A Strongly Convergent Numerical Scheme from Ensemble Kalman Inversion

Dirk Blömker, Claudia Schillings, Philipp Wacker

Angaben zur Veröffentlichung / Publication details:

Blömker, Dirk, Claudia Schillings, and Philipp Wacker. 2018. "A Strongly Convergent Numerical Scheme from Ensemble Kalman Inversion." *SIAM Journal on Numerical Analysis* 56 (4): 2537–62. https://doi.org/10.1137/17m1132367.

SIAM J. NUMER. ANAL.
Vol. 56, No. 4, pp. 2537–2562

A STRONGLY CONVERGENT NUMERICAL SCHEME FROM ENSEMBLE KALMAN INVERSION*

DIRK BLÖMKER†, CLAUDIA SCHILLINGS‡, AND PHILIPP WACKER§

Abstract. The Ensemble Kalman methodology in an inverse problems setting can be viewed as an iterative scheme, which is a weakly tamed discretization scheme for a certain stochastic differential equation (SDE). Assuming a suitable approximation result, dynamical properties of the SDE can be rigorously pulled back via the discrete scheme to the original Ensemble Kalman inversion. The results of this paper make a step towards closing the gap of the missing approximation result by proving a strong convergence result in a simplified model of a scalar SDE. We focus here on a toy model with similar properties to the one arising in the context of an Ensemble Kalman filter. The proposed model can be interpreted as a single particle filter for a linear map and thus forms the basis for further analysis. The difficulty in the analysis arises from the formally derived limiting SDE with nonglobally Lipschitz continuous nonlinearities both in the drift and in the diffusion. Here the standard Euler–Maruyama scheme might fail to provide a strongly convergent numerical scheme and taming is necessary. In contrast to the strong taming usually used, the method presented here provides a weaker form of taming. We present a strong convergence analysis by first proving convergence on a domain of high probability by using a cutoff or localization, which then leads, combined with bounds on moments for both the SDE and the numerical scheme, by a bootstrapping argument to strong convergence.

Key words. Ensemble Kalman filter, stochastic differential equations, Bayesian inverse problems

AMS subject classifications. 60H10, 65N75, 62F15

DOI. 10.1137/17M1132367

1. Introduction.

We consider throughout the whole paper the following simplified model given by the one-dimensional SDE

$$\mathrm{d}u = -u^3\mathrm{d}t + u^2\mathrm{d}W\ . \tag{1}$$

For this equation, we will prove that it is strongly approximated for $t \in [0,T]$ by the numerical scheme arising in the context of the Ensemble Kalman filter (EnKF) for inverse problems

$$u_{n+1} = u_n - h\cdot\frac{u_n^3}{1+h\cdot u_n^2} + \frac{u_n^2}{1+h\cdot u_n^2}\cdot\Delta W_{n+1} \tag{2}$$

with $n = 0,\dots,T/h$, where u_n is supposed to correspond to $u(t_n)$ with $t_n = n\cdot h$ in the strong sense at a single fixed time, and in a uniform strong sense, where moments of the uniform error in time are taken.

We are also able to extract a rate of convergence for both the strong and the uniform strong error. The exact statement of the rate can be found in our main result Theorem 8. In section 2 we will comment in detail on how this approximation scheme arises as a toy model in the theory of EnKF.

*Received by the editors May 30, 2017; accepted for publication (in revised form) June 12, 2018; published electronically August 7, 2018.
http://www.siam.org/journals/sinum/56-4/M113236.html
†Institute für Mathematics, Universität Augsburg, Augsburg 86135, Germany (dirk.bloemker@math.uni-augsburg.de).
‡Institute of Mathematics, Universität Mannheim, Mannheim 68131, Germany (c.schillings@uni-mannheim.de).
§Universität Erlangen, Erlangen 91058, Germany (phkwacker@gmail.com).

Note that a convergence result like the one above does not hold if we exchange the EnKF numerical scheme u_n by the naive Euler–Maruyama discretization scheme

$$\tilde{u}_{n+1} = \tilde{u}_n - h \cdot \tilde{u}_n^3 + \tilde{u}_n^2 \cdot \Delta W_{n+1} \tag{3}$$

as was shown by Hutzenthaler, Jentzen, and Kloeden [12]. In this setting we only obtain pathwise convergence but it can even be proven that the Euler–Maruyama discretization diverges strongly. This is due to an exponentially rare (in h) family of events (paths of $\tilde{u}_n$) which grows biexponentially strongly (in h) while paths of u stay near the origin and have p-moments at least up to $p < 3$. Then, any moment of the difference between u and $\tilde{u}_n$ will explode for $h \to 0$.

There has been significant progress in the field of strongly convergent numerical schemes for SDEs with nonglobal Lipschitz-continuous nonlinearities. Standard references for numerical methods for SDEs, like [17] and [25, 22], show strong convergence of the Euler–Maruyama method only for globally Lipschitz-continuous drift and diffusion terms. Higham, Mao, and Stuart [10] proved a conditional result about strong convergence of the Euler–Maruyama discretization for nonglobally Lipschitz SDEs given that moments of both solutions and the discretization stay bounded. This means that the question of strong convergence was replaced by the question of whether moments of the numerical scheme stay bounded, but Hutzenthaler, Jentzen, and Kloeden [12] answered the latter to the negative, even proving that moments of the Euler–Maruyama scheme always explode in finite time if either drift or diffusion term are not globally Lipschitz. Instead, they proposed a slight modification of the Euler–Maruyama (EM) method, the so-called tamed EM method (and implicit equivalents) in [13, 11].

The numerical scheme (2) arising in the EnKF analysis bears resemblance to the "tamed" methods used throughout the literature. In the case only the drift is nonglobally Lipschitz a similar idea to (2) was used already in [13], but there the drift-tamed nonlinearity is strictly bounded by one, while in EnKF it is still allowed to grow linearly. In increment-tamed schemes (see Hutzenthaler and Jentzen [11]) even the whole increment on the right-hand side of (3) is cut to have modulus of at most one.

Thus the differences to other tamed schemes are still too large to carry out their analysis analogously. In particular, we cannot choose an arbitrary scheme with the right taming for a given SDE, but we face a specific iterative scheme directly given by the EnKF. We emphasize that even the increment-tamed scheme cannot provably approximate the SDE

$$\mathrm{d}u = -u^3 \mathrm{d}t + u^2 \mathrm{d}W\,,$$

which we focus on in this paper. This is easily checked by Corollary 3.18 in [11]. For $\mu(x) = -x^3$ and $\sigma(x) = x^2$, we cannot find a positive number q such that the conditions for strong convergence are met.

Remark 1. We want to point out again that the goal of this manuscript is not to propose a new class of numerical discretization schemes for SDEs, although further research might reveal whether this is possible. Rather, we are initially "stuck" with a iterative scheme which has a formal continuous form (the SDE). We will show that the formal continuous limit is rigorous using the strong convergence notion. We believe that our results should be extendable in some situations to the case of multivariate $u \in \mathbb{R}^d$,

$$\mathrm{d}u = f(u)\mathrm{d}t + \sigma(u)\mathrm{d}B,$$

where B is a Brownian motion in $\mathbb{R}^d$ and f and σ are nonlinearities which are only locally Lipschitz, although we haven't checked this properly. Especially, the correct taming of drift and diffusion that is weaker than the usual taming in the literature, but still preserves all the properties that allow for moment bounds for the numerical scheme and the SDE, like one-sided Lipschitz properties, is open.

Our method of proof in showing strong convergence is similar in spirit to the method deployed in the first chapter of [11], although in a continuous setting instead for a discrete process. We show strong convergence on a domain of high probability (governed by a stopping time in our case instead of restricting to events of high probability in the method of Hutzenthaler and Jentzen) and combined with bounds on moments for both the SDE and the numerical scheme, a bootstrapping argument facilitated by Hölder's inequality proves convergence on the whole domain. Our method also bears resemblance to the trick employed by Higham, Mao, and Stuart [10]. The key difference is that we need to avoid using Gronwall's lemma in its integrated form. Rather we exploit the dissipative structure of the SDE explicitly and also obtain, in addition to the uniform bounds, bounds on time-averaged quantities. This helps us to eliminate the need for high-order moments. This is reflected in (23): The integral term originates as a negative term on the right-hand side as a result of the stability of the SDE. The key results are Lemmas 5 and 7.

