1)\, dx_1 \right] \phi_2^\theta(y_2 - h_2(x_2))\, dx_2 \\ &= \int\limits_D \left[ \int\limits_D \rho_1^\theta(x_1) u_1(x_1) u^{(2)}(x_2, x_1)\, dx_1 \right] \rho_2^\theta(x_2)\, dx_2. \end{aligned}$$

Please cite this article in press as: B. Kaltenbacher, B. Pedretscher, Parameter estimation in SDEs via the Fokker–Planck equation: Likelihood function and adjoint based gradient computation, J. Math. Anal. Appl. (2018), https://doi.org/10.1016/j.jmaa.2018.05.048

By Lemma 5, this can be reformulated as:

$$\psi(y) = \int\limits_D \tilde{u}_2(t_2, x_2)\rho_2^\theta(x_2)\, dx_2,$$

where $\tilde{u}_2$ is the solution of the FP equation (15) on the sub-time interval $(t_1, t_2)$ with initial condition $\tilde{u}_2(t_1, x) = \lim_{t\to t_1^-} \tilde{u}_1(t, x)\rho_1(x)$.

Inductive generalization to $n$ time instances $\{t_1, \dots, t_n\}$ gives the reformulation (22). Namely, with the notation

$$\psi^{(\ell)}(y_1, \dots, y_\ell; x_\ell) = \int\limits_{D^{m-1}} u_1(x_1) \prod_{j=2}^{\ell} u(x_j \,|\, X_{j-1} = x_{j-1}) \prod_{i=1}^{\ell} \phi_i^\theta \left(y_i - h_i^\theta(x_i)\right)\, d(x_1, \dots, x_{\ell-1})$$

for $\ell \in \{2, \dots, n\}$, we have the recursion

$$\psi^{(\ell+1)}(y_1, \dots, y_{\ell+1}, x_{\ell+1}) = \int\limits_D u(x_{\ell+1} \,|\, X_\ell = x_\ell)\, \phi_{\ell+1}^\theta \left(y_{\ell+1} - h_{\ell+1}^\theta(x_{\ell+1})\right) \psi^{(\ell)}(y_1, \dots, y_\ell; x_\ell)\, dx_\ell\,,$$

which according to the induction hypothesis

$$\psi^{(\ell)}(y_1, \dots, y_\ell; x_\ell) = \tilde{u}_\ell(t_\ell, x_\ell)\, \rho_\ell^\theta(x_\ell)\,,$$

and with the notation (24) yields

$$\psi^{(\ell+1)}(y_1, \dots, y_{\ell+1}; x_{\ell+1}) = \left[\int\limits_D \tilde{u}_\ell(t_\ell, x_\ell)\, \rho_\ell^\theta(x_\ell)\, u^{(\ell+1)}(x_{\ell+1}, x_\ell)\, dx_\ell\right] \rho_{\ell+1}^\theta(x_{\ell+1})\,,$$

where by Lemma 5, the term in brackets is the value at time $t_{\ell+1}$ of the solution $\tilde{u}_{\ell+1}$ to the FP equation (15) on $(t_\ell, t_{\ell+1})$ with initial data $\lim_{t\to t_\ell^-} \tilde{u}_\ell(t, x_\ell)\rho_\ell^\theta(x_\ell)$ at time $t_\ell$. □

## 3. Gradient computation by adjoint approach

As already outlined above, our aim is to reconstruct the parameters $\theta$ of the degradation model by maximizing the likelihood of the observation data. Thus, the optimization problem is of the form

$$\max_{\theta,\, \tilde{u}} \Psi(\theta, \tilde{u}; y)$$
$$\text{s.t. } \tilde{\Sigma}(\theta, \tilde{u}; y)$$

with $\Psi$ defined by the likelihood function derived in the previous section, more precisely, the representation (22), i.e.,

$$\Psi(\theta, \tilde{u}, y) = \int\limits_D \phi_n^\theta(y_n - h_n^\theta(x_n))\,(x_n)\; \tilde{u}_n\,(t_n,\, x_n)\; dx_n\,,$$

