On parameter identification in stochastic differential equations by penalized maximum likelihood

Fabian Dunker[1,2] **and Thorsten Hohage**[1]

[1]Institut für Numerische und Angewandte Mathematik, Georg-August Universität Göttingen, Lotzestr. 16–18, 37083 Göttingen, Germany
[2]Department of Economics, Boston College,
140 Commonwealth Avenue, Chestnut Hill, MA 02467, USA

E-mail: dunker@math.uni-goettingen.de

Abstract. In this paper we present nonparametric estimators for coefficients in stochastic differential equation if the data are described by independent, identically distributed random variables. The problem is formulated as a nonlinear ill-posed operator equation with a deterministic forward operator described by the Fokker-Planck equation. We derive convergence rates of the risk for penalized maximum likelihood estimators with convex penalty terms and for Newton-type methods. The assumptions of our general convergence results are verified for estimation of the drift coefficient. The advantages of log-likelihood compared to quadratic data fidelity terms are demonstrated in Monte-Carlo simulations.

1. Introduction

Many dynamical processes in physics, social sciences and economics can be modeled by systems of stochastic differential equations

$$d\mathbf{X}_t = \boldsymbol{\mu}(t, \mathbf{X}_t)dt + \sigma(t, \mathbf{X}_t)d\mathbf{W}_t. \tag{1}$$

Here $t \in [0, T]$ with $T > 0$ is interpreted as time, $\mathbf{X}_t$ is a family of random variables with values in $\mathbb{R}^d$, and $\mathbf{W}_t$ is a standard Wiener process in $\mathbb{R}^d$. The function $\boldsymbol{\mu} : [0, T] \times \mathbb{R}^d \to \mathbb{R}^d$ is called drift coefficient while $\sigma : [0, T] \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is the volatility or diffusion. Observations of the process give values of one or more paths $(\mathbf{X}_t)_{t\geq 0}$ at one or many times t. In many applications there is an interest to estimate the drift or the diffusion either non-parametrically or parametrically to gain a better understanding of the modeled process.

In this paper we consider the time homogenous differential equations, i.e. $\boldsymbol{\mu}$ and σ are independent of t. Nevertheless, the framework of our estimators also allows for time dependent coefficients. We are particularly interested in the case where σ is known while $\boldsymbol{\mu}$ should be estimated. Let us describe two kinds of observations sufficient for the time homogenous case:

(i) An ensemble of independent paths $\mathbf{X}_t^{(i)}$, $i = 1, \dots, n$ is observed at a fixed time $t = T$. I.e. the observations are the random variables $\mathbf{Y}_i = \mathbf{X}_T^{(i)}$. The starting points of the paths $\mathbf{X}_0^{(i)}$ are assumed to be sampled from a known distribution u_0

(ii) We observe only one path of a strictly stationary, ergodic process at equidistant points in time. I.e. our observations are $\mathbf{Y}_i = \mathbf{X}_{(i+i_0)\Delta t}$ for $i = 1, \dots, n$ and $i_0 > 0$.

Our approach to the problem is based on the Fokker-Planck equation, also called forward Kolmogorov equation. Assume $\mathbf{X}_t$ has a sufficiently smooth density $u(t, \cdot)$ for all $t \in [0, T]$.Then (1) holds true if and only if u solves the initial value problem

$$\begin{aligned} \frac{\partial}{\partial t} u = & \text{ div } \left(-\boldsymbol{\mu} u + \frac{1}{2} \sigma \sigma^\top \text{ grad } u \right) \\ u(0, \cdot) = & \, u_0 \end{aligned} \tag{2}$$

(see e.g. [31]). Hence, we can define the deterministic coefficient-to-solution operator $F(\boldsymbol{\mu}) := u(T, \cdot)$. This operator is nonlinear.

In case of an ergodic process with $\boldsymbol{\mu}$, σ not depending on t, solutions to eq. (2) tend to a stationary solution as $t \to \infty$ which solves the elliptic equation

$$\begin{aligned} 0 = & \text{ div } \left(-\boldsymbol{\mu} u + \frac{1}{2} \sigma \sigma^\top \text{ grad } u \right) \\ \int & u(x) dx = 1. \end{aligned} \tag{3}$$

Here the coefficient-to-solution operator is defined by $F(\boldsymbol{\mu}) := u$. The operator F and its properties will be discussed in Section 2.

We will derive convergence results for general operators F with values in a set of probability densities. The unknown of the inverse problem will be denoted by f in this general case. In the setting above we have $f = \boldsymbol{\mu}$, but in other applications $f = \sigma$ or $f = (\boldsymbol{\mu}, \sigma)$ are possible as well. If parametric estimation is preferred over non-parametric estimation, f can be a parameter in a model of $\boldsymbol{\mu}$ or σ. Suppose that $f^\dagger$ is the exact solution and $u^\dagger := F(f^\dagger)$ the corresponding probability density. We assume that the observed data are described by independent random variables $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ each of which has probability density $u^\dagger$. Note that equidistant observations $\mathbf{Y}_i = \mathbf{X}_{(i+i_0)\Delta t}$ of one path are actually not independent. Therefore, our results apply immediately only to the first scenario where an ensemble of independent paths is observed. In the second scenario additional information is contained in the order of the data $\mathbf{Y}_i$ which will be neglected here. This is justified if Δ_t is so large that the dependence of $\mathbf{Y}_i$ and $\mathbf{Y}_{i+1}$ is negligible or if only the set $\{\mathbf{Y}_1, \dots, \mathbf{Y}_n\}$ without ordering is available. (Alternatively, one could consider pairs $(\mathbf{Y}_i, \mathbf{Y}_{i+1})$ as sampled from the transitional probability density function weighted by u and use the forward operator F mapping f to this weighted transitional probability density function (see [22]). F is then characterized by the parabolic equation (2).)

Our estimator follows the idea to seek an estimator $\widehat{f}$ which maximizes the likelihood of the given observations $\mathbf{Y}_i = y_i$. It is convenient to describes these observations by

the empirical measure

$$\Phi_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{y_i}. \tag{4}$$

Since $\mathbb{P}_u[y_1, \dots, y_n] = \prod_{i=1}^n u(y_i)$, the negative log-likelihood is given by

$$\mathcal{S}_0(\Phi_n, u) = -\frac{1}{n}\ln \mathbb{P}_u[y_1, \dots, y_n] = -\frac{1}{n}\sum_{i=1}^{n} \ln u(y_i) = -\int \ln(u)\, d\Phi_n. \tag{5}$$

Due to ill-posedness a simple maximum likelihood estimator, i.e. a minimizer of $\mathcal{S}_0(\Phi_n, F(f))$ over f in some convex set $\mathfrak{B}$, is unstable. Therefore, we have to regularize. In (generalized) Tikhonov regularization one adds a penalty term $\mathcal{R} : \mathfrak{B} \to \mathbb{R} \cup \{\infty\}$, which we assume to be convex, lower semi-continuous, and not identically ∞. It is weighted by a regularization parameter $\alpha > 0$:

$$\widehat{f}_\alpha \in \operatorname{argmin}_{f\in\mathfrak{B}} \left[\mathcal{S}(\Phi_n; F(f)) + \alpha\mathcal{R}(f)\right]. \tag{6}$$

If the operator F is linear, this is a convex minimization problem. But for non-linear F this is in general a non-convex minimization. The non-convexity can be avoided by locally approximating F around a current iterate by its Fréchet derivative $F'[\widehat{f}_k]$. This yields the iteratively regularized Newton method

$$\widehat{f}_k \in \operatorname{argmin}_{f\in\mathfrak{B}} \left[\mathcal{S}\left(\Phi_n; F'[\widehat{f}_{k-1}](f - \widehat{f}_{k-1}) + F(\widehat{f}_{k-1})\right) + \alpha_k\mathcal{R}(f)\right]. \tag{7}$$

Here (α_k) is a sequence of positive regularization parameters converging monotonically to 0 for increasing k such that α_k/α_{k+1} remains bounded. To assure well-posedness of these optimization problems and to analyze convergence, it is often necessary to "regularize" the data fidelity term $\mathcal{S}$. This is of particular importance when u is negative on a set of positive measure which implies $\mathcal{S}(\Phi_n, u) = \infty$. A further discussion is contained in Section 3.

All known convergence rate results for regularization methods involving F' under source conditions weaker than $f^\dagger \in \operatorname{ran}(F'[f^\dagger]^*)$ require additional assumptions on F' such as the tangential cone condition

$$\|F(g) - F(f) - F'[f](g - f)\|_{L^2} \le \eta\|F(g) - F(f)\|_{L^2}. \tag{8}$$

For KL-type data fidelity terms a related formulation (20) suggested recently in [21] is required. For parameter identification problems for which $D(F)$ and $\operatorname{ran}(F)$ are function spaces over different domains these conditions are typically very difficult to verify, but if the domains coincide the L^2 tangential cone condition has been shown for a number of problems (see e.g. [18, 5]). To the best of our knowledge for drift estimation in the stationary Fokker-Planck equation (3) both the L^2-version and in particular the KL-version of the tangential cone condition are unknown so far, and we will prove them below.