The many-particle setting makes the analysis very complicated, so we will work with the simplified model (1) in one dimension which still shows the main characteristics of the SDE and the numerical scheme in the EnKF setting. For this model we will show that the numerical scheme which arises from the EnKF iteration is in fact a strongly converging discretization.

This is a surprising fact given that for the specific equation $\mathrm{d}u = -u^3\mathrm{d}t + u^2\mathrm{d}W$, both the Euler–Maruyama discretization and the increment-tamed scheme proposed by Hutzenthaler, Jentzen, and Kloeden, do not share this property. We also note that nonstrongly converging numerical schemes are really only an issue for SDEs with nonglobally Lipschitz drift and/or diffusion terms, which is the case here.

In order to prove strong convergence we consider an intermediate model which is the regularization of (1):

$$\mathrm{d}v = -\frac{v^3}{1+\varepsilon\cdot v^2}\mathrm{d}t + \frac{v^2}{1+\varepsilon\cdot v^2}\mathrm{d}W. \tag{4}$$

Its Euler–Maruyama discretization is then

$$\begin{aligned} v_{n+1} &= v_n + h\cdot\frac{-v_n^3}{1+\varepsilon\cdot v_n^2} + \Delta W_n\cdot\frac{v_n^2}{1+\varepsilon\cdot v_n^2} \\ &= v_n + \int_{t_n}^{t_{n+1}}\frac{-v_n^3}{1+\varepsilon\cdot v_n^2}\mathrm{d}t + \int_{t_n}^{t_{n+1}}\frac{v_n^2}{1+\varepsilon\cdot v_n^2}\mathrm{d}W \end{aligned} \tag{5}$$

and from standard theory we know that the discretization converges strongly to the SDE. The main point is that for $h=\varepsilon$ this numerical scheme is identical to the one we are interested in, i.e., u_n. We will set $\varepsilon = h$ in what follows so we are actually considering a sequence of numerical schemes for which we want to prove that they approximate a sequence of SDEs (the family of paths of v parameterized by ε). We set $f(v) = -\frac{v^3}{1+\varepsilon\cdot v^2}$ and $\sigma(v) = \frac{v^2}{1+\varepsilon\cdot v^2}$ and define the interpolation of v_n as

$$\bar{v}(t) = v_n + \int_{t_n}^{t} f(v_n)\mathrm{d}s + \int_{t_n}^{t}\sigma(v_n)\mathrm{d}W. \tag{6}$$

Note that this is not actually an interpolation in the normal sense because it has "wiggly" paths from the stochastic integral. We will show that

1. v converges strongly to u. More precisely we prove

$$\sup_{t\in[0,T]} \lim_{\varepsilon\to 0} \mathbb{E}|u(t)-v(t)|^\alpha = 0 \quad \text{for any } 0<\alpha<3 \tag{7}$$

and

$$\lim_{\varepsilon\to 0} \mathbb{E} \sup_{t\in[0,T]} |u(t)-v(t)|^\beta = 0 \quad \text{for any } 0<\beta<3/2. \tag{8}$$

We also obtain rates of this convergence; see Lemma 10;

2. $\bar v$ and v become arbitrarily close in the strong sense. More precisely,

$$\lim_{h=\varepsilon\to 0} \sup_{t\in[0,T]} \mathbb{E}|v(t)-\bar v(t)|^\alpha = 0 \quad \text{for any } 0<\alpha<2 \tag{9}$$

and also

$$\lim_{h=\varepsilon\to 0} \mathbb{E} \sup_{t\in[0,T]} |v(t)-\bar v(t)|^\beta = 0 \quad \text{for all } 0<\beta<1, \tag{10}$$

again with rates as shown in Lemma 11.

This means that paths of the regularization v are arbitrarily well-approximated (strongly) by the numerical scheme. By the triangle inequality we obtain our main result, keeping in mind that $\bar v(t_n) = u_n$ for $\varepsilon = h$.

Key for our analysis are a priori moment bounds on all quantities involved, i.e., $\sup_{t\in[0,T]} \mathbb{E}|w|^\kappa < C$ and $\mathbb{E}\sup_{t\in[0,T]} |w|^\xi$ for $w \in \{u, v, \bar v\}$ and $\kappa, \xi > 0$ chosen suitably (cf. Lemma 5).

Remark 2. Note that we are able to show

$$\lim_{\varepsilon\to 0} \mathbb{E} \sup_{t\in[0,T]} |u(t)-v(t)|^\beta = 0 \quad \text{for } \beta \in (0,3/2)\,,$$

but a similar result for the difference $|v(t)-\bar v(t)|$ holds only for $\beta \in (0,1)$. The difference $v-\bar v$ is a similar bottleneck for the $\sup\mathbb{E}$-case, where we only achieve $0<\alpha<2$ instead of $0<\alpha<3$ for the difference $u-v$. In the case "$\mathbb{E}\sup$", we expect that the result can be shown to hold for $\beta \in (0,3/2)$. However, the presented analysis does not include the uniform boundedness (in ε) of

$$\mathbb{E} \sup_{t\in[0,T]} |\bar v(t)|^\beta < C \quad \text{or, equivalently,} \quad \mathbb{E} \sup_{n=1,\dots,T/h} |v_n|^\beta < C$$

for $\beta \in (1,3/2)$.

Remark 3. We took care to formulate "$\bar v$ and v become arbitrarily close." We will from now on sometimes use the intuitive notion of "$\bar v$ converges to v" for simplicity, although this is not rigorously defined as v varies with $\varepsilon = h$ as well.

2. Motivation from the EnKF and its continuum analysis. The EnKF was first introduced by Evensen [6]. Since then, it has been widely used in the context of data assimilation and recently also of inverse problems; see, e.g., [19] for more details. Despite its great success story in various areas of application, the analysis of the EnKF is still in its infancy. Well-posedness and convergence results in the case of a fully observed system are discussed in [15, 32, 31, 16] and in [3] in the data assimilation context. The analysis of the large ensemble size limit can be found in [23, 18, 9]. The generalization to inverse problems is presented in [14]. Ernst, Sprungk, and Starkloff [5] showed that the EnKF is not consistent with the Bayesian

perspective in the nonlinear setting, but can be interpreted as a point estimator of the unknown parameters. Most of the analysis is confined to the large ensemble size limit, i.e., the case where the number of particles in the ensemble goes to infinity. However, the EnKF is usually used with a rather small ensemble size in practice. This finite-ensemble scenario was analyzed by Kelly et al. [15, 16, 32, 31]. Stannat [29] analyzed the behavior of the EnKF for the filtering problem of a partially observed solution of an SDE. This setting is different from ours in that the scaling of the data (necessarily) is completely different: In [29], more frequent observations lead to more available data for filtering. In our scenario, the data are artifically augmented to "live" on a time axis and are used repeatedly (and redundantly). To compensate for this redundancy, an artificial noise term is added to the data (cf. (16)). For a more complete discussion of this, see [28, 14].

The authors presented an analysis of the long-time behavior and ergodicity of the EnKF with arbitrary ensemble size establishing time uniform bounds to control the filter divergence and ensuring in addition the existence of an invariant measure. In the linear Gaussian setting, Del Moral and Tugaut [4] investigated the convergence of the ensemble Kalman–Bucy filter and provided time uniform error estimates for the empirical mean and covariance. Schillings and Stuart [28] (and prior to that, Bergemann and Reich in [2]) conducted a continuum limit analysis of the EnKF methodology, shedding light on the properties of the EnKF in the small ensemble size setting. They propose to interpret the Ensemble Kalman method as a numerical discretization scheme of a continuous process driven by a (stochastic) differential equation by formally conducting the continuum limit. In [1], the behavior of deterministic EnKF limits has been studied to develop stable and efficient integration methods. The purpose of this paper is to provide a step towards making this continuum limit rigorous (in a probabilistically strong meaning). In the following, we shortly introduce the EnKF for inverse problems and refer to [14, 28] for a more detailed exposition.

The use of the EnKF in Bayesian inverse problems. The inverse problem is defined as follows: The goal of computation is to recover the unknown parameters $u \in X$ from a noisy observation $y \in Y$, where

$$y = G(u) + \eta \tag{11}$$

with the so-called forward response map $G : X \to Y$. We denote by X and Y separable Hilbert spaces and by the random variable η the observational noise.