and $\tilde{\Sigma}$ by (23), (24). By means of the *parameter-to-state* map

$$\theta \mapsto \tilde{\mathbf{u}}(\theta), \tag{31}$$

Please cite this article in press as: B. Kaltenbacher, B. Pedretscher, Parameter estimation in SDEs via the Fokker–Planck equation: Likelihood function and adjoint based gradient computation, J. Math. Anal. Appl. (2018), https://doi.org/10.1016/j.jmaa.2018.05.048

which maps the parameters to the solution $\tilde{u} = (\tilde{u}_1, \dots, \tilde{u}_n)$ of (23) with $\rho_j$ as in (24), the above optimization problem can be formulated as an unconstrained problem

$$\max_\theta \Psi(\theta, \tilde{\mathbf{u}}(\theta); y) = \max_\theta j(\theta)\,. \tag{32}$$

Efficient parameter identification in (32) requires the gradient $\nabla j$, i.e., the derivatives of the likelihood with respect to all the parameters $\theta_1, \dots, \theta_d$. For this purpose, an adjoint approach, based on the Lagrange function, is applied. Basically, the adjoint approach is used to avoid the expensive computation of the state sensitivities from $d$ linearized versions of the state equation over the whole time interval. Instead, we solve the adjoint equation, a linear PDE backwards in time, which will be defined piecewise in time, i.e., on the subintervals $(t_{j-1}, t_j)$ to take into account the discrete observation time instances. This will allow to compute the full gradient by means of just one additional linear PDE solution, instead of $d$.

To define the Lagrange function, we split the time interval according to the observation time instances $t_j$ and introduce Lagrange multipliers (adjoint states) $p_j$ on the subintervals $(t_{j-1}, t_j]$. Therewith, the Lagrange function reads as follows:

$$\begin{aligned}
&\mathcal{L}(\theta, \tilde{u}_1, \dots, \tilde{u}_n, p_1, \dots, p_n)\\
&\quad= \int_D \rho_n^\theta\,(x_n)\;\tilde{u}_n\,(t_n,\,x_n)\;dx_n\\
&\qquad+ \sum_{j=1}^n \int_{t_{j-1}}^{t_j} \int_D \left[\frac{\partial \tilde{u}_j}{\partial t}(t,x) p_j(t,x) + \left(\frac{1}{2}\,\nabla \cdot \left(\left(b^\theta b^{\theta^T}\right) \tilde{u}_j\right) - a^\theta \tilde{u}_j\right)(t,x) \cdot \nabla p_j(t,x)\right] d(x,t)\\
&\qquad+ \int_D \left(\tilde{u}_1(0,x) - u_0^\theta(x)\right) \lim_{t \to 0^+} p_1(t,x)\, dx\\
&\qquad+ \sum_{j=2}^n \int_D \left(\tilde{u}_j(t_{j-1},x) - \lim_{t \to t_{j-1}^-} \tilde{u}_{j-1}(t,x) \rho_{j-1}^\theta(x)\right) \lim_{t \to t_{j-1}^+} p_j(t,x)\, dx
\end{aligned}$$

with $t_0 := 0$. Integration by parts with respect to space and time gives:

$$\begin{aligned}
&\mathcal{L}(\theta, \tilde{u}_1, \dots, \tilde{u}_n, p_1, \dots, p_n)\\
&\quad= \int_D \rho_n^\theta\,(x_n)\;\tilde{u}_n\,(t_n,\,x_n)\;dx_n\\
&\qquad+ \sum_{j=1}^n \int_{t_{j-1}}^{t_j} \left\{ \int_D \tilde{u}_j(t,x) \left[-\frac{\partial p_j}{\partial t} - a^{\theta^T} \nabla p_j - \frac{1}{2}\left(b^\theta b^{\theta^T}\right) : \nabla^2 p_j\right](t,x)\, dx \right.\\
&\qquad\qquad\qquad \left. + \int_{\partial D} \tilde{u}_j(t,x) \left(\frac{1}{2} b^\theta b^{\theta^T} \nabla p_j\right)(t,x) \cdot n_D\, dS \right\} dt\\
&\qquad+ \sum_{j=2}^n \int_D \left(\tilde{u}_j(t_j,x)\, p_j(t_j,x) - \lim_{t \to t_{j-1}^-} \tilde{u}_{j-1}(t,x)\, \rho_{j-1}^\theta(x)\, p_j(t_{j-1},x)\right) dx\\
&\qquad+ \int_D \left(\tilde{u}_1\,(t_1,x)\; p_1(t_1,x) - u_0^\theta(x)\, p_1(0,x)\right)\, dx
\end{aligned}$$