The modeling by stochastic differential equations became standard in financial econometrics since the work of Black& Scholes [2]. The parametric and non-parametric estimation of drift and diffusion has attracted a lot of interest since then. We just

mention the text book by Kutoyants [24] and references therein. More recent works on nonparametric estimation of the drift are those by Hoffmann [19] using wavelets, Spokoiny [37] using kernel methods, Gobet, Hoffmann & Reiß [16] using wavelet estimation of an eigenvalue-eigenfunction pair of the transition operator, Comte, Genon-Catalot & Rozenholc [6] using penalized least squares, Schmisser [33] applying penalized least squares to high dimensional problems, Papaspiliopoulos et al. [27], Pokern, Stuart & van Zanten [28] using Bayesian methods. A parametric estimator related to our approach was developed by Hurn, Jeismann & Lindsay [22]. They propose a maximum likelihood estimator which relies on the computation of (9) by finite elements. Due to a parametric model for $\boldsymbol{\mu}$ their problem is not ill-posed. Furthermore, we mention Crépey [7, 8], Egger & Engl [12] and De Cezaro, Scherzer & Zubelli [9] for nonparametric volatility estimation using partial differential equations.

We will show convergence in expectation results with rates as $n \to \infty$ both for generalized Tikhonov regularization (6) and the iteratively regularized Newton method (7) by adapting corresponding results for inverse problems with Poisson data in [21, 39]. Here we make essential use of a version of Talagrand's concentration inequality formulated by Massart [25].

Deterministic variational regularization with general convex penalty terms have recently been investigated in a number of papers. We just mention Eggermont [13], Burger & Osher [4], Resmerita & Anderssen [29], Hofmann et al. [20], Grasmair [17], Kaltenbacher & Hofmann [23], and the monographs by Scherzer et al. [32], Schuster, Kaltenbacher, Hofmann & Kazimierski [35] and Flemming [14] and the references therein.

The remainder of this paper is organized as follows: In the next section we present some properties of the Fokker-Planck equation and prove a tangential cone condition for the corresponding forward operator F. In Section 3 general convergence rates results for variational regularization methods with Kullback-Leibler-type data fidelity and convex penalty term are presented. These results are applied to our estimator of the drift in Section 4. Results of numerical simulations are shown in Section 5. We end this paper with some conclusions.

2. Fokker-Planck equation

In this section we collect some properties of the stationary Fokker-Planck equation and prove the L^2 tangential cone condition for the corresponding operator F. We consider this equation on a bounded Lipschitz domain $D \subset \mathbb{R}^d$ with the no-flux boundary condition. I.e. in terms of probability densities no probability mass enters or leaves through the boundary. It is the natural boundary condition for the Fokker-Planck

equation:

$$\begin{aligned} &\operatorname{div}\left(-\boldsymbol{\mu}u + \frac{1}{2}\sigma\sigma^\top \operatorname{grad} u\right) = 0 \qquad \text{in } D \\ &-u(\boldsymbol{\mu}\cdot\mathbf{n}) + \frac{1}{2}\left(\sigma\sigma^\top \operatorname{grad} u\right)\cdot\mathbf{n} = 0 \qquad \text{on } \partial D \\ &\int_D u(x)dx = 1. \end{aligned} \tag{9}$$

We assume that $\boldsymbol{\mu} \in L^\infty(D,\mathbb{R}^d)$ and $\sigma \in L^\infty(D)^{d\times d}$ with well-defined L^∞ traces on ∂D which appear in the boundary condition. Moreover, we assume that there exists a constant $C_\sigma > 0$ such that

$$|\sigma(x)^\top \xi|_2 \geq C_\sigma |\xi|_2 \quad \text{for all } \xi \in \mathbb{R}^d, \text{ and all } x \in \overline{D}. \tag{10}$$

Let us comment on the natural boundary condition of the Fokker-Planck equation:

- In case $d = 1$ we can assume w.l.o.g. that $D = (-1, 1)$. Extend $\boldsymbol{\mu}$ by $\boldsymbol{\mu}(x) := \boldsymbol{\mu}(1)$ and $\boldsymbol{\mu}(-x) := \boldsymbol{\mu}(-1)$ for $x > 1$ and similarly for σ. Since the constant coefficient differential equation $-\boldsymbol{\mu}u' + \frac{\sigma^2}{2}u'' = 0$ with $\boldsymbol{\mu} \neq 0$ has the linearly independent solutions 1 and $\exp\left(\frac{2\boldsymbol{\mu}}{\sigma^2}x\right)$, the Fokker-Planck equation on $\mathbb{R}$ has an integrable solution if and only if $\boldsymbol{\mu}(1) < 0$ and $\boldsymbol{\mu}(-1) > 0$. In this case every integrable solution satisfies
  $$u(x) = u(1)\exp\left(\frac{2\boldsymbol{\mu}(1)}{\sigma(1)^2}(x-1)\right), \quad u(-x) = u(-1)\exp\left(\frac{2\boldsymbol{\mu}(-1)}{\sigma(-1)^2}(1-x)\right), \quad x \geq 1.$$
  Therefore, these solutions satisfy the boundary condition in (9). Hence, the restrictions of solutions to (3) restricted to $D = (-1, 1)$ are solutions to (9) up to a scaling factor, i.e. the boundary condition is an exact transparent boundary condition. This is how the boundary condition will be interpreted in our numerical experiments.
- For $d > 1$ exact transparent boundary conditions are always non-local. Since the boundary condition in (9) is local, we may at best hope for convergence to a solution of the Fokker-Planck equation in $\mathbb{R}^d$ as the size of D tends to ∞.
- In other applications, e.g. diffusion in biological cells the solution paths $\mathbf{X}_t$ are naturally contained in a subdomain D of $\mathbb{R}^d$. In this case the behavior at the boundary has to be modeled separately. E.g. when a path hits the boundary, it may be reflected in a certain way with a certain probability and otherwise destroyed. As discussed in [36, 34] and references therein, the behavior of the probability densities at the boundary may be rather complex involving boundary layers, but no-flux boundary conditions often appear as the limiting model.

The weak formulation of the elliptic problem (9) is to find $u \in H^1(D)$ such that

$$\int_D u dx = 1, \qquad a_{\boldsymbol{\mu}}(u, v) = 0 \quad \text{for all } v \in H^1(D) \tag{11}$$

where

$$a_{\boldsymbol{\mu}}(u, v) := \int_D \left(-\boldsymbol{\mu}u \cdot \operatorname{grad} v + \frac{1}{2}\sigma\sigma^\top \operatorname{grad} u \cdot \operatorname{grad} v\right) dx.$$

Let $L_{\boldsymbol{\mu}} : H^1(D) \to H_0^{-1}(D)$ denote the operator associated to $a_{\boldsymbol{\mu}}$, i.e. $\langle L_{\boldsymbol{\mu}} u, v\rangle = a_{\boldsymbol{\mu}}(u,v)$ for all $u, v \in H^1(D)$. It was proven by Droniou and Vázquez [10] that every function in the kernel of $L_{\boldsymbol{\mu}}$ is either a.e. positive, a.e. negative, or a.e. 0. Therefore, the kernel is either trivial or one-dimensional. For the convenience of the reader we collect some further properties of $L_{\boldsymbol{\mu}}$ all of which are more or less explicitly contained in [10].

Lemma 1. *Assume (10) for σ and let $\boldsymbol{\mu} \in L^\infty(D, \mathbb{R}^d)$.*

(i) The following Gårding inequality holds with $\gamma > \|\boldsymbol{\mu}\|_\infty^2/(2C_\sigma)$ and $0 < c < \min\left\{\gamma - \frac{\|\boldsymbol{\mu}\|_\infty^2}{2C_\sigma}, \frac{C_\sigma}{2} - \frac{\|\boldsymbol{\mu}\|_\infty^2}{4\gamma}\right\}$

$$a_{\boldsymbol{\mu}}(u,u) + \gamma\|u\|_{L^2}^2 \geq c\|u\|_{H^1}^2, \qquad u \in H^1(D).$$

(ii) Eq. (11) has a unique solution.

(iii) Let $H^1_\diamond(D) := \{u \in H^1(D) | \int u dx = 0\}$, let $\tilde{a}_{\boldsymbol{\mu}} : H^1_\diamond(D) \times H^1_\diamond(D) \to \mathbb{R}$ denote the restriction of $a_{\boldsymbol{\mu}}$ to $H^1_\diamond(D)$, and let $\tilde{L}_{\boldsymbol{\mu}} : H^1_\diamond(D) \to H^1_\diamond(D)^*$ denote the operator associated to $\tilde{a}_{\boldsymbol{\mu}}$. Then $\tilde{L}_{\boldsymbol{\mu}}$ is bijective and has a bounded inverse.