Inverse problems arise in a multitude of applications, for example, in geosciences, medical imaging, reservoir modeling, astronomy, and signal processing. A probabilistic approach has been undertaken first by Franklin [8], Mandelbaum [24], and others, and a formulation of inverse problems in the context of Bayesian probability theory was done by Fitzpatrick [7]. More recent literature about inverse problems in the Bayesian setting can be found in [26, 30].

In the following, we assume that the number of observations is finite, i.e., $Y = \mathbb{R}^K$ for some $K \in \mathbb{N}$, whereas the unknown is a distributed quantity, which is a typical setting for many applications mentioned above. Furthermore, we assume that the observational noise is Gaussian, i.e., $\eta \sim N(0, \Gamma)$ with a symmetric, positive definite matrix $\Gamma \in \mathbb{R}^{K \times K}$. We consider the least squares "error" (or model-data misfit) functional

$$\Phi(u; y) = \frac{1}{2} \|\Gamma^{-1/2}(y - G(u))\|_Y^2, \tag{12}$$

where Γ normalizes the model-data misfit and is normally chosen as the covariance

operator of the noise. Plain infimization of this cost functional is not feasible due to the ill-posedness of the problem. For a given prior $u \sim \mu_0$, we derive the posterior measure $u|y \sim \mu$, where

$$\mu(du) = \frac{1}{Z} \exp(-\Phi(u; y))\mu_0(du). \tag{13}$$

Sequential version of the EnKF. By interpolating the step $\mu_0 \square \mu$ as $\mu_n(du) \propto \exp(-nh\Phi(u; y))\mu_0(du)$ with a step size $h = 1/N$ and $n \in \{1, \dots, N\}$, a finite sequence of measures is obtained with initial measure μ_0 (the prior) and final measure $\mu_N = \mu$. Note that, by introducing the artificial time step h, we recast the problem into a data assimilation problem by using the observational data sequentially. We account for the repeated use of the data by amplifying the noise covariance $1/h \cdot \Gamma$. It is important to keep in mind that the time we refer to in the following is the artificial time introduced by the sequence of intermediate measures μ_n.

The EnKF approximates each of those measures μ_n by a sum of Dirac masses (their centers form the ensemble of particles giving its name to the EnKF):

$$\mu_n \approx \nu_n = \frac{1}{J} \sum_{j=1}^{J} \delta_{u_n^{(j)}} . \tag{14}$$

Then the problem of mapping $\mu_n \mapsto \mu_{n+1}$ reduces to mapping the particles $u_n^{(j)}$ in "time" step $n \mapsto n + 1$.

The EnKF chooses a linear transformation of the particles such that the mean and covariance are consistent updates with the Kalman filter. Various variants of the EnKF realizing different transformations of the particles exist; see, e.g., [27] for more details. We focus here on the oldest transformation, the EnKF with perturbed observations leading to the iteration of the form

$$u_{n+1}^{(j)} = u_n^{(j)} + C^{up}(u_n)\left[C^{pp}(u_n) + h^{-1}\Gamma\right]^{-1}(y_{n+1}^{(j)} - G(u_n^{(j)})), \quad j = 1, \dots, J, \tag{15}$$

for each particle $u^{(j)}$, $j - 1, \dots, J,\ J \in \mathbb{N}$, in the nth iteration, where

$$y_{n+1}^{(j)} = y + \xi_{n+1}^{(j)} \tag{16}$$

with $\xi_{n+1}^{(j)} \sim N(0, h^{-1}\Gamma)$, is the artificially perturbed observation and the operators[1] C^{pp} and C^{up} are defined for some $u = (u^{(j)})_{j=1}^{J}$ with each $u^{(j)} \in X$, as

$$C^{pp}(u) = \frac{1}{J} \sum_{j=1}^{J} (G(u^{(j)}) - \overline{G}) \otimes (G(u^{(j)}) - \overline{G}), \tag{17}$$

$$C^{up}(u) = \frac{1}{J} \sum_{j=1}^{J} (u^{(j)} - \overline{u}) \otimes (G(u^{(j)}) - \overline{G}), \tag{18}$$

$$\overline{u} = \frac{1}{J} \sum_{j=1}^{J} u^{(j)}, \quad \overline{G} = \frac{1}{J} \sum_{j=1}^{J} G(u^{(j)}). \tag{19}$$

[1]The "p" in the superscript is notation from [14] where quantities in parameter space are called u and objects in observation space are called p.

Equivalently, we can rewrite the iteration in the form

$$u_{n+1}^{(j)} = u_n^{(j)} + hC^{up}(u_n)\left[hC^{pp}(u_n) + \Gamma\right]^{-1}\left(y - G(u_n^{(j)})\right) + \sqrt{h}C^{up}(u_n)\left[hC^{pp}(u_n) + \Gamma\right]^{-1}\Gamma^{1/2}\zeta_{n+1}^{(j)}, \quad j = 1, \dots, J, \tag{20}$$

with standard Gaussians $\zeta_{n+1}^{(j)} \sim N(0,1)$ independent and identically distributed. Formally, in the limit $h \to 0$ one can see (cf. [28]) that this is some nonstandard approximation scheme for the following system of SDEs:

$$\mathrm{d}u^{(j)} = C^{up}(u)\Gamma^{-1}(y - G(u^{(j)}))\mathrm{d}t + C^{up}(u)\Gamma^{-1/2}\mathrm{d}W^{(j)}, \quad j = 1, \dots, J. \tag{21}$$

Due to the derivation, we are particularly interested in the dynamics on $t \in [0,1]$ (with $t = 0$ corresponding to the prior measure and $t = 1$ denoting the posterior), but we will work with a general bounded time domain $t \in [0,T]$ from here on. A well-studied example for an approximation scheme is the Euler–Maruyama discretization, which in our example is of the form

$$\tilde{u}_{n+1}^{(j)} = \tilde{u}_n^{(j)} + hC^{up}(\tilde{u}_n)\Gamma^{-1}(y - G(\tilde{u}_n^{(j)})) + \sqrt{h}C^{up}(\tilde{u}_n)\Gamma^{-1/2}\zeta_{n+1}^{(j)}. \tag{22}$$

In contrast to the usual direction of thought when working with SDEs and discretizations of their dynamics (i.e., start with a dynamics and construct a discrete numerical scheme with "good" approximation properties), we are traveling in the opposite direction: The EnKF methodology yields a numerical iteration which looks like some approximation scheme for an SDE, and it is interesting to understand whether this numerical scheme is in fact well-approximated by the SDE.

Schillings and Stuart [28] formally derived the limiting SDE and focused on the analysis of the dynamical behavior of the particles establishing convergence results for the fixed ensemble size limit. Motivation for this manuscript is the question of what is the nature of this continuum limit and what is the convergence behavior of the numerical scheme going to the limit. This problem is not only of interest for the analysis of the EnKF, but also possibly opens up the perspective of using the EnKF as a numerical discretization scheme for SDEs with nonglobally Lipschitz-continuous nonlinearities.

Note at this point that the Euler–Maruyama discretization is not a suitable approximation of the limiting SDE. Although for a given realization it converges pathwise, it was shown [12] to diverge in the strong sense (i.e., moments of the difference between numerical scheme and the solution of the SDE will explode in the limit $h \to 0$). This is mainly due to the fact that the SDE does not have a globally Lipschitz continuous drift and diffusion term. At this point, we only manage to carry out the necessary analysis for an extremely simplified toy model exhibiting (arguably) a similar structure to the SDE (21) arising in the EnKF context. The simplified model is a good starting point for the analysis and will form the basis for the more general setting, which will be the subject of future work. Due to the subspace property of the EnKF, the dynamics can be described in the finite-dimensional coordinate system of the particles, thus the analysis of the finite-dimensional case can be straightforwardly generalized to the EnKF setting. However, the interaction of the particles and the nonlinearity of the forward response operator results in more complicated nonlinearities of the drift and diffusion term, so that we anticipate the need of additional assumptions on their stability in the more general setting.