Please cite this article in press as: B. Kaltenbacher, B. Pedretscher, Parameter estimation in SDEs via the Fokker–Planck equation: Likelihood function and adjoint based gradient computation, J. Math. Anal. Appl. (2018), https://doi.org/10.1016/j.jmaa.2018.05.048

$$
\begin{aligned}
&= \int_D \rho_n^\theta\left(x_n\right) \tilde{u}_n\left(t_n, x_n\right) dx_n \\
&\quad + \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \left\{ \int_D \tilde{u}_j(t,x) \left[ -\frac{\partial p_j}{\partial t} - a^{\theta T} \nabla p_j - \frac{1}{2}\left(b^\theta b^{\theta T}\right) : \nabla^2 p_j \right] (t,x)\, dx \right. \\
&\qquad\qquad \left. + \int_{\partial D} \tilde{u}_j(t,x) \left( \frac{1}{2} b^\theta b^{\theta T} \nabla p_j \right)(t,x) \cdot n_D \, dS \right\} dt \\
&\quad + \sum_{j=1}^{n-1} \int_D \tilde{u}_j(t_j,x) \left[ p_j(t_j,x) - \rho_j^\theta(x)\, p_{j+1}(t_j,x) \right] dx \\
&\quad - \int_D u_0^\theta(x) p_1(0,x)\, dx \; + \int_D \tilde{u}_n\left(t_n,x\right) p_n\left(t_n,x\right) dx.
\end{aligned}
$$

Thus, the requirement

$$
\mathcal{L}'_{\tilde{u}}(\theta, \tilde{u}_1, \dots, \tilde{u}_n, p_1, \dots, p_n) = 0
$$

results in the adjoint equation

$$
\begin{aligned}
-\frac{\partial p_j}{\partial t} - \frac{1}{2}\left(b^\theta b^{\theta T}\right) : \nabla^2 p_j - a^{\theta T} \nabla p_j &= 0, \quad \text{in } (t_{j-1}, t_j) \times D, \\
\left( \frac{1}{2} b^\theta b^{\theta T} \nabla p_j \right) \cdot n_D &= 0, \quad \text{on } (t_{j-1}, t_j) \times \partial D,
\end{aligned} \tag{33}
$$

with the final conditions on each subinterval

$$
p_j\left(t_j, x\right) = \begin{cases} \lim\limits_{t \to t_j^+} p_{j+1}(t,x)\, \rho_j^\theta(x), & j = 1, \dots, n-1, \\ -\rho_n^\theta(x), & j = n. \end{cases} \tag{34}
$$

By means of the adjoint states $\mathbf{p} = (p_1, \dots, p_n) = \mathbf{p}(\theta)$, the gradient of the likelihood can be computed without computing state sensitivities. To this end, note that by definition of the parameter-to-state map (namely such that the state equation constraint is satisfied) and by the chain rule, the gradient of the likelihood is given as follows:

$$
\nabla_\theta j(\theta) = \nabla_\theta \Psi(\theta, \tilde{\mathbf{u}}(\theta); y) = \nabla_\theta \mathcal{L}\left(\theta, \tilde{\mathbf{u}}(\theta), \mathbf{p}(\theta)\right) \tag{35}
$$

$$
= \mathcal{L}'_\theta\left(\theta, \tilde{\mathbf{u}}(\theta), \mathbf{p}(\theta)\right) + \mathcal{L}'_{\tilde{u}}\left(\theta, \tilde{\mathbf{u}}(\theta), \mathbf{p}(\theta)\right) \tilde{\mathbf{u}}'_\theta(\theta) + \mathcal{L}'_p\left(\theta, \tilde{\mathbf{u}}(\theta), \mathbf{p}(\theta)\right) \mathbf{p}'(\theta). \tag{36}
$$