Proof. (i) We have

$$\begin{aligned} a_{\boldsymbol{\mu}}(u,u) + \gamma\|u\|_{L^2}^2 &= \int_D -\boldsymbol{\mu} u \operatorname{grad} u + \frac{1}{2}\left|\sigma^\top \operatorname{grad} u\right|_2^2 dx + \gamma\|u\|_{L^2}^2 \\ &\geq -\|\boldsymbol{\mu}\|_\infty \|u\|_{L^2} \|\operatorname{grad} u\|_{L^2} + \frac{C_\sigma}{2}\|\operatorname{grad} u\|_{L^2}^2 + \gamma\|u\|_{L^2}^2 \\ &\geq \left(\gamma - \frac{\|\boldsymbol{\mu}\|_\infty^2}{4\varepsilon}\right)\|u\|_{L^2}^2 + \left(\frac{C_\sigma}{2} - \varepsilon\right)\|\operatorname{grad} u\|_{L^2}^2. \end{aligned}$$

The last step uses Young's inequality $ab \leq a^2/(4\epsilon) + \epsilon b^2$, which holds for $a, b \geq 0$ and $\varepsilon > 0$. Choosing $\varepsilon < C_\sigma/2$ and $\gamma > \|\boldsymbol{\mu}\|_\infty^2/(4\varepsilon)$ gives the Gårding inequality.
(ii) As a consequence of part 1, $L_{\boldsymbol{\mu}}$ is a Fredholm operator of index 0, i.e. $\dim(\ker(L_{\boldsymbol{\mu}})) = \dim(\operatorname{ran}(L_{\boldsymbol{\mu}})^\perp)$ (where orthogonality is understood with respect to the dual pairing of $H^1(D)$ and $H_0^{-1}(D)$) and $\operatorname{ran}(L_{\boldsymbol{\mu}})$ is closed. As argued above, $\dim(\ker(L_{\boldsymbol{\mu}})) \in \{0,1\}$. As $a_{\boldsymbol{\mu}}(u,1) = 0$ for all $u \in H^1(D)$, i.e. $1 \in \operatorname{ran}(L_{\boldsymbol{\mu}})^\perp$, we have $\dim(\ker(L_{\boldsymbol{\mu}})) = 1$. Since the elements of $\ker(L_{\boldsymbol{\mu}})$ are positive a.e. or negative a.e., there exists a unique $u \in \ker(L_{\boldsymbol{\mu}})$ satisfying $\int_D u\, dx = 1$.
(iii) We also have $\dim(\operatorname{ran}(L_{\boldsymbol{\mu}})^\perp) = 1$, so by the proof of part 2 $\operatorname{ran}(L_{\boldsymbol{\mu}}) = \{1\}^\perp = H^1_\diamond(D)^*$ as $\operatorname{ran}(L_{\boldsymbol{\mu}})$ is closed. By the characterization of $\ker(L_{\boldsymbol{\mu}})$, the operator $L_{\boldsymbol{\mu}}$ is injective on $H^1_\diamond(D)$. Moreover, $\operatorname{ran}(\tilde{L}_{\boldsymbol{\mu}}) = \operatorname{ran}(L_{\boldsymbol{\mu}})$ as $H^1_\diamond(D) \oplus \operatorname{span}\{1\} = H^1(D)$, so $\tilde{L}_{\boldsymbol{\mu}}$ is surjective. Boundedness of $\tilde{L}_{\boldsymbol{\mu}}^{-1}$ follows from the open mapping theorem. □

The differentiability of F and the tangential cone condition stated in the next theorem are crucial for the Gauß-Newton method.

Theorem 2. *The operator $F : L^\infty(D, \mathbb{R}^d) \to L^2(D)$ is Fréchet differentiable, and $F'[\boldsymbol{\mu}]\mathbf{h} = u'_{\boldsymbol{\mu},\mathbf{h}}$ where $u'_{\boldsymbol{\mu},\mathbf{h}} \in H^1_\diamond(D)$ is the unique solution to the variational problem*

$$\tilde{a}_{\boldsymbol{\mu}}(u'_{\boldsymbol{\mu},\mathbf{h}}, v) = \int_D F(\boldsymbol{\mu})\, \mathbf{h} \cdot \operatorname{grad} v\, dx, \qquad v \in H^1_\diamond(D). \tag{12}$$

Furthermore, the strong tangential cone condition holds true:

$$\|F(\boldsymbol{\mu}+\mathbf{h})-F(\boldsymbol{\mu})-F'[\boldsymbol{\mu}]\mathbf{h}\|_{L^2} \le \tilde{C}_{\boldsymbol{\mu}}\|\mathbf{h}\|_\infty \|F(\boldsymbol{\mu}+\mathbf{h})-F(\boldsymbol{\mu})\|_{L^2} \tag{13}$$

for all $\boldsymbol{\mu}, \mathbf{h} \in L^\infty(D, \mathbb{R}^d)$ *with* $\tilde{C}_{\boldsymbol{\mu}} := \|\tilde{L}_{\boldsymbol{\mu}}^{-1}\|$.

Proof. Note that $\tilde{u} := F(\boldsymbol{\mu}+\mathbf{h}) - F(\boldsymbol{\mu})$ belongs to $H^1_\diamond(D)$ and satisfies

$$\tilde{a}_{\boldsymbol{\mu}}(\tilde{u}, v) = \int_D (F(\boldsymbol{\mu})+\tilde{u})\ \mathbf{h} \cdot \operatorname{grad} v \, dx, \qquad v \in H^1_\diamond(D).$$

For $v \neq 0$ the functional on the right hand side is bounded by

$$\frac{1}{\|v\|_{H^1}} \int_D (F(\boldsymbol{\mu})+\tilde{u})\ \mathbf{h} \cdot \operatorname{grad} v \, dx \le \|\mathbf{h}\|_\infty \|F(\boldsymbol{\mu})+\tilde{u}\|_{L^2} \le \|\mathbf{h}\|_\infty \left(\|F(\boldsymbol{\mu})\|_{L^2} + \|\tilde{u}\|_{H^1}\right)$$

Therefore, $\|\tilde{u}\|_{H^1} \le \tilde{C}_{\boldsymbol{\mu}}\|\mathbf{h}\|_\infty \left(\|F(\boldsymbol{\mu})\|_{L^2} + \|\tilde{u}\|_{H^1}\right)$, which implies

$$(1 - \tilde{C}_{\boldsymbol{\mu}}\|\mathbf{h}\|_\infty)\|\tilde{u}\|_{H^1} \le \tilde{C}_{\boldsymbol{\mu}}\|\mathbf{h}\|_\infty \|F(\boldsymbol{\mu})\|_{L^2}.$$

Hence, F is continuous since $\|F(\boldsymbol{\mu}+\mathbf{h}) - F(\boldsymbol{\mu})\|_{H^1} = \|\tilde{u}\|_{H^1}$ tends to 0 as $\|\mathbf{h}\|_\infty$ tends to 0. As

$$\tilde{a}_{\boldsymbol{\mu}}(\tilde{u} - u_{\boldsymbol{\mu},\mathbf{h}}, v) = \int_D \tilde{u}\ \mathbf{h} \cdot \operatorname{grad} v \, dx, \qquad v \in H^1_\diamond(D),$$

a similar estimate of the right hand side as above yields the bound

$$\|F(\boldsymbol{\mu}+\mathbf{h}) - F(\boldsymbol{\mu}) - u'_{\boldsymbol{\mu},\mathbf{h}}\|_{L^2} = \|\tilde{u} - u'_{\boldsymbol{\mu},\mathbf{h}}\|_{L^2} \le \|\tilde{u} - u'_{\boldsymbol{\mu},\mathbf{h}}\|_{H^1} \le \tilde{C}_{\boldsymbol{\mu}}\|\mathbf{h}\|_\infty \|\tilde{u}\|_{L^2},$$

which shows the tangential cone condition. Together with the continuity of F this implies that F is Fréchet differentiable, and $F'[\boldsymbol{\mu}]\mathbf{h} = u'_{\boldsymbol{\mu},\mathbf{h}}$. □

Example 3. *If* $\boldsymbol{\mu}$ *has a representation of the form*

$$\boldsymbol{\mu} = \sigma\sigma^\top \operatorname{grad} \phi \tag{14}$$

for some $\phi \in H^1(D)$ *the solution of the stationary Fokker-Planck equation (11) is given explicitly by*

$$u = \frac{1}{\int_D \exp(2\phi)\, dx} \exp(2\phi),$$

since

$$\operatorname{grad} u = \frac{2}{\int_D \exp(2\phi)\, dx} \operatorname{grad} \phi \exp(2\phi) = 2(\sigma\sigma^\top)^{-1}\boldsymbol{\mu} u.$$

The normalization constant $\int_D \exp(2\phi)\, dx$ *ensures that* u *is a density. In particular, we obtain the following explicit formula for the inverse of* F*:*

$$\boldsymbol{\mu} = \frac{\sigma\sigma^\top \operatorname{grad} u}{2u.} \tag{15}$$

The methods discussed below do not rely on this formula and the assumption (14).

3. General convergence results for inverse problems with i.i.d. sample data

In this section we consider the following general setting:

- $\mathbb{X}$ is a Banach space, $\mathfrak{B} \subset \mathbb{X}$ a convex subset, $D \subset \mathbb{R}^d$ a bounded Lipschitz domain, and $H^s(D)$ with $s > \frac{d}{2}$ an L^2-based Sobolev space.
- The range of operator $F : \mathfrak{B} \to H^s(D)$ consists of probability densities, i.e. $F(f) \geq 0$ and $\int_D F(f)\,dx = 1$ for all $f \in \mathfrak{B}$.
- There exists $R > 1$ such that $\sup_{f\in\mathfrak{B}} \|F(f)\|_{H^s} \leq R$.
- $f^\dagger \in \mathfrak{B}$ is the exact solution, $u^\dagger := F(f^\dagger)$, and observations are described by independent random variables $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ with density $u^\dagger$. Recall the definition of the empirical measure Φ_n in (4).