Remark 4. Note that we when we say "continuum limit," we mean the following: We artificially augment the state space with an artificial (discrete) time and consider the limit for continuous artificial time. This is not to be confused with either the continuous-time limit in the data assimilation setting (where time is explicitly present in the data) and also not with the mean-field approach where the limit $J \to \infty$ (a continuum of particles) is considered. For the latter, see, for example, [21, 23, 20]. We also do not comment on a "fully interacting infinite-dimensional PDE system" and "mean-field limiting system" which also treat the case of $J \to \infty$, while we keep J fixed and possibly even quite small.

Constructing a toy model. As the full EnKF iteration (20) is too difficult to analyze, we make a series of simplifying steps in order to reduce the complexity of the problem, thereby deriving a toy model on which to test the tools we plan on eventually applying to the full EnKF setting.

We consider a linear map G and set the dimension of the parameter space to one, i.e., $d = 1$, and the number of particles to two, i.e., $J = 2$. Note that in this case

$$u^{(1)} - \bar{u} = -(u^{(2)} - \bar{u}) =: q.$$

Now the SDE for the EnKF is

$$\mathrm{d}u^{(j)} = \frac{1}{2} \sum_{k=1}^{2} (u^{(k)} - \bar{u}) \cdot (Gu^{(k)} - G\bar{u}) \cdot \Big((y - Gu^{(j)})\mathrm{d}t + \Gamma^{1/2}\mathrm{d}W^j \Big)$$

for $j = 1, 2$. The SDE for the particle mean is then

$$\mathrm{d}\bar{u} = \frac{1}{2} \sum_{k=1}^{2} (u^{(k)} - \bar{u}) \cdot (Gu^{(k)} - G\bar{u}) \cdot \Big((y - G\bar{u})\mathrm{d}t + \Gamma^{1/2}\mathrm{d}\bar{W} \Big)$$

and the equation for the particles distance to the mean is

$$\mathrm{d}q = -q(Gq)^2\mathrm{d}t + q \cdot Gq\mathrm{d}B$$

(note that $(u^{(k)} - \bar{u}) \cdot (Gu^{(k)} - G\bar{u}) = q \cdot Gq$ for $k = 1, 2$) by setting $B = W^{(1)} - \bar{W} = -(W^{(2)} - \bar{W})$. For the identity map $G(u) = u$ we get exactly our toy model (1). This means that our toy model is a simplification of the dynamics of the particles' distance to their joint mean. To recover the dynamics of the particles, we can use

$$\mathrm{d}\bar{u} = q \cdot Gq \cdot (y - G\bar{u})\mathrm{d}t + q \cdot Gq \cdot \Gamma^{1/2}\mathrm{d}\bar{W},$$

which we do not consider here in our further analysis.

To see how the discretization scheme (2) arises, we start from (20) and use the same simplifying assumptions (i.e., G the identity map, one-dimensional parameter space, two particles, $\Gamma = \gamma = 1$). This yields

$$\begin{aligned} u_{n+1}^{(1)} &= u_n^{(1)} + \frac{\frac{1}{2}\sum_{k=1}^{2}(Gu_n^{(1)} - G\bar{u}_n) \cdot (u^{(1)} - \bar{u})}{h\frac{1}{2}\sum_{k=1}^{2}(Gu_n^{(1)} - G\bar{u}_n)^2 + \gamma^2} \cdot \Big(h \cdot (y - Gu_n^{(1)}) + \sqrt{h} \cdot \xi_{n+1}^{(1)} \Big) \\ &= u_n^{(1)} + \frac{q_n^2}{hq_n^2 + 1} \cdot \Big(h \cdot (y - Gu_n^{(1)}) + \sqrt{h} \cdot \xi_{n+1}^{(1)} \Big), \\ \bar{u}_{n+1} &= \bar{u}_n + \frac{q_n^2}{hq_n^2 + 1} \cdot \Big(h \cdot (y - \bar{G}u_n) + \sqrt{h} \cdot \bar{\xi}_{n+1} \Big), \end{aligned}$$

and thus

$$q_{n+1} = q_n + \frac{q_n^2}{hq_n^2+1} \cdot \Big[h \cdot (-q) + \sqrt{h} \cdot (\xi_{n+1}^{(1)} - \bar{\xi}_{n+1}) \Big] ,$$

which is exactly (2).

3. Organization of the paper. The remaining part of the paper is organized as follows. We present in section 4 the main results of the paper, the strong convergence of the EnKF scheme to the SDE. The proof essentially splits into two parts: the analysis of the error between the solution of the SDE and the solution of the Lipschitz regularized SDE, which is presented in section 5 and the analysis of the error between the solution of the Lipschitz regularized SDE and its Euler–Maruyama discretization presented in section 6. The analysis relies on a priori estimates on the moments of the processes, which are derived in section 7.

4. Statement of the main results. We will present in the following the main result, the strong convergence of the EnKF numerical discretization scheme to the SDE. The proof of the main statement essentially relies on two lemmas on bounding differences (and extracting rates) between stochastic processes up to a deterministic time T in the case that the difference is small up to a stopping time and the moments of the stochastic processes are small. This approach is based on ideas from Higham, Mao, and Stuart [10].

LEMMA 5. *Let $v_i, i = 1, 2$, be two stochastic processes with continuous paths and let $e = v_1 - v_2$ denote the difference between both processes. Further, we denote by τ the stopping time defined for $T > 0$ and $\gamma_h > 0$ by*

$$\tau = T \wedge \inf\{t > 0 \ : \ |v_1(t)| > \gamma_h^{-1} \text{ or } |v_2(t)| > \gamma_h^{-1}\} .$$

Assume that for some $C_\star > 0$ and $p > 0$ it holds

$$\mathbb{E} \sup_{[0,T]} |v_1|^p \le C_\star \quad \text{and} \quad \mathbb{E} \sup_{[0,T]} |v_2|^p \le C_\star ,$$

and that for some $\delta_h > 0$ it holds

$$\mathbb{E} \sup_{[0,\tau]} |e|^2 \le \delta_h^2.$$

Then for $\eta \in (0, p)$ there exists a constant $K := 2^{(p-\eta)/p} 2^\eta$ such that

$$\mathbb{E} \sup_{[0,T]} |e|^\eta \le K C_\star \gamma_h^{p-\eta} + \delta_h^\eta .$$

Thus, for sufficiently small $\eta > 0$ the term δ_h determines the rate of the strong convergence. If $\delta_h = h^\delta$ and $\gamma_h = h^\gamma$, the effective rate is given by

$$\Big[\mathbb{E} \sup_{[0,T]} |e|^\eta \Big]_{1/\eta} \le \text{Const} \cdot h^{\min\{\gamma(p-\eta)/\eta, \delta\}} .$$

It is easy to check, that the optimal rate is δ if $\eta \le p\gamma/(\gamma + \delta)$.

Proof. For the stopping time, it holds that

$$\mathbb{P}(\tau < T) \le \mathbb{P} \Big[\sup_{[0,T]} |v_1| > \gamma_h^{-1} \text{ or } \sup_{[0,T]} |v_2| > \gamma_h^{-1} \Big] .$$

Thus, Chebychev's inequality yields

$$\mathbb{P}(\tau < T) \leq 2C_\star \gamma_h^p \,.$$

It directly follows that

$$\mathbb{E}\sup_{[0,T]} |e|^p \leq \left(\mathbb{E}\sup_{[0,T]} |v_1|^p\right)^{1/p} + \left(\mathbb{E}\sup_{[0,T]} |v_1|^p\right)^{1/p} \leq 2^p C_\star \,.$$

Now we obtain

$$\begin{aligned}
\mathbb{E}\sup_{[0,T]} |e|^\eta &= \int_{\tau<T} \sup_{[0,T]} |e|^\eta d\mathbb{P} + \int_{\tau=T} \sup_{[0,T]} |e|^\eta d\mathbb{P} \\
&= \int_{\tau<T} \sup_{[0,T]} |e|^\eta d\mathbb{P} + \int_{\tau=T} \sup_{[0,\tau]} |e|^\eta d\mathbb{P} \\
&\leq \left(\int_{\tau<T} d\mathbb{P}\right)^{(p-\eta)/p} \left(\mathbb{E}\sup_{[0,T]} |e|^p d\mathbb{P}\right)^{\eta/p} + \mathbb{E}\sup_{[0,\tau]} |e|^\eta \\
&\leq (2C_\star\gamma_h^p)^{(p-\eta)/p} \left(2^p C_\star\right)^{\eta/p} + \left(\mathbb{E}\sup_{[0,\tau]} |e|^2\right)^{\eta/2} \\
&\leq 2^{(p-\eta)/p} C_\star 2^\eta \cdot \gamma_h^{p-\eta} + \delta_h^\eta \,.
\end{aligned}$$

∎

COROLLARY 6. *Given the assumptions of Lemma 5, $\gamma_h = h^\gamma$, and, additionally, $\delta_h^2 = h^{1-\rho\gamma}$ for some $\rho > 0$, we obtain*

$$\left(\mathbb{E}\sup_{[0,T]} |e|^\eta\right)^{1/\eta} \leq C h^{\frac{1}{2}\cdot\frac{p\ \ \eta}{p\ \ \eta+\eta\rho/2}},$$

in particular, for $\eta \to 0$ we recover a rate of $h^{1/2}$.