Since $\mathbf{p}(\theta)$ solves the adjoint problem (33)–(34), the second summand in equation (36) vanishes. So does the last summand in (36), due to the fact that $\tilde{\mathbf{u}}(\theta)$ satisfies the state equation. Thus, it suffices to compute the direct derivatives of $\mathcal{L}$ with respect to $\theta$. More precisely, we end up with

$$
\nabla_\theta j(\theta) = \mathcal{L}'_\theta\left(\theta, \tilde{\mathbf{u}}(\theta), \mathbf{p}(\theta)\right), \tag{37}
$$

where $\tilde{u}_1, \dots, \tilde{u}_n$ solve (23) and $p_1, \dots, p_n$ solve the adjoint system (33)–(34). Therewith, under the smoothness conditions given in Assumption 3, the gradient can be given explicitly by means of the adjoint state.

**Assumption 3.** For $\theta \in Q$ and $i \in \{1, \dots, d\}$, $(\vartheta, y) \mapsto \phi_1^\vartheta(y), \dots, \phi_n^\vartheta(y)$ is differentiable with respect to $\vartheta_i$ and with respect to $y$ in $\{\theta\} \times \mathbb{R}^k$. Moreover, all $f^\vartheta \in \{a^\vartheta, b^\vartheta, u_0^\vartheta, h_1^\vartheta, \dots, h_n^\vartheta\}$ satisfy the following conditions:

(i) For all $\vartheta$ in a neighborhood of $\theta$, the function $z \mapsto f^\vartheta(z)$ is integrable.

(ii) There exists an integrable function $g : D \to \mathbb{R}_0^+$ (or $g : (0, T) \times D \to \mathbb{R}_0^+$) such that for all $\vartheta_1, \vartheta_2$ in a neighborhood of $\theta$ and all $z \in D$ (or $z \in (0, T) \times D$)

$$|f^{\vartheta_1}(z) - f^{\vartheta_2}(z)| \leq g(z)|\vartheta_1 - \vartheta_2|\,.$$

(iii) For almost all $z \in D$ (or $z \in (0, T) \times D$), the function $\vartheta \mapsto f^\vartheta(z)$ is differentiable with respect to $\vartheta_i$ in $\theta$.

**Proposition 4.** *Under the assumptions of Proposition 2, Assumption 3, and if* $a^\theta$, $\frac{\partial a^\theta}{\partial \theta_i}$, $b^\theta$, $\nabla b^\theta$, $\frac{\partial b^\theta}{\partial \theta_i}$, $\nabla \frac{\partial b^\theta}{\partial \theta_i} \in L^\infty((0, T) \times D)$, $u_0^\theta$, $\frac{\partial u_0^\theta}{\partial \theta_i} \in L^2(D)$, $\frac{\partial \phi_j^\theta}{\partial \theta_i} \in L^\infty(\mathbb{R}^k)$, $\frac{\partial \phi_j^\theta}{\partial y} \in L^\infty(\mathbb{R}^k)$, $\frac{\partial h_j^\theta}{\partial \theta_i} \in L^\infty(D)$, *and* $B^\theta$ *as defined in* (11) *is uniformly positive definite on* $(0, T) \times D$*, we have*

$$\begin{aligned}
\frac{\partial}{\partial \theta_i} j(\theta) &= \int\limits_D \frac{\partial \rho_n^\theta}{\partial \theta_i}(x) \tilde{u}_n(t_n, x)\, dx \\
&\quad + \sum_{j=1}^{n} \int\limits_{t_{j-1}}^{t_j} \int\limits_D \left[ \nabla \left( \frac{1}{2} \left( \frac{\partial b^\theta}{\partial \theta_i} b^{\theta T} + b^\theta \frac{\partial b^{\theta T}}{\partial \theta_i} \right) \tilde{u}_j \right) - \frac{\partial a^{\theta T}}{\partial \theta_i} \tilde{u}_j \right] (t, x) \cdot \nabla p_j(t, x)\, d(x, t) \\
&\quad - \int\limits_D \frac{\partial u_0^\theta}{\partial \theta_i}(x)\, p_1(0, x)\, dx - \sum_{j=1}^{n-1} \int\limits_D \lim_{t \to t_j^-} \tilde{u}_j(t, x) \frac{\partial \rho_j^\theta}{\partial \theta_i}(x) p_{j+1}(t_j, x)\, dx,
\end{aligned}$$