A concentration inequality. Note that

$$\mathbb{E}\left[\int_D \varphi\, d\Phi_n\right] = \int_D \varphi u^\dagger\, dx, \qquad \text{and} \qquad \mathbb{V}ar\left[\int_D \varphi\, d\Phi_n\right] = \frac{1}{n}\int_D \varphi^2 u^\dagger\, dx$$

whenever the right hand sides are well-defined. We will need a concentration inequality which is uniform in φ. Our starting point is a version of the concentration inequality in the seminal work by Talagrand [38], which is due to Massart [25] and has explicit constants. In our notation a special case of this inequality can be stated as follows:

Theorem 4 (Theorem 3 in [25])**.** *Let $\mathcal{F} \subset L^\infty(D)$ be a countable family of functions with $\|\varphi\|_\infty \leq b$ for all $\varphi \in \mathcal{F}$. Moreover, let*

$$Z := n \sup_{\varphi\in\mathcal{F}} \left| \int_D \varphi(d\Phi_n - u^\dagger dx) \right|$$

and $v := n \sup_{\varphi\in\mathcal{F}} \int_D \varphi^2 u^\dagger\, dx$. Then

$$\mathbb{P}\left[Z \geq (1+\epsilon)\mathbb{E}[Z] + \sqrt{8v\xi} + \kappa(\epsilon)b\xi\right] \leq \exp(-\xi)$$

for all $\epsilon, \xi > 0$ where $\kappa(\epsilon) = 2.5 + 32/\epsilon$.

Massart also proved a similar inequality for the left tail of Z, but we only need the inequality above, so we might rather speak of a deviation inequality.

In analogy to [39] where similar results were derived using a concentration inequality for Poisson processes in [30] instead of Theorem 4, we show the following corollary:

Corollary 5. *There exists a constant $C_\mathrm{c} \geq 1$ depending only on D and s such that for $\rho \geq RC_\mathrm{c}$ and for all $n \in \mathbb{N}$*

$$\mathbb{P}\left[\sup_{\|\varphi\|_{H^s(D)}\leq R} \left| \int_D \varphi\left(d\Phi_n - u^\dagger dx\right) \right| \geq \frac{\rho}{\sqrt{n}}\right] \leq \exp\left(-\frac{\rho}{RC_\mathrm{c}}\right). \tag{16}$$

Proof. (Sketch) The most difficult part in the derivation of Corollary 5 from Theorem 4 is the estimation of $\mathbb{E}[Z]$. In analogy to [39, Lemma A.2] we can prove that

$$\mathbb{E}[Z] \leq \sqrt{n} C_1 R$$

with a constant C_1 depending only on s and D. As $H^s(D)$ is continuously embedded in $L^\infty(D)$, we have $\|\varphi\|_\infty \le C_2 R$ for all $\varphi \in H^s(D)$ with $\|\varphi\|_{H^s} \le R$ where C_2 is the norm of the embedding operator. Moreover, $v \le n(C_2R)^2$ as $\|u^\dagger\|_{L^1} = 1$. Using the separability of balls in $H^s(D)$ and choosing $\epsilon = 1$ in Theorem 4 we obtain

$$\mathbb{P}\left[\sup_{\|\varphi\|_{H^s(D)}\le R}\left|\int_D \varphi\left(d\Phi_n - u^\dagger dx\right)\right| \ge \left(\frac{2C_1}{\sqrt{n}} + \frac{C_2\sqrt{8\xi}}{\sqrt{n}} + \frac{34.5C_2\xi}{n}\right)R\right] \le \exp(-\xi).$$

As $\frac{1}{n} \le \frac{1}{\sqrt{n}}$ and $\sqrt{\xi} \le \xi$ for $\xi \ge 1$, this yields (16) with $C_{\mathrm{c}} := 2C_1 + (34.5 + \sqrt{8})C_2$ and $\rho = RC_{\mathrm{c}}\xi$. □

Distance measures. To state our convergence theorems we need two distance measures one in $\mathbb{X}$ and one in $L^1(D)$. As usual, convergence rates of variational regularization methods are proved for the Bregman distance associated to the penalty functional. The Bregman distance with respect to $\mathcal{R}$ and $f^* \in \partial\mathcal{R}(f^\dagger)$ is defined as

$$D_{\mathcal{R}}^{f^*}(f,\, f^\dagger) := \mathcal{R}(f) - \mathcal{R}(f^\dagger) - \langle f^*, f - f^\dagger\rangle.$$

For quadratic penalty in Hilbert spaces we have $D_{\mathcal{R}}^{f^*}(f,\, f^\dagger) = \|f - f^\dagger\|^2$. In general, $D_{\mathcal{R}}^{f^*}$ is nonnegative with $D_{\mathcal{R}}^{f^*}(f^\dagger, f^\dagger) = 0$, but it is neither symmetric nor does it satisfy a triangle inequality.

In $L^1(D)$ we use the distance measure which corresponds to the negative log-likelihood introduced in (5) is the Kullback-Leibler divergence

$$\mathrm{KL}(u; v) := \int_D v - u - u\ln\left(\frac{v}{u}\right) dx$$

with the convention $0\ln 0 := 0$ and $\ln(x) := -\infty$ for $x \le 0$. Note that $\mathrm{KL}(u^\dagger; v) = \mathbb{E}\left[\mathcal{S}_0(\Phi_n; v) - \mathcal{S}_0(\Phi_n; u^\dagger)\right]$, in other words KL is the expectation of the negative log-likelihood functional with an additive constant chosen in a way such that $\mathrm{KL}(u^\dagger; v) \ge 0$ for all v and $\mathrm{KL}(u^\dagger; u^\dagger) = 0$. If u and v are probability densities, the formula above simplifies to $\mathrm{KL}(u; v) = \int_D u\ln(v/u)\, dx$, but since the values of the linearization of F are not densities in general, we have to use the general formula.

Note that

$$\mathcal{S}_0(\Phi_n; v) - \mathcal{S}_0(\Phi_n; u^\dagger) - \mathrm{KL}(u^\dagger; v) = \int -\ln\frac{v}{u^\dagger}\left(d\Phi_n - u^\dagger dx\right).$$

To prove rates of convergence we have to bound the absolute value of the right hand side with sufficiently large probability. In principle, this can be done by applying Corollary 5 with $\varphi = -\ln\frac{v}{u^\dagger}$. However, this corollary is only applicable if we have uniform bounds $0 < c \le \frac{v}{u^\dagger} \le C < \infty$ for all $v \in F(\mathfrak{B})$, which is not always the case. Therefore, we introduce a shift parameter $\tau > 0$ and use $\mathrm{KL}(u^\dagger + \tau, v + \tau)$ as limiting data fidelity term and the corresponding empirical data fidelity term

$$\mathcal{S}_\tau(\Phi_n; v) = \int_D v dx - \int_D \ln(v + \tau)(d\Phi_n + \tau dx)$$

such that

$$\mathcal{S}_\tau(\Phi_n; v) - \mathcal{S}_\tau(\Phi_n; u^\dagger) - \mathrm{KL}(u^\dagger + \tau; v + \tau) = \int -\ln\frac{v+\tau}{u^\dagger + \tau}\left(d\Phi_n - u^\dagger dx\right).$$

Now we can bound

$$\mathrm{err} := \sup_{v \in F(\mathfrak{B})} \left| \mathcal{S}_\tau(\Phi_n; v) - \mathcal{S}_\tau(\Phi_n; u^\dagger) - \mathrm{KL}(u^\dagger + \tau; v + \tau) \right|$$

with high probability using Corollary 5 since $\sup_{v \in F(\mathfrak{B})} \| - \ln \frac{v+\tau}{u^\dagger + \tau} \|_{H^s} < \infty$ under our assumptions.

Convergence rate results. To obtain rates of convergence we need some kind of smoothness condition on the solution. Source conditions are commonly used for this purpose. In the regularization theory for Banach spaces they are formulated as variational inequalities (see [20] and [14] for relations to other formulations of source conditions). We assume that there exists a constant $\beta > 0$, $f^* \in \partial \mathcal{R}(f^\dagger)$ and a concave, strictly increasing function $\Lambda : [0, \infty[\to [0, \infty[$ with $\Lambda(0) = 0$ such that

$$\beta D_{\mathcal{R}}^{f^*}(f,\, f^\dagger) \leq \mathcal{R}(f) - \mathcal{R}(f^\dagger) + \Lambda\Big(\mathrm{KL}\left(u^\dagger + \tau; F(f) + \tau\right) \Big) \quad \text{for all } f \in \mathfrak{B}. \tag{17}$$

The proof of the following theorem is now completely analogous to the proof of [39, Theorem 4.3], but we point out that in [39, eq. (10)] on the left hand side $\mathbb{E}\left[\mathcal{S}(G_t; g^\dagger)\right]$ should be replaced by $\mathcal{S}(G_t; g^\dagger)$ and on the right hand side $\ln(g + \sigma)$ by $\ln \frac{g+\sigma}{g^\dagger + \sigma}$.