Proof. From Lemma 5 we get

$$\mathbb{E}\sup_{[0,T]} |e|^\eta \leq K\gamma_h^{p-\eta} + \delta_h^\eta \leq C \cdot (h^{\gamma(p-\eta)} + h^{\frac{\eta}{2}(1-r\gamma)}).$$

The rate is optimal if $\gamma = \frac{\eta}{2p+\eta(r-2)}$, i.e.,

$$\mathbb{E}\sup_{[0,T]} |e|^\eta \leq C \cdot h^{\frac{\eta}{2}\frac{p\ \ \eta}{p\ \ \eta+\eta r/2}} \,.$$

Taking the ηth root on both sides yields the claim. ∎

LEMMA 7. *Let $v_i, i = 1, 2$, be two stochastic processes with continuous paths and let $e = v_1 - v_2$ denote the difference between both processes. Further, we denote by τ the stopping time defined for $T > 0$ and $\gamma > 0$ by*

$$\tau = T \wedge \inf\{t > 0 \ : \ |v_1(t)| > h^{-\gamma} \text{ or } |v_2(t)| > h^{-\gamma}\} \,.$$

Suppose for some $C_\star > 0$ and $p > 0$

$$\mathbb{E}\sup_{[0,T]} |v_1|^p \leq C_\star \quad \text{and} \quad \mathbb{E}\sup_{[0,T]} |v_2|^p \leq C_\star$$

and for $s > p$

$$\sup_{[0,T]} \mathbb{E}|v_1|^s \leq C_\star \quad \textit{and} \quad \sup_{[0,T]} \mathbb{E}|v_2|^s \leq C_\star \, .$$

Moreover, we assume that

$$\mathbb{E} \sup_{[0,\tau]} |e|^2 \leq h^{2\delta}$$

for some $\delta > 0$. Then,

$$\sup_{t\in[0,T]} \mathbb{E}|e(t)|^q \leq Const \cdot \left[h^{(\delta - \frac{\gamma p}{2})\cdot q} + h^{\gamma p \frac{s-q}{s}} \right]$$

and the effective optimal rate is the minimum of both exponents.

If, in addition, $\delta = \frac{1-\rho\gamma}{2}$ for some $\rho > 0$ such that $\delta > 0$, then

$$\left(\sup_{t\in[0,T]} \mathbb{E}|e(t)|^q \right)^{\frac{1}{q}} \leq C \cdot h^{\frac{p(s-q)}{2p(s-q)+(p+\rho)qs}} \, ,$$

which amounts to an effective rate of $1/2$ for $q \to 0$.

Proof. Similarly to the proof of Lemma 5, we obtain

$$\mathbb{P}(\tau < T) \leq 2C_\star h^{\gamma p}$$

and

$$\mathbb{E} \sup_{[0,T]} |e|^p \leq 2^p C_\star \, .$$

Also, $\{\sup_{t\leq T} |e|^2 \geq \kappa_h^2\} \subset \{\tau < T\} \cup \{\sup_{t\leq\tau} |e|^2 \geq \kappa_h^2\}$, where $\kappa_h = h^\kappa$ with a parameter $\kappa > 0$ yet to be determined. Thus,

$$\mathbb{P}\left(\sup_{t\leq T} |e|^2 \geq \kappa_h^2 \right) \leq \mathbb{P}(\tau < T) + \frac{\mathbb{E}\sup_{t\leq\tau} |e|^2}{h^{2\kappa}} \leq 2C_\star \cdot h^{\gamma p} + h^{2\delta - 2\kappa}.$$

The optimal (i.e., lowest) parameter κ is $\kappa = \delta - \frac{\gamma p}{2}$, then

$$\mathbb{P}\left(\sup_{t\leq T} |e|^2 \geq \kappa_h^2 \right) \leq (2C_\star + 1) h^{\gamma p}.$$

Then

$$\begin{aligned} \mathbb{E}|e(t)|^q &= \left(\chi_{\{\sup_{t\leq\tau} |e(t)|^2 \leq \kappa_h^2\}} |e(t)|^q \right) + \left(\chi_{\{\sup_{t\leq\tau} |e(t)|^2 > \kappa_h^2\}} |e(t)|^q \right) \\ &\leq \kappa_h^q + \left(\mathbb{P}\left(\sup_{t\leq T} |e|^2 \geq \kappa_h^2 \right)^{\frac{s-q}{s}} \cdot \left(\mathbb{E}|e(t)|^s \right)^{\frac{q}{s}} \right) . \end{aligned}$$

Now we can take the supremum over time $[0, T]$, and set our optimal parameter $\kappa = \delta - \gamma p/2$ to obtain

$$\sup_{t\in[0,T]} \mathbb{E}|e(t)|^q \leq C \cdot [h^{(\delta - \frac{\gamma p}{2})\cdot q} + h^{\gamma p \frac{s-q}{s}}] \, ,$$

which proves the first claim. If also $\delta = \frac{1-r\gamma}{2}$, then

$$\sup_{t\in[0,T]} \mathbb{E}|e(t)|^q \leq C \cdot [h^{\frac{1}{2}(1-r\gamma-\gamma p)\cdot q} + h^{\gamma p \frac{s-q}{s}}]$$

and we can extract the optimal value for $\gamma = \frac{qs}{p(qs+2(s-q))+qrs}$, which makes both rates identical. Thus, we derive

$$\sup_{t\in[0,T]} \mathbb{E}|e(t)|^q \le C \cdot h^{q\cdot \frac{p(s\quad q)}{2p(s\quad q)+(p+r)qs}} .$$

Taking the qth root yields the claim. ∎

We can now present the main result of this paper.

THEOREM 8 (strong convergence of the EnKF numerical scheme (5) to the SDE (1)). *For any* $0<\alpha<2$ *and* $0<\eta<1$,

$$\lim_{h=\varepsilon\to 0} h^{-\frac{1}{2}\cdot\frac{3\quad\alpha}{3+25/3\cdot\alpha}} \cdot \left(\sup_{t\in[0,T]} \mathbb{E}|\bar{v}(t)-u(t)|^{\alpha}\right)^{\frac{1}{\alpha}} = 0,$$

$$\lim_{h=\varepsilon\to 0} h^{-\frac{1}{2}\cdot\frac{1\quad\eta}{1+3\eta}} \cdot \left(\mathbb{E}\sup_{t\in[0,T]} |\bar{v}(t)-u(t)|^{\eta}\right)^{\frac{1}{\eta}} = 0.$$

Idea of proof. For the proof we will rely on Lemmas 5, 7 and Corollary 6. First we need to verify a priori estimates for the approximation $\bar{v}$, the regularization v, and the solution of the SDE u, which we will do in section 7. For the error estimate up to a stopping time, we use the triangle inequality $|\bar{v}-u| \le |\bar{v}-v|+|v-u|$ and bound $|v-u|$ in Lemma 10 and $|\bar{v}-v|$ in Lemma 11 in the following two sections. ∎

Remark 9. Note that the bottleneck for η is the a priori bound on $\bar{v}$. In the case $\mathbb{E}\sup_{t\in[0,T]}|\bar{v}(t)|^{\kappa} < C$ for $0<\kappa<3/2$, then $\eta<1$ will get replaced by $\eta<3/2$ in Theorem 8.