*where for all* $j \in \{1, \dots, n\}$,

$$\frac{\partial \rho_j^\theta}{\partial \theta_i}(x) = \frac{\partial \phi_j^\theta}{\partial \theta_i}(y_j - h_j^\theta(x)) - \frac{\partial \phi_j^\theta}{\partial y}(y_j - h_j^\theta(x)) \frac{\partial h_j^\theta}{\partial \theta_i}(x)\,.$$

*Here* $\tilde{u}_j$ *and* $p_j$ *solve* (23)*, and* (33)–(34)*, respectively.*

**Proof.** After the derivations above, it only remains to verify differentiability of $j$ as well as exchangeability of integration and differentiation, as can be done according to, e.g., [5, Proposition 5.108] under Assumption 3. Moreover, the given conditions on the coefficients in the Fokker–Planck equation and the initial conditions guarantee well-definedness, uniqueness and $L^2(0, T; H^1(D)) \cap H^1(0, T; (H^1(D))^*) \subseteq C(0, T; L^2(D))$ regularity of solutions to (23), and (33)–(34). (The latter, by making use of Sobolev embedding results of $H^1(D)$ in $L^p(D)$ for appropriate dimension dependent $p > 2$ for $\tilde{u}_j$, $p_j$, actually enables a slight relaxation of assumptions on the summability index of derivatives of the coefficients $a^\theta$, $b^\theta$.) □

## 4. Conclusions and remarks

In this paper we considered the problem of identifying parameters in stochastic differential equations. The main challenges in this setting lie in the fact that observations $h^\theta(X)$ are only indirect and we are in a low frequency regime [9] in the sense that the observation times are too far away from each other to admit applicability of conventional drift and diffusion estimators. On the other hand, they are not far enough away from each other to justify a mutual independence assumption, which would enable to work with a conventional likelihood function. Therefore, in this paper we first of all derived an expression for the

Please cite this article in press as: B. Kaltenbacher, B. Pedretscher, Parameter estimation in SDEs via the Fokker–Planck equation: Likelihood function and adjoint based gradient computation, J. Math. Anal. Appl. (2018), https://doi.org/10.1016/j.jmaa.2018.05.048

correct likelihood. In addition, an adjoint approach for computing gradients of this likelihood, as required for sensitivity based optimization in parameter fitting, was presented. This approach could as well be transferred to the setting of an infinite dimensional parameter space, including even the task of nonparametric drift and diffusion estimation by defining $\theta = (a, b)$, with $a$, $b$ denoting drift and diffusion, respectively.

Future research will be concerned with a numerical implementation of this approach and with investigating the possibility of applying likelihood profiles [4] for quantifying the uncertainty in the estimated parameters.

## Acknowledgments

The authors gratefully acknowledge support by the Karl Popper Kolleg "Modeling-Simulation-Optimization", funded by the Alpen-Adria-Universität Klagenfurt and by the Carinthian Economic Promotion Fund (KWF), as well as by the Austrian Research Promotion Agency (FFG, Project No. 860424). Further, both authors wish to thank Olivia Bluder (KAI GmbH, Villach, Austria), Michael Nelhiebel (KAI GmbH, Villach, Austria), Max-Hermann Poech (Fraunhofer-Institut für Siliziumtechnologie (ISIT), Itzehoe, Germany), Thomas Svensson (Ingenjörsstatistik, Borås, Sweden), and Johannes Zechner (KAI GmbH, Villach, Austria) for fruitful discussions.

Moreover, the authors wish to thank the reviewer for valuable remarks that led to an improved version of the manuscript.

## Appendix A

To reformulate the likelihood function the following convolution property is required.