Theorem 6. *If $u^\dagger$ satisfies the variational source condition (17) for some $\tau > 0$, the nonlinear Tikhonov regularization (6) with $\mathcal{S} = \mathcal{S}_\tau$ has a global minimizer $\widehat{f}_\alpha$, and the regularization parameter is chosen such that*

$$\alpha^{-1} \in -\partial(-\Lambda)\left(\frac{2\rho}{\sqrt{n}}\right), \tag{18}$$

then we have

$$\mathbb{E}\left[D_{\mathcal{R}}^{f^*}(\widehat{f}_\alpha, f^\dagger) \right] = \mathcal{O}\left(\Lambda\left(\frac{1}{\sqrt{n}} \right) \right), \qquad n \to \infty. \tag{19}$$

For the convergence theorem of the Newton-type iteration we additionally have to impose a tangential cone condition adapted to our data fidelity term. Let

$$\mathcal{T}_\tau(u; v) := \begin{cases} \mathrm{KL}(u + \tau, v + \tau) & \text{if } v \geq -\tau/2 \\ \infty & \text{else.} \end{cases}$$

We assume that for all $f, g \in \mathfrak{B}$

$$\begin{aligned} \frac{1}{C_{\mathrm{tcc}}} \mathcal{T}_\tau\left(u^\dagger; F(g)\right) - \eta \mathcal{T}_\tau\left(u^\dagger; F(f)\right) &\leq \mathcal{T}_\tau\left(u^\dagger; F(f) + F'[f](g - f)\right) \\ &\leq C_{\mathrm{tcc}} \mathcal{T}_\tau\left(u^\dagger; F(g)\right) + \eta \mathcal{T}_\tau\left(u^\dagger; F(f)\right) \end{aligned} \tag{20}$$

with η sufficiently small and $C_{tcc} > 1$. We also set $\mathcal{S}_\tau(\Phi_n; v) := \infty$ if $v < -\tau/2$. Then we can show in analogy to [21]:

Theorem 7. *Let assumptions (17), (20) hold true. If $\widehat{f}_k$ is defined by the iteratively regularized Newton method (7) where $k \in \mathbb{N}$ is the largest index such that*

$$\alpha_k^{-1} \leq \sup -\partial(-\Lambda)\left(\frac{2\rho}{\sqrt{n}}\right), \tag{21}$$

then

$$\mathbb{E}\left[D_{\mathcal{R}}^{f^*}(\widehat{\boldsymbol{\mu}}_k, f^\dagger)\right] = \mathcal{O}\left(\Lambda\left(\frac{1}{\sqrt{n}}\right)\right). \tag{22}$$

Remark 1. (i) Related results exist for the iteratively regularized Gauß-Newton method with L^2 data fidelity term. Instead of (20), these theorems assume the L^2 tangential cone condition (8). Results like this were proven by Kaltenbacher and Hofmann [23], Hohage and Werner [21], or Dunker et al. [11]. The convergence rates for quadratic data fidelity terms compare to the rates in (22).

(ii) The selection rule (21) uses *a priori* information about the index function Λ which is usually not available in practice. It was shown in [21] that a data driven Lepskiĭ type parameter choice can be used instead. Only a logarithmic factor gets lost in the resulting convergence rate:

$$\mathbb{E}\left[D_{\mathcal{R}}^{f^*}(\widehat{f}_{k_{\text{Lepskii}}}, f^\dagger)\right] = \mathcal{O}\left(\Lambda\left(\frac{\ln(n^{-1})}{\sqrt{n}}\right)\right).$$

4. Convergence of the drift estimator

In order to apply Theorems 6 and 7 to the drift estimation problem we have to discuss the assumptions (17) and (20). For this purpose we need the following estimates for the Kullback-Leibler divergence:

Lemma 8. *The inequality*

$$\|\varphi - \psi\|_{L^2}^2 \le \left(\frac{2}{3}\|\varphi\|_\infty + \frac{4}{3}\|\psi\|_\infty\right) \mathrm{KL}(\varphi;\psi). \tag{23}$$

holds for all nonnegative functions $\varphi, \psi \in L^\infty(D)$ with $\varphi - \psi \in L^2(D)$. If ψ is bounded away from 0 then

$$\mathrm{KL}(\varphi;\psi) \le \left\|\frac{1}{\psi}\right\|_\infty \|\varphi - \psi\|_{L^2}^2. \tag{24}$$

Proof. The lower bound can be found in [3]. The upper bound follows from the simple estimation $x - 1 \ge \ln x$ which entails $(x-1)^2 \ge x \ln x - x + 1$. Setting $x = \varphi/\psi$ implies

$$\frac{1}{\psi}(\varphi - \psi)^2 \ge \psi - \varphi - \varphi \ln\left(\frac{\psi}{\varphi}\right).$$

Integrating this inequality over D and using $(1/\psi)(\varphi - \psi)^2 \le \|1/\psi\|_\infty (\varphi - \psi)^2$ yields (24). □

Proposition 9. *Let $s > d/2 + 1$, $\tau > 0$, and assume that D and σ are smooth. Then for every $\boldsymbol{\mu}^\dagger \in H^s(D;\mathbb{R}^d)$ there exists a ball $\mathfrak{B} \subset \{\boldsymbol{\mu} : \|\boldsymbol{\mu} - \boldsymbol{\mu}^\dagger\|_{H^s} < r\}$ such that F satisfies the Kullback-Leibler tangential cone condition (20) in $\mathfrak{B}$.*

Proof. As shown in [21, Lemma 5.2], the classical tangential cone condition (13) is equivalent to

$$\begin{aligned}\frac{1}{C}\left\|u^\dagger - F(g)\right\|_{L^2} - \tilde{\eta}\left\|u^\dagger - F(f)\right\|_{L^2} &\leq \left\|u^\dagger - F(f) - F'[f](g-f)\right\|_{L^2} \\ &\leq C\left\|u^\dagger - F(g)\right\|_{L^2} + \tilde{\eta}\left\|u^\dagger - F(f)\right\|_{L^2}\end{aligned}$$

for some constants $\tilde{\eta}, C > 0$ and all $f, g \in \mathfrak{B}$.

Next we are going to show that F is also continuously differentiable as a mapping from the Hölder space $C^{1,\beta}(\overline{D}, \mathbb{R}^d) \to L^\infty(D)$. Note that the solution u to (11) satisfies

$$\begin{pmatrix} \tilde{L}_{\boldsymbol{\mu}} & \mathbb{1} \\ \mathbb{1}^* & 0 \end{pmatrix} \begin{pmatrix} u \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

where $\mathbb{1}$ maps a constant $\lambda \in \mathbb{R}$ to the constant function with value λ on D, and $\mathbb{1}^*$ is its L^2-adjoint. By Schauder estimates (see e.g. [15]) the (block-)operator as a mapping from $C^{2,\beta}(\overline{D}) \times \mathbb{R} \to C^{0,\beta}(\overline{D}) \times \mathbb{R}$ has a bounded inverse if $\boldsymbol{\mu} \in C^{1,\beta}(\overline{D}, \mathbb{R}^d)$. Since the block operator depends continuously and affinely linear on $\boldsymbol{\mu}$ in these topologies and since the operator inversion is continuously differentiable, F is continuously Fréchet differentiable from $C^{1,\beta}(\overline{D}, \mathbb{R}^d)$ to $C^{2,\beta}(\overline{D})$ and hence from $C^{1,\beta}(\overline{D}, \mathbb{R}^d)$ to $L^\infty(D)$.

Choose $0 < \beta < s - d/2$. Then every ball $\mathfrak{B}$ in $H^s(D)$ is compact in $C^{1,\beta}(\overline{D})$, and the mappings $\boldsymbol{\mu} \mapsto \|F(\boldsymbol{\mu})\|_{L^\infty}$ and $\boldsymbol{\mu} \mapsto \|F'[\boldsymbol{\mu}]\|_{C^{1,\beta} \to L^\infty}$ are bounded on $\mathfrak{B}$ as continuous functions on a compact set. Together with Lemma 8 this implies (20) after possibly decreasing the radius of $\mathfrak{B}$. □

Proposition 10. *If $\mathcal{R}(\boldsymbol{\mu}) = \|\boldsymbol{\mu}\|_{H^s}^2$ with $s > d/2 + 1$, then every $\boldsymbol{\mu}^\dagger \in H^s(D; \mathbb{R}^d)$ satisfies a variational source condition of the form (17) in some H^s-ball.*