5. Bounding the difference between the SDE and its regularization. We denote by $r=u-v$ the difference between the solution of the SDE (1) and its regularization (4). Further, we define the bounded stopping time

$$\tau = \inf\{t>0 : \ \max\{|u|,|v|\} \ge h^{-\gamma}\} \wedge T ,$$

i.e., the first time that any of u and v become "large." This section is devoted to the establishment of bounds on the difference of the SDE and the Lipschitz regularized version, in particular, we will present rates w.r.t. the regularization parameter $h=\epsilon$.

For simplicity, we introduce the notation h^{1-} which is supposed to mean $h^{1-\kappa}$ for some $\kappa\in(0,1)$ (and κ close to 0).

LEMMA 10. *For any* $0<\alpha<3$ *and* $0<\beta<3/2$, *we have*

$$\lim_{h\to 0} h^{-\frac{1}{2}\frac{3/2\quad\beta}{3/2+2\beta}} \cdot \left(\mathbb{E}\sup_{[t\in[0,T]} |r|^{\beta}\right)^{\frac{1}{\beta}} = 0,$$

$$\lim_{h\to 0} h^{-\frac{1}{2}\frac{3\quad\alpha}{3+13/2\cdot\alpha}} \cdot \left(\sup_{[t\in[0,T]} \mathbb{E}|r|^{\alpha}\right)^{\frac{1}{\alpha}} = 0.$$

Proof. The proof will follow these steps:

1. First we use Ito's formula to obtain a bound on moments of r of the form $\mathbb{E}r^2(t\wedge\tau)$. Doing this, we will even obtain a better estimate, i.e.,

$$\mathbb{E}r^2(t\wedge\tau) + \mathbb{E}\left(\int_0^{t\wedge\tau} r^2(s)\cdot\frac{u^2+v^2}{1+hv^2}\mathrm{d}s\right) \le C\cdot h^{1-} . \tag{23}$$

2. Next, we use the previous result to get a bound on moments of suprema of r, i.e.,

$$\mathbb{E} \sup_{t\in[0,\tau]} r^2(t) \leq C \cdot h^{1-} \,. \tag{24}$$

3. Finally, we employ Corollary 6 to bootstrap our estimates, which are up to a stopping time. This yields moments up to time T with a rate of strong convergence.

5.1. Step 1. If $r = u - v$, and recalling

$$\mathrm{d}v = -\frac{v^3}{1+hv^2}\mathrm{d}t + \frac{v^2}{1+hv^2}\mathrm{d}W$$

and

$$\mathrm{d}u = -u^3\mathrm{d}t + u^2\mathrm{d}W,$$

we have that

$$\begin{aligned}
\mathrm{d}r &= -\frac{u^3-v^3}{1+hv^2}\mathrm{d}t - h\frac{u^3v^2}{1+hv^2}\mathrm{d}t + \frac{u^2-v^2}{1+hv^2}\mathrm{d}W + h\frac{u^2v^2}{1+hv^2}\mathrm{d}W \\
&= -r\frac{u^2+uv+v^2}{1+hv^2}\mathrm{d}t - h\frac{u^3v^2}{1+hv^2}\mathrm{d}t + r\frac{u+v}{1+hv^2}\mathrm{d}W + h\frac{u^2v^2}{1+hv^2}\mathrm{d}W \\
&=: -r\cdot T_1(u,v)\mathrm{d}t - h\frac{u^3v^2}{1+hv^2}\mathrm{d}t + r\cdot T_2(u,v)\mathrm{d}W + h\frac{u^2v^2}{1+hv^2}\mathrm{d}W \,.
\end{aligned} \tag{25}$$

Thus, by Ito's formula

$$\begin{aligned}
\mathrm{d}r^2 &= 2r\cdot\mathrm{d}r + (\mathrm{d}r)^2 \\
&= -2r^2\cdot T_1\mathrm{d}t - 2hr\frac{u^3v^2}{1+hv^2}\mathrm{d}t + r^2\cdot T_2^2\mathrm{d}t + 2h\cdot r\cdot T_2\frac{u^2v^2}{1+hv^2}\mathrm{d}t \\
&\quad + h^2\cdot\frac{u^4v^4}{(1+hv^2)^2}\mathrm{d}t + 2r^2\cdot T_2\mathrm{d}W + 2h\cdot r\cdot\frac{u^2v^2}{1+hv^2}\mathrm{d}W.
\end{aligned}$$

In integral form this is

$$\begin{aligned}
r^2(s\wedge\tau) = &-\int_0^{s\wedge\tau} r^2\cdot(2T_1 - T_2^2)\mathrm{d}t \\
&+ h\cdot\int_0^{s\wedge\tau} -2r\frac{u^3v^2}{1+hv^2} + 2\cdot r\cdot T_2\frac{u^2v^2}{1+hv^2} + h\cdot\frac{u^4v^4}{(1+hv^2)^2}\mathrm{d}t \\
&+ \int_0^{s\wedge\tau} 2e^2\cdot T_2\mathrm{d}W + h\cdot\int_0^t 2r\cdot\frac{u^2v^2}{1+hv^2}\mathrm{d}W \,.
\end{aligned}$$

Now

$$\begin{aligned}
2T_1 + T_2^2 &= \frac{2u^2+2uv+2v^2}{1+hv^2} - \frac{(u+v)^2}{(1+hv^2)^2} \\
&\geq \frac{2u^2+2uv+2v^2}{(1+hv^2)^2} - \frac{u^2+2uv+v^2}{(1+hv^2)^2} = \frac{u^2+v^2}{(1+hv^2)^2} \geq 0
\end{aligned}$$

and thus we know that the first integral is negative. We can summarize the part in square brackets by noting that $\tau \leq T$ and that the integrand is bounded by some

power of h, more accurately, $h^{-6\gamma}$ (because u, v, r, T_2 all are bounded as we only integrate up to stopping time τ). Finally, we get

$$r^2(s\wedge\tau)+\int_0^{s\wedge\tau} r^2\cdot\frac{u^2+v^2}{1+hv^2}\mathrm{d}t \le h^{1-6\gamma}\cdot C\cdot T+\int_0^{s\wedge\tau} 2e^2\cdot T_2\mathrm{d}W+h\cdot\int_0^{s\wedge\tau} 2r\cdot\frac{u^2v^2}{1+hv^2}\mathrm{d}W \tag{26}$$

and thus

$$\mathbb{E}r^2(s\wedge\tau)+\mathbb{E}\int_0^{s\wedge\tau} r^2\cdot\frac{u^2+v^2}{1+hv^2}\mathrm{d}t \le h^{1-6\gamma}\cdot C\cdot T$$

which is (23).

5.2. Step 2. Going one step back to (26) (where we drop for now the second positive term on the left-hand side), we take a look at the supremum of r^2, of which we know now

$$\sup_{t\in[0,\tau]} r^2(t) \le h^{1-6\gamma}\tilde{C}+2\sup_{t\in[0,\tau]}\int_0^t r^2\cdot T_2\mathrm{d}W + 2h\sup_{t\in[0,\tau]}\int_0^{s\wedge\tau} r\cdot\frac{u^2v^2}{1+hv^2}\mathrm{d}W\,.$$

Applying the expectation, we have

$$\mathbb{E}\sup_{t\in[0,\tau]} r^2(t) \le h^{1-6\gamma}\tilde{C}+2\mathbb{E}\sup_{t\in[0,\tau]}\int_0^t r^2\cdot T_2\mathrm{d}W + 2h\mathbb{E}\sup_{t\in[0,\tau]}\int_0^{s\wedge\tau} r\cdot\frac{u^2v^2}{1+hv^2}\mathrm{d}W\,.$$

The last term's integrand can be bound again by some power of h as done before for the deterministic integral. Then we just have an expectation of the supremum of the Brownian motion up to a bounded stopping time (which is a constant). The first term behaves nicely under $h\to 0$ so all we need to take care of is the middle term. For this we need the Burkholder–Davis–Gundy inequality

$$\mathbb{E}\sup_{t\in[0,\tau]}\int_0^t r^2\cdot T_2\mathrm{d}W \le C\mathbb{E}\left[\int_0^\tau r^4\cdot T_2^2\mathrm{d}t\right]^{\frac{1}{2}}$$

if the right-hand side is well-defined (i.e., finite). We show that it indeed is finite by further bounding

$$\mathbb{E}\sup_{t\in[0,\tau]}\int_0^t r^2\cdot T_2\mathrm{d}W \le C\mathbb{E}\left[\sup_{t\in[0,\tau]} r^2(t)\cdot\int_0^\tau r^2\cdot T_2^2\mathrm{d}t\right]^{\frac{1}{2}}$$

and an application of the Cauchy–Schwarz inequality yields

$$\mathbb{E}\sup_{t\in[0,\tau]}\int_0^t r^2\cdot T_2\mathrm{d}W \le C\left[\mathbb{E}\sup_{t\in[0,\tau]} r^2(t)\right]^{\frac{1}{2}}\cdot\left[\mathbb{E}\int_0^\tau r^2\cdot T_2^2\mathrm{d}t\right]^{\frac{1}{2}}.$$

Observe that the left term is what we started with. We do not worry that the term might be infinite because we only go up to a stopping time. We rather want to show that this term goes to 0 for $h \to 0$. Inserting the just obtained bound above, we get

$$\mathbb{E} \sup_{t\in[0,\tau]} r^2(t) \le h^{1-6\gamma}\tilde{C} + 2C \left(\mathbb{E} \sup_{t\in[0,\tau]} r^2(t)\right)^{\frac{1}{2}} \cdot \left(\mathbb{E} \int_0^\tau r^2 \cdot T_2^2 \mathrm{d}t\right)^{\frac{1}{2}} + 2Ch^{1-5\gamma}.$$

The second expectation on the right-hand side can be bounded as follows:

$$\mathbb{E} \int_0^\tau r^2 \cdot T_2^2 \mathrm{d}t = \mathbb{E} \int_0^\tau r^2 \cdot \frac{(u+v)^2}{(1+hv^2)^2} \mathrm{d}t \le 2\mathbb{E} \int_0^\tau \frac{u^2+v^2}{(1+hv^2)^2} \cdot r^2 \mathrm{d}t \le Ch^{1-6\gamma},$$

where the last inequality comes from (23). All in all,

$$\mathbb{E} \sup_{t\in[0,\tau]} r^2(t) \le C \cdot h^{1-6\gamma} + 2C \cdot h^{\frac{1}{2}-3\gamma} \cdot \left(\mathbb{E} \sup_{t\in[0,\tau]} r^2(t)\right)^{\frac{1}{2}}.$$

This is equivalent to finding a bound on A given that $A \le Ch^{1-\kappa} + C'\sqrt{A}h^{\frac{1-\kappa}{2}}$, which yields $A \le C'' \cdot h^{1-\kappa}$ and thus for our problem,

$$\mathbb{E} \sup_{t\in[0,\tau]} r^2(t) \le C \cdot h^{1-6\gamma},$$

which is (24).

5.3. Step 3. Now we employ Corollary 6 with $v_1 = u$, $v_2 = v$, $p = 3/2$, $\rho = 6$. This yields

$$\left(\mathbb{E} \sup_{[t\in[0,T]} |r|^\beta\right)^{\frac{1}{\beta}} \le Ch^{\frac{1}{2}\frac{3/2-\beta}{3/2+2\beta}}.$$

For the slightly weaker convergence condition we can apply Lemma 7 with $v_1 = u$, $v_2 = v$, $p = 3/2$, $s = 3$, and $\rho = 6$, in order to obtain

$$\left(\sup_{[t\in[0,T]} \mathbb{E}|r|^\alpha\right)^{\frac{1}{\alpha}} \le Ch^{\frac{3-\alpha}{6+13\alpha}}.$$

∎

6. Bounding the difference between the regularization and its Euler–Maruyama discretization. We define $e = \bar{v} - v$, i.e., we consider now the difference between the regularization (4) and its Euler–Maruyama discretization (5). The main result of this section is the following lemma. The whole section is devoted to its proof. The idea of the proof is based on the a priori bounds in section 7 and the application of Lemmas 5, 7 and Corollary 6.

LEMMA 11. *For any $\eta \in (0,1)$ and $\alpha \in (0,3)$, we have*

$$\lim_{h\to 0} h^{-\frac{1}{2}\cdot\frac{1-\eta}{1+3\eta}} \cdot \left(\mathbb{E} \sup_{t\in[0,T]} |e(t)|^\eta\right)^{\frac{1}{\eta}} = 0,$$

$$\lim_{h\to 0} h^{-\frac{1}{2}\cdot\frac{3-\alpha}{3+25/3\cdot\alpha}} \cdot \left(\mathbb{E} \sup_{t\in[0,T]} |e(t)|^\alpha\right)^{\frac{1}{\alpha}} = 0.$$

6.1. A priori bounds for the error. We can write, defining $\eta(t) = k$ and $\eta^+(t) = k+1$ for $t \in [t_k, t_{k+1})$,

$$\bar{v}(t) = v_0 + \int_0^t f(\bar{v}(s))\mathrm{d}s + \int_0^t \sigma(\bar{v}(s))\mathrm{d}W + \underbrace{\sum_{k=1}^{\eta^+(t)} \int_{t_{k-1}}^{t\wedge t_k} f(v_k) - f(\bar{v}(s))\mathrm{d}s + \int_{t_{k-1}}^{t\wedge t_k} \sigma(v_k) - \sigma(\bar{v}(s))\mathrm{d}W_s}_{\mathrm{Res}(t)}.$$

Now,

$$\mathrm{d}\,\mathrm{Res}(t) = [f(v_{\eta(t)}) - f(\bar{v}(t))]\mathrm{d}t + [\sigma(v_{\eta(t)}) - \sigma(\bar{v}(t))]\mathrm{d}W_t.$$

Next, with the error between solution and interpolated scheme $e = \bar{v} - v$,

$$\mathrm{d}e = [f(v+e) - f(v)]\mathrm{d}t + [\sigma(v+e) - \sigma(v)]\mathrm{d}W + \mathrm{d}\,\mathrm{Res}(t)$$

and

$$\frac{1}{2}\mathrm{d}|e|^2 = [f(v+e) - f(v)]e\mathrm{d}t + [\sigma(v+e) - \sigma(v)]e\mathrm{d}W + e \cdot \mathrm{d}\,\mathrm{Res} + \frac{1}{2}[\mathrm{d}\,\mathrm{Res}]^2 + \frac{1}{2}[\sigma(v+e) - \sigma(v)]^2\mathrm{d}t + [\sigma(v+e) - \sigma(v)]\mathrm{d}W\mathrm{d}\,\mathrm{Res}.$$

A calculation shows that

$$\begin{aligned}(27)\qquad &[f(v) - f(w)] \cdot (v-w) + \frac{1}{2}[g(v) - g(w)]^2 \\ &\quad = -\frac{(v-w)^2}{2} \cdot \frac{v^2 + w^2 + \varepsilon v^2 w^2 + (v+w)^2 \cdot \left(1 - \frac{1}{(1+\varepsilon v^2)(1+\varepsilon w^2)}\right)}{1 + \varepsilon v^2 + \varepsilon w^2 + \varepsilon^2 v^2 w^2}\end{aligned}$$

and using this above yields

$$\begin{aligned}\frac{1}{2}\mathrm{d}|e|^2 &= -\frac{|e|^2}{2} \cdot T(v,e) \cdot \mathrm{d}t + [\sigma(v+e) - \sigma(v)]e\mathrm{d}W \\ &\quad + e \cdot [f(v_{\eta(t)}) - f(\bar{v}(t))]\mathrm{d}t + e \cdot [\sigma(v_{\eta(t)}) - \sigma(\bar{v}(t))]\mathrm{d}W \\ &\quad + \frac{1}{2}[\sigma(v_{\eta(t)}) - \sigma(\bar{v}(t))]^2\mathrm{d}t + [\sigma(v+e) - \sigma(v)][\sigma(v_{\eta(t)}) - \sigma(\bar{v}(t))]\mathrm{d}t\end{aligned}$$

with $T(v,e) \geq 0$ defined as

$$\begin{aligned}T(v,e) &= \frac{v^2 + (v+e)^2 + \varepsilon^2 v^2 (v+e)^2 + (v + (v+e))^2 \cdot \left(1 - \frac{1}{(1+\varepsilon v^2)(1+\varepsilon(v+e)^2)}\right)}{1 + \varepsilon v^2 + \varepsilon(v+e)^2 + \varepsilon^2 v^2 (v+e)^2} \\ &\geq C\tilde{T}(v,\bar{v})\end{aligned}$$

where $\min\{\frac{1}{\epsilon};\ \bar{v}^2 + v^2\} =: \tilde{T}(v,\bar{v})$ and the inequality holds from Calculation 15.