**Lemma 5.** *For arbitrary $x_0 \in D$, suppose $\hat{v}^{x_0}$ solves the homogeneous linear PDE with Dirac delta initial condition:*

$$\begin{cases} L\hat{v}^{x_0}(t,x) = 0, & (t,x) \in (0,T) \times D, \\ \hat{v}^{x_0}(0,x) = \delta_{x_0}(x), & x \in D. \end{cases}$$

*Then, for any $w \in C(D)$, the convolution*

$$v(t,x) := \int_D \hat{v}^{x_0}(t,x) w(x_0)\, dx_0$$

*solves the initial value problem*

$$\begin{cases} Lv(t,x) = \int_D L\hat{v}^{x_0}(t,x) w(x_0)\, dx_0 = 0, \\ v(0,x) = \int_D \delta_{x_0}(x) w(x_0)\, dx_0 = \int_D \delta_x(x_0) w(x_0)\, dx_0 = w(x). \end{cases}$$

**Proof.** Due to definition of $v$ combined with the convolution property of the Dirac delta function. □

## References

[1] T.L. Anderson, Fracture Mechanics: Fundamentals and Applications, 3rd edition, CRC Press, 2005.

[2] A. Arnold, E. Carlen, Q. Ju, Large-time behavior of non-symmetric Fokker–Planck type equations, Commun. Stoch. Anal. (2008) 153–175.

[3] M. Avlonitis, M. Zaiser, E.C. Aifantis, A stochastic approach to microstructural evolution during plastic deformation and its application to texture, in: P.S. Theocaris, D.I. Fotiadis, C.V. Massalas (Eds.), Proceedings of the 5th National Congress on Mechanics, University of Ioannina Press, 1998, pp. 907–914.

[4] R. Boiger, J. Hasenauer, S. Hroß, B. Kaltenbacher, Integration based profile likelihood calculation for PDE constrained parameter estimation problems, Inverse Probl. 32 (2016) 125009, https://doi.org/10.1088/0266-5611/32/12/125009, arXiv: 1604.02894v1 [math.NA].

[5] J. Bonnans, A. Shapiro, Perturbation Analysis of Optimization Problems, Springer Series in Operations Research and Financial Engineering, Springer, New York, 2000.

[6] J.A. Carrillo, S. Cordier, S. Mancini, A decision-making Fokker–Planck model in computational neuroscience, J. Math. Biol. 63 (5) (2011) 801–830, https://doi.org/10.1007/s00285-010-0391-3.

[7] F. Dunker, T. Hohage, On parameter identification in stochastic differential equations by penalized maximum likelihood, Inverse Probl. 30 (9) (2014) 095001, https://doi.org/10.1088/0266-5611/30/9/095001, arXiv:1404.0651 [stat.CO].

[8] L.C. Evans, Partial Differential Equations, Graduate Studies in Mathematics, vol. 19, American Mathematical Society, Providence, RI, 1998.

[9] E. Gobet, M. Hoffmann, M. Reiß, Nonparametric estimation of scalar diffusions based on low frequency data, Ann. Statist. 32 (5) (2004) 2223–2253, https://doi.org/10.1214/009053604000000797.

[10] G. Gottstein, Physikalische Grundlagen der Materialkunde, 2nd edition, Springer-Lehrbuch, Springer, Berlin, Heidelberg, 2001.

[11] F.B. Hanson, Applied Stochastic Processes and Control for Jump Diffusions: Modeling, Analysis, and Computation, Advances in Design and Control, Society for Industrial and Applied Mathematics, 2007.

[12] P.E. Kloeden, E. Platen, Numerical Solution of Stochastic Differential Equations, Applications of Mathematics – Stochastic Modelling and Applied Probability, vol. 23, Springer, Berlin, Heidelberg, 2010.

[13] Y. Kutoyants, Statistical Inference for Ergodic Diffusion Processes, Springer Series in Statistics, Springer, 2004.

[14] M. Mohammadi, Analysis of Discretization Schemes for Fokker–Planck Equations and Related Optimality Systems, Ph.D. thesis, Julius-Maximilians-Universität Würzburg, 2015.

[15] H. Risken, T. Frank, The Fokker–Planck Equation: Methods of Solution and Applications, Springer Series in Synergetics, Springer, Berlin, Heidelberg, 1996.

Please cite this article in press as: B. Kaltenbacher, B. Pedretscher, Parameter estimation in SDEs via the Fokker–Planck equation: Likelihood function and adjoint based gradient computation, J. Math. Anal. Appl. (2018), https://doi.org/10.1016/j.jmaa.2018.05.048