Proof. Due to the results in [26], $\boldsymbol{\mu}^\dagger$ satisfies a spectral source condition

$$\boldsymbol{\mu}^\dagger = \Theta(F'[\boldsymbol{\mu}^\dagger]^* F'[\boldsymbol{\mu}^\dagger])w$$

for some $w \in H^s(D; \mathbb{R}^d)$ and some index function Θ. Therefore, $\boldsymbol{\mu}^\dagger$ also satisfies a variational source condition for the linear operator $F'[\boldsymbol{\mu}^\dagger]$

$$\beta D_{\mathcal{R}}^{\boldsymbol{\mu}}(\boldsymbol{\mu}, \boldsymbol{\mu}^\dagger) \leq \mathcal{R}(\boldsymbol{\mu}) - \mathcal{R}(\boldsymbol{\mu}^\dagger) + \tilde{\Lambda}\Big(\|F'[\boldsymbol{\mu}]'(\boldsymbol{\mu} - \boldsymbol{\mu}^\dagger)\|_{L^2}^2\Big)$$

for all $\boldsymbol{\mu} \in H^s(D; \mathbb{R}^d)$ with another index function $\tilde{\Lambda}$ (see [14]). Note that the L^2 tangential cone condition in Theorem 2 implies

$$\|F'[\boldsymbol{\mu}^\dagger](\boldsymbol{\mu} - \boldsymbol{\mu}^\dagger)\|_{L^2} \leq (1 + \tilde{C}_{\boldsymbol{\mu}^\dagger}\|\boldsymbol{\mu}^\dagger - \boldsymbol{\mu}\|_\infty)\|F(\boldsymbol{\mu}) - F(\boldsymbol{\mu}^\dagger)\|_{L^2}$$

for all $\boldsymbol{\mu} \in H^s(D; \mathbb{R}^d)$. Therefore, $\boldsymbol{\mu}^\dagger$ also satisfies the variational source condition for the nonlinear operator F

$$\beta D_{\mathcal{R}}^{\boldsymbol{\mu}}(\boldsymbol{\mu}, \boldsymbol{\mu}^\dagger) \leq \mathcal{R}(\boldsymbol{\mu}) - \mathcal{R}(\boldsymbol{\mu}^\dagger) + \tilde{\Lambda}\Big(4\|u^\dagger - F(\boldsymbol{\mu})\|_{L^2}^2\Big)$$

for all $\boldsymbol{\mu} \in H^s(D; \mathbb{R}^d)$ with $\tilde{C}_{\boldsymbol{\mu}^\dagger}\|\boldsymbol{\mu} - \boldsymbol{\mu}^\dagger\|_\infty \leq 1$. Together with Lemma 8 and the continuous embedding of $H^s(D, \mathbb{R}^d)$ in $L^\infty(D, \mathbb{R}^d)$ this entails the KL related source condition (17). □

To sum up, all assumptions of Theorems 6 and 7 are satisfied for our problem. It would be interesting to have explicit characterizations of the index function Λ when $\boldsymbol{\mu}$ satisfies certain classical smoothness conditions. We intend to address this question in future research.

5. Numerical simulations

Implementation. The implementation of the iteration scheme (7) requires the evaluation of the forward operator F and its derivative F'. We did this for both operators by finite elements of degree 3. The convex minimization problem which occurs in every Newton step is solved by a nested Newton iteration as described in [21].

In addition to the iteration (7) we implemented the classical Gauß-Newton method with quadratic data fidelity term. As both methods were equipped with an H^1-quadratic penalty term, this setup allows for a comparison of the two methods. For the latter inversion scheme the minimization problem in every Newton step becomes quadratic and can be solved by a conjugate gradient method.

Test example. To test the algorithm we considered a one-dimensional stochastic differential equation (1) with diffusion $\sigma = 0.5$ and drift

$$\mu^\dagger(x) = -5x^3 - 2x - 0.25 \qquad \text{for } x \in [-1, 1], \tag{25}$$

$\mu^\dagger(x) = \mu^\dagger(1)$ for $x \geq 1$, and $\mu^\dagger(x) = \mu^\dagger(-1)$ for $x \leq -1$. The drift is plotted in Figures 6 and 7. We simulated a path of the stochastic process with the Euler-Maruyama method on a large time interval $[0, T]$ with $T = 1000$ and with 10^5 Euler steps. But we used only 125 to 1000 points in the time domain as observations of the path. This drift (25) is rather large in absolute values for $x = 1$ and $x = -1$ with a negative sign for $x = -1$. When the path jumped outside $[-1, 1]$ in the simulations it jumped back into the interval in a very small number of steps. The probability to have an observation of the simulated path outside of the interval is close to 0. To implement the forward operator we used transparent boundary conditions at -1 and 1 as described in section 2.

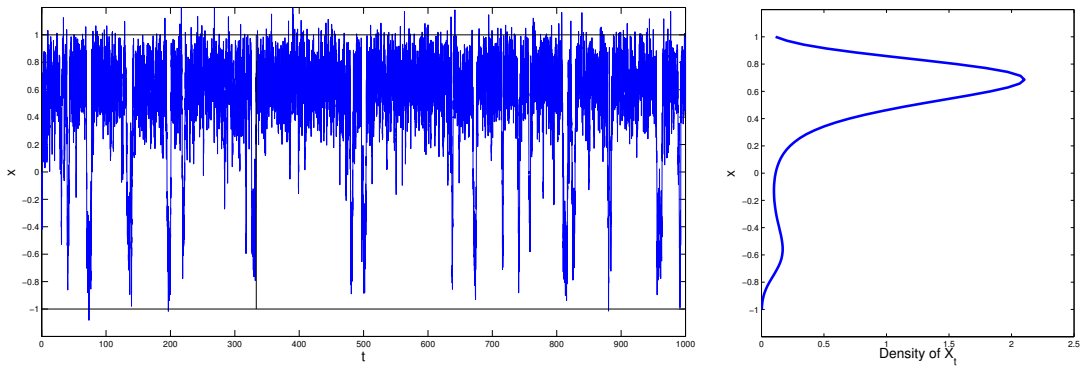

Figure 1: A simulated path and the corresponding limit density of the process X_t for $t \to \infty$.

Results. We reconstructed the drift using 4 different numbers of equidistant observations of a path namely 125, 250, 500, and 1000 points. For each set of observations we reconstructed the drift using the iteratively regularized Newton method (7) with KL data fidelity term and additionally using the iteratively regularized Gauß-Newton method. In both reconstruction methods we assumed that the drift is known in the semi-infinite intervals $(-\infty, -1]$ and $[1, \infty)$. In order to compare both methods independent of a stopping rule, in both cases an oracle choice of the stopping index was used. I.e. the stopping index was chosen such that the average L^2-error was minimal.

Due to the random error in the data, a statistic evaluation of the inversion methods is needed. For this purpose we repeated the procedure of simulating a path, drawing observations from it and conducting the estimations 1000 times. The following histograms show the distribution of the L^2 error of both methods. The error is normalized in a way such that the error of the initial guess is 1.

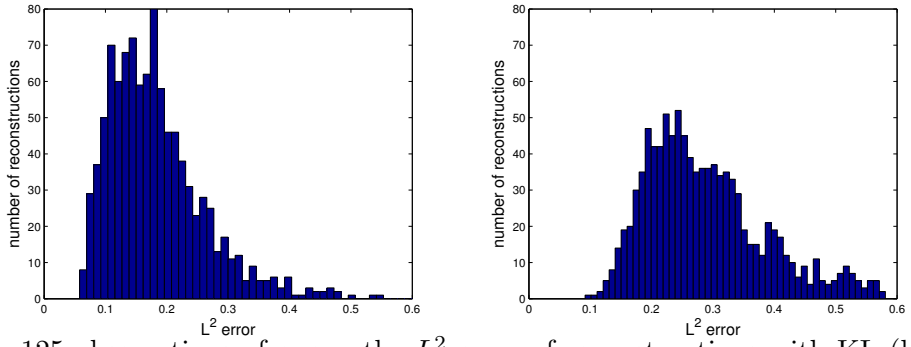

Figure 2: 125 observations of one path: L^2 error of reconstructions with KL (left) and L^2 (right) data fidelity term.

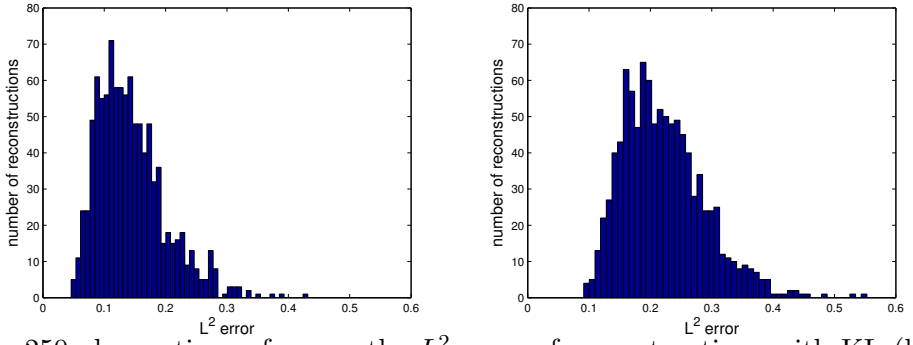

Figure 3: 250 observations of one path: L^2 error of reconstructions with KL (left) and L^2 (right) data fidelity term.