In proving the convergence of the scheme to v, we go through very similar steps 1–4 as in the previous section (but this time for $e =$ the discrete error instead of the difference between both solutions). We know

$$|\bar{v}(t) - v_n| \leq h \cdot |f(v_n)| + |W(t) - W(t_n)| \cdot \sigma(v_n)$$

and thus

$$|\bar{v}(t) - v_n|^2 \leq Ch^2|f(v_n)|^2 + C|W(t) - W(t_n)|^2 \cdot |\sigma(v_n)|^2. \tag{28}$$

In particular,

$$\mathbb{E}|\bar{v}(t) - v_n|^2 \leq Ch^2\mathbb{E}|f(v_\eta)|^2 + Ch\mathbb{E}|\sigma(v_\eta)|^2 \ . \tag{29}$$

With this, we derive

$$\begin{aligned}\tilde{T}(v_\eta, \bar{v}) &= \min\{\varepsilon^{-1}, |v_\eta|^2 + |\bar{v}|^2\} \leq C\min\{\varepsilon^{-1}, |v_\eta - \bar{v}|^2 + |\bar{v}|^2\} \\ &\leq \tilde{T}(v, \bar{v}) + \min\{\varepsilon^{-1}, Ch^2|f(v_n)|^2 + C|W(t) - W(t_n)|^2 \cdot \sigma(v_n)\} \ .\end{aligned}$$

We use this in our expression for $\mathrm{d}e^2$ to get rid of the term $\delta\tilde{T}(v_\eta, \bar{v})|e|^2\mathrm{d}t$:

$$\begin{aligned}\frac{1}{2}\mathrm{d}|e|^2 \leq &-\frac{|e|^2}{2} \cdot \tilde{T}(v, \bar{v}) \cdot \mathrm{d}t + \delta\tilde{T}(v_\eta, \bar{v})|e|^2\mathrm{d}t + \delta\tilde{T}(\bar{v}, v)|e|^2\mathrm{d}t \\ &+ 2C_\delta\tilde{T}(v_\eta, \bar{v})|v_\eta - \bar{v}|^2\mathrm{d}t + \frac{1}{2}\tilde{T}(v_\eta, \bar{v})|v_\eta - \bar{v}|^2\mathrm{d}t \\ &+ (\sigma(v + e) - \sigma(v)) \cdot e \ \mathrm{d}W + (\sigma(v_\eta) - \sigma(v + e)) \cdot e \ \mathrm{d}W \\ \leq &-\frac{|e|^2}{2} \cdot \tilde{T}(v, \bar{v}) \cdot \mathrm{d}t \\ &+ \delta\tilde{T}(v, \bar{v})|e|^2\mathrm{d}t \\ &+ \delta\min\{\varepsilon^{-1}, Ch^2|f(v_n)|^2 + C|W(t) - W(t_n)|^2 \cdot |\sigma(v_n)|^2\}|e|^2\mathrm{d}t \\ &+ \delta\tilde{T}(\bar{v}, v)|e|^2\mathrm{d}t \\ &+ 2C_\delta\tilde{T}(v_\eta, \bar{v})|v_\eta - \bar{v}|^2\mathrm{d}t + \frac{1}{2}\tilde{T}(v_\eta, \bar{v})|v_\eta - \bar{v}|^2\mathrm{d}t \\ &+ (\sigma(v + e) - \sigma(v)) \cdot e \ \mathrm{d}W + (\sigma(v_\eta) - \sigma(v + e)) \cdot e \ \mathrm{d}W \ .\end{aligned}$$

Now for small enough δ we can bring all terms of the form $C \cdot \tilde{T}(v, \bar{v})|e|^2\mathrm{d}t$ as a positive summand to the left-hand side because of the negatively dominating first term:

$$\begin{aligned}&\frac{1}{2}\mathrm{d}|e|^2 + \frac{1}{2}(1 - 4\delta) \cdot |e|^2 \cdot \tilde{T}(v, \bar{v})\mathrm{d}t \\ &\quad\leq \delta\min\{\varepsilon^{-1}, Ch^2|f(v_n)|^2 + C|W(t) - W(t_n)|^2 \cdot |\sigma(v_n)|^2\}|e|^2\mathrm{d}t \\ &\quad\quad + 2C_\delta\tilde{T}(v_\eta, \bar{v})|v_\eta - \bar{v}|^2\mathrm{d}t + \frac{1}{2}\tilde{T}(v_\eta, \bar{v})|v_\eta - \bar{v}|^2\mathrm{d}t \\ &\quad\quad + (\sigma(v + e) - \sigma(v)) \cdot e \ \mathrm{d}W + (\sigma(v_\eta) - \sigma(v + e)) \cdot e \ \mathrm{d}W \ ,\end{aligned}$$

and merging similar terms,

$$\begin{aligned}&\leq \delta\min\{\varepsilon^{-1}, Ch^2|f(v_n)|^2 + C|W(t) - W(t_n)|^2 \cdot |\sigma(v_n)|^2\}|e|^2\mathrm{d}t \\ &\quad + C_\delta\tilde{T}(v_\eta, \bar{v})|v_\eta - \bar{v}|^2\mathrm{d}t \\ &\quad + (\sigma(v + e) - \sigma(v)) \cdot e \ \mathrm{d}W + (\sigma(v_\eta) - \sigma(v + e)) \cdot e \ \mathrm{d}W \ .\end{aligned}$$

6.2. Bounds up to the stopping time. Now for M fixed later we set

$$\tau = T \wedge \inf\{t > 0 : |v(t)| \geq M \vee |\bar{v}(t)| \geq M \vee |v_\eta(t)| \geq M\} \ .$$

Up to this stopping time, $|f(v_\eta)| \le M^3$, $|\sigma(v_\eta)| \le M^2$, and $|\tilde{T}(v_\eta, \bar{v})| \le CM^2$, as well as (from (29))

$$\mathbb{E}|\bar{v}(t) - v_n|^2 \le ChM^6 \,. \tag{30}$$

Then

$$\begin{aligned}\frac{1}{2}|e(\tau)|^2 + \frac{1}{2}(1-4\delta)\cdot \int_0^\tau |e(t)|^2 \cdot \tilde{T}(v,\bar{v})\mathrm{d}t &\le C\delta h^2 \cdot M^6 \int_0^\tau |e(t)|^2 \mathrm{d}t \\ &+ C\delta M^4 \cdot \int_0^\tau |W(t) - W(t_n)|^2 \cdot |e(t)|^2 \mathrm{d}t + C_\delta M^2 \cdot \int_0^\tau |v_\eta - \bar{v}|^2 \mathrm{d}t \\ &+ \int_0^\tau (\sigma(v+e) - \sigma(v)) \cdot e \,\mathrm{d}W + \int_0^\tau (\sigma(v_\eta) - \sigma(v+e)) \cdot e \,\mathrm{d}W\end{aligned} \tag{31}$$

or better, with a different time argument,

$$\begin{aligned}\frac{1}{2}|e(s\wedge\tau)|^2 + \frac{1}{2}(1-4\delta)\cdot \int_0^{s\wedge\tau} |e(t)|^2 \cdot \tilde{T}(v,\bar{v})\mathrm{d}t &\le C\delta h^2 \cdot M^6 \int_0^{s\wedge\tau} |e(t)|^2 \mathrm{d}t \\ &+ C\delta M^4 \cdot \int_0^{s\wedge\tau} |W(t) - W(t_n)|^2 \cdot |e(t)|^2 \mathrm{d}t + C_\delta M^2 \cdot \int_0^{s\wedge\tau} |v_\eta - \bar{v}|^2 \mathrm{d}t \\ &+ \int_0^{s\wedge\tau} (\sigma(v+e) - \sigma(v)) \cdot e \,\mathrm{d}W + \int_0^{s\wedge\tau} (\sigma(v_\eta) - \sigma(v+e)) \cdot e \,\mathrm{d}W \,.\end{aligned} \tag{32}$$

Up to stopping time τ, we can brute-force bound $|e(t)|^2 \le 2M^2$. Second, in order to be able to apply the expectation on both sides in the next step, we estimate all positive integrals