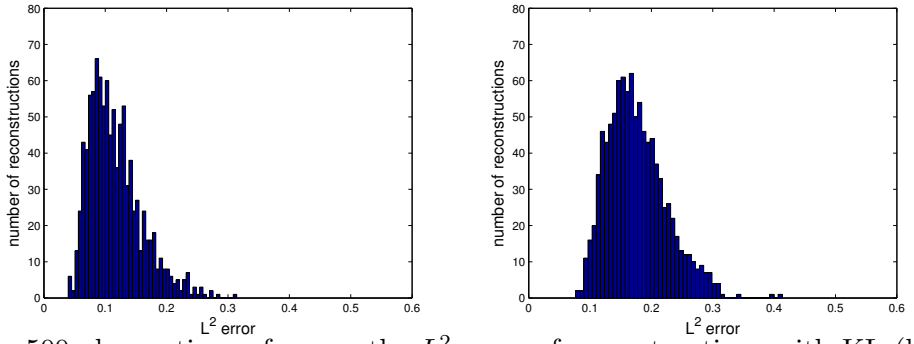

Figure 4: 500 observations of one path: L^2 error of reconstructions with KL (left) and L^2 (right) data fidelity term.

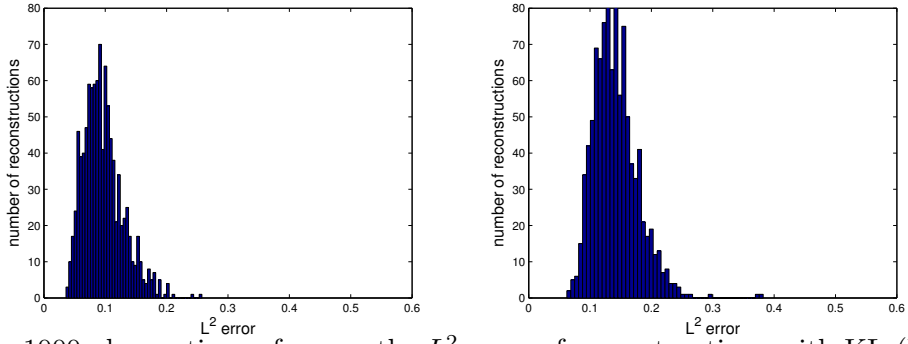

Figure 5: 1000 observations of one path: L^2 error of reconstructions with KL (left) and L^2 (right) data fidelity term.

The histograms suggest that the reconstructions with KL-type data fidelity term have a smaller mean error and smaller variance. This is made explicit by the following table:

observations	KL mean	KL variance	L^2 mean	L^2 variance
125	0.1832	0.0063	0.2870	0.0093
250	0.1439	0.0031	0.2212	0.0044
500	0.1160	0.0018	0.1759	0.0023
1000	0.0963	0.0010	0.1417	0.0013

Table 1: Mean and variance of the error distributions when one path is observed.

The following plots are typical reconstructions of the drift using a KL-type data fidelity term. We chose results with a median L^2 error for each sample size.

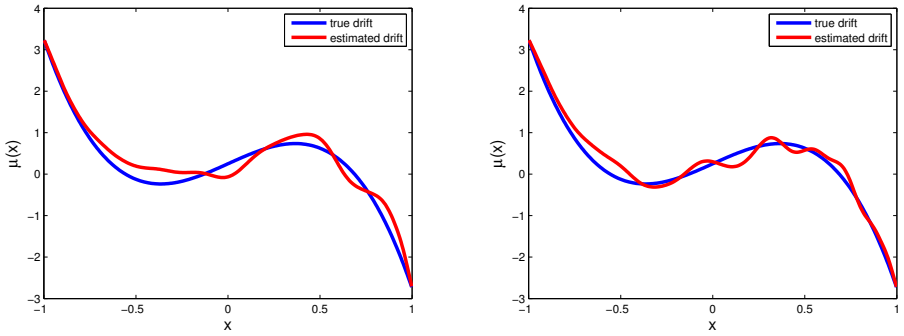

Figure 6: Median reconstructions with KL data fidelity term using 125 (left) and 250 (right) observations of one path.

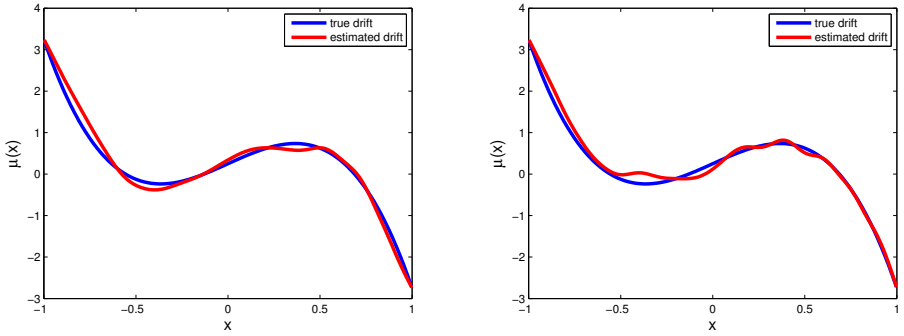

Figure 7: Median reconstructions with KL data fidelity term using 500 (left) and 1000 (right) observations of one path.

We summarize that in our numerical simulations the iteratively regularized Newton method with KL-type or with L^2 data fidelity term works well as nonparametric estimator of the drift coefficient. Reduction of mean and variance of the L^2 error with increasing number of data is observable. The advantage of a KL-type data fidelity term is a significantly smaller mean and variance of the L^2 error compared to the inversion with L^2 data fidelity term.

Modifications of the setup. In addition to the systematic numerical study above we tested the inversion scheme in two modified setups. The first variation of the setting above is to assume that the true values of the drift for $x \geq 1$ and $x \leq -1$ are unknown. Naturally, this makes the estimation of the drift close to the boundary more difficult. In addition, observations in this regions are rare in our examples as can be seen in the limit density of the process. Furthermore, the values of the drift at the boundaries are rather large in absolute values which amplifies the problem. The following plots show typical reconstructions in this case.

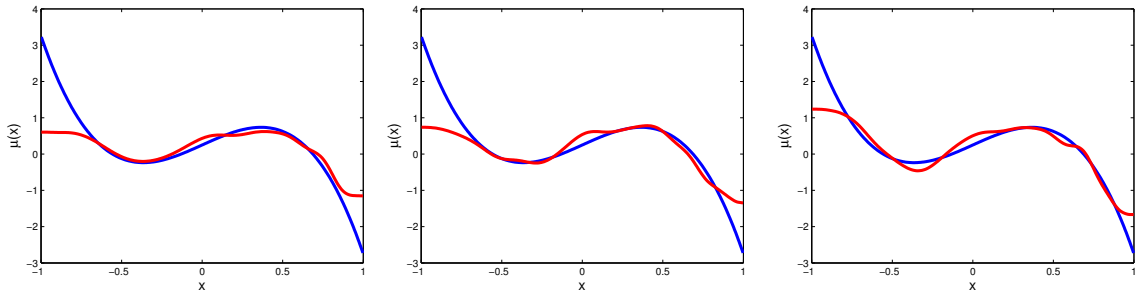

Figure 8: Reconstructions with KL-type data fidelity term using 250 (left), 500 (middle), and 1000 (right) observations of one path. red – reconstruction, blue – true drift

As a second modification of the setup we implemented the first scenario discussed in the introduction. I.e. we simulated a number of paths with common starting point over a smaller period of time instead of simulating one path over a long period of time. Each path is observed at one single time point T. The operator F must be modified for this setting. Instead of solving the elliptic problem (3) we have to solve the parabolic problem (2) in each Newton step. We implemented this by finite elements of order three together with an implicit Euler scheme. The following plots show examples for simulated paths on the time interval $[0, 1]$, the density of the process X_t, and reconstructions of the drift. All paths start at 0 and observations where made at $T = 1$. As above we assumed that the boundary values of the drift are unknown.

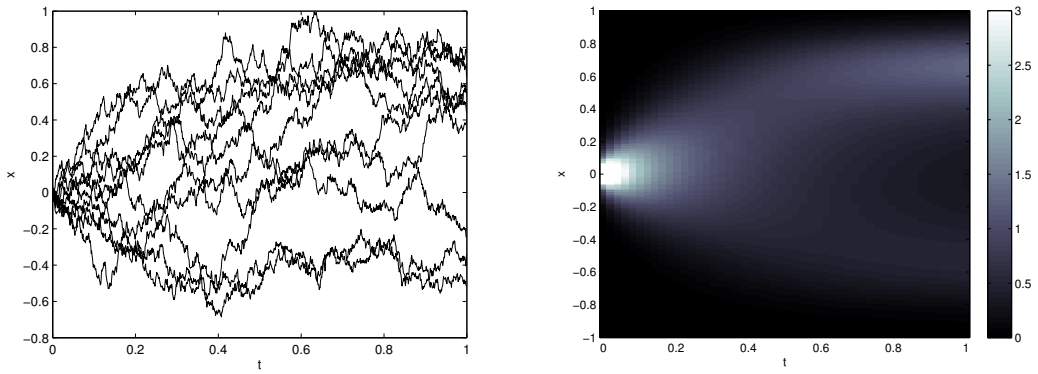

Figure 9: 10 simulated paths (left), density of X_t (right)

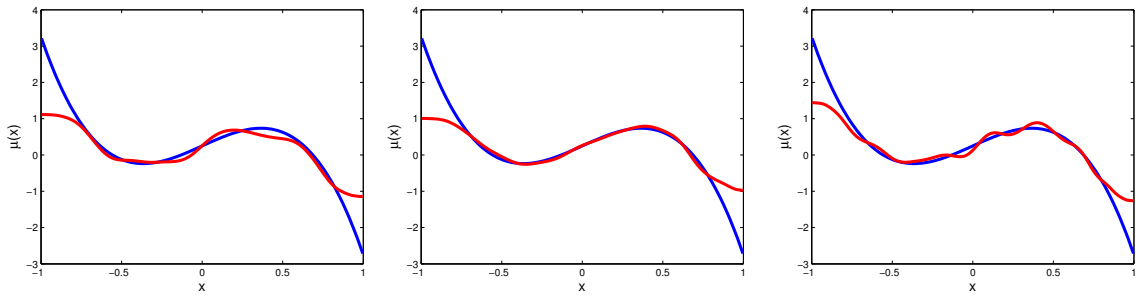

Figure 10: Reconstructions with KL-type data fidelity term using 250 (left), 500 (middle), and 1000 (right) simulated paths. red – reconstruction, blue – true drift.

We can conclude that the algorithm works well in the modified setups. The problems with estimation close to the boundary are typical in nonparametric methods. Furthermore, our test examples are particularly prone to these problem. Nevertheless,

our algorithm produces good results in the interior of the interval.

6. Conclusions

We presented general convergence rate results for estimating parameters in stochastic differential equations by variational regularization methods using Kullback-Leibler-type data fidelity terms. Such terms naturally appear as negative log-likelihood functionals if the observations of paths are described by independent identically distributed random variables. An advantage of this approach is its flexibility. For example, it can also be used to estimate the volatility, initial conditions or coefficients in boundary conditions, and it can handle observations only in part of the domain (see [1]), observations of many paths, and equidistant high frequency data. However, in each situation the conditions of our convergence theorems have to be checked, which may not always be an easy task.

Here we showed that the assumptions of our general convergence theorems are fulfilled for the estimation of the drift in arbitrary space dimensions. A more explicit characterization of the conditions for rates of convergence and comparisons with lower bounds would be desirable, but have to be left for future research.

We demonstrated by Monte-Carlo experiments that Kullback-Leibler-type data fidelity terms yield significantly better results than quadratic data fidelity terms.

Acknowledgments

The authors would like to thank Christian Bender and Thomas Schuster for helpful discussions. Financial support by German Research Foundation DFG through the German-Swiss Research Group FOR 916 is gratefully acknowledged.

References

[1] S. Albeverio, P. Blanchard, S. Kusuoka, and L. Streit. An inverse problem for stochastic differential equations. *Journal of Statistical Physics*, 57:347–356, 1989.

[2] F. Black and M. S. Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654, 1973.

[3] J. M. Borwein and A. S. Lewis. Convergence of best entropy estimates. *SIAM J. Optim.*, 1(2):191–205, 1991.

[4] M. Burger and S. Osher. Convergence rates of convex variational regularization. *Inverse Problems*, 20(5):1411–1421, 2004.

[5] A. D. Cezaro and J. P. Zubelli. The tangential cone condition for the iterative calibration of local volatility surfaces. *IMA J. Applied Math.*, 2013.

[6] F. Comte, V. Genon-Catalot, and Y. Rozenholc. Penalized nonparametric mean square estimation of the coefficients of diffusion processes. *Bernoulli*, 13(2):514–543, 2007.

[7] S. Crépey. Calibration of the local volatility in a generalized Black-Scholes model using Tikhonov regularization. *SIAM J. Math. Anal.*, 34(5):1183–1206 (electronic), 2003.

[8] S. Crépey. Calibration of the local volatility in a trinomial tree using Tikhonov regularization. *Inverse Problems*, 19(1):91–127, 2003.

[9] A. De Cezaro, O. Scherzer, and J. P. Zubelli. Convex regularization of local volatility models from option prices: convergence analysis and rates. *Nonlinear Anal.*, 75(4):2398–2415, 2012.

[10] J. Droniou and J.-L. Vázquez. Noncoercive convection-diffusion elliptic problems with Neumann boundary conditions. *Calc. Var. Partial Differential Equations*, 34(4):413–434, 2009.

[11] F. Dunker, J.-P. Florens, T. Hohage, J. Johannes, and E. Mammen. Iterative estimation of solutions to noisy nonlinear operator equations in nonparametric instrumental regression. *Journal of Econometrics*, 178:444–455, 2014.

[12] H. Egger and H. W. Engl. Tikhonov regularization applied to the inverse problem of option pricing: convergence analysis and rates. *Inverse Problems*, 21(3):1027–1045, 2005.

[13] P. P. B. Eggermont. Maximum entropy regularization for Fredholm integral equations of the first kind. *SIAM J. Math. Anal.*, 24:1557–1576, 1993.

[14] J. Flemming. *Generalized Tikhonov regularization and modern convergence rate theory in Banach spaces*. Shaker Verlag, Aachen, 2012.

[15] D. Gilbarg and N. S. Trudinger. *Elliptic Partial Differential Equations of Second Order*. Springer, 1977.

[16] E. Gobet, M. Hoffmann, and M. Reiß. Nonparametric estimation of scalar diffusions based on low frequency data. *Ann. Statist.*, 32(5):2223–2253, 2004.

[17] M. Grasmair. Generalized Bregman distances and convergence rates for non-convex regularization methods. *Inverse Problems*, 26:115014 (16pp), 2010.

[18] M. Hanke, A. Neubauer, and O. Scherzer. A convergence analysis of the Landweber iteration for nonlinear ill-posed problems. *Numer. Math.*, 72:21–37, 1995.

[19] M. Hoffmann. Adaptive estimation in diffusion processes. *Stochastic Process. Appl.*, 79(1):135–163, 1999.

[20] B. Hofmann, B. Kaltenbacher, C. Pöschl, and O. Scherzer. A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators. *Inverse Problems*, 23(3):987–1010, 2007.

[21] T. Hohage and F. Werner. Iteratively regularized Newton-type methods for general data misfit functionals and applications to Poisson data. *Numer. Math.*, 123(4):745–779, 2013.

[22] A. Hurn, J. Jeisman, and K. Lindsay. Teaching an old dog new tricks: Improved estimation of the parameters of stochastic differential equations by numerical solution of the Fokker-Planck equation. NCER Working Paper Series 9, National Centre for Econometric Research, Feb. 2007.

[23] B. Kaltenbacher and B. Hofmann. Convergence rates for the iteratively regularized Gauss-Newton method in Banach spaces. *Inverse Problems*, 26(3):035007, 21, 2010.

[24] Y. A. Kutoyants. *Statistical inference for ergodic diffusion processes*. Springer Series in Statistics. Springer-Verlag London Ltd., London, 2004.

[25] P. Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Prob.*, 28(2):863–884, 2000.

[26] P. Mathé and B. Hofmann. How general are general source conditions? *Inverse Problems*, 24(1):015009, 5, 2008.

[27] O. Papaspiliopoulos, Y. Pokern, G. O. Roberts, and A. M. Stuart. Nonparametric estimation of diffusions: a differential equations approach. *Biometrika*, 99(3):511–531, 2012.

[28] Y. Pokern, A. M. Stuart, and J. H. van Zanten. Posterior consistency via precision operators for Bayesian nonparametric drift estimation in SDEs. *Stochastic Processes and their Applications*, 123(2):603 – 628, 2013.

[29] E. Resmerita and R. S. Anderssen. Joint additive Kullback-Leibler residual minimization and regularization for linear inverse problems. *Math. Methods Appl. Sci.*, 30(13):1527–1544, 2007.

[30] P. Reynaud-Bouret. Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. *Probab. Theory Related Fields*, 126(1):103–153, 2003.

[31] H. Risken. *The Fokker-Planck equation*, volume 18 of *Springer Series in Synergetics*. Springer-Verlag, Berlin, second edition, 1989. Methods of solution and applications.

[32] O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier, and F. Lenzen. *Variational methods in imaging*, volume 167 of *Applied Mathematical Sciences*. Springer, New York, 2009.

[33] E. Schmisser. Penalized nonparametric drift estimation for a multidimensional diffusion process.

Statistics, 47(1):61–84, 2013.

[34] Z. Schuss. *Theory and applications of stochastic processes*, volume 170 of *Applied Mathematical Sciences*. Springer, New York, 2010. An analytical approach.

[35] T. Schuster, B. Kaltenbacher, B. Hofmann, and K. Kazimierski. *Regularization Methods in Banach Spaces*. Radon Series on Computational and Applied Mathematics. deGruyter, Berlin, 2012.

[36] A. Singer, Z. Schuss, A. Osipov, and D. Holcman. Partially reflected diffusion. *SIAM J. Appl. Math.*, 68(3):844–868, 2007/08.

[37] V. G. Spokoiny. Adaptive drift estimation for nonparametric diffusion model. *Ann. Statist.*, 28(3):815–836, 2000.

[38] M. Talagrand. New concentration inequalities in product spaces. *Invent. Math.*, 126(3):505–563, 1996.

[39] F. Werner and T. Hohage. Convergence rates in expectation for Tikhonov-type regularization of inverse problems with Poisson data. *Inverse Problems*, 28(10):104004, 15, 2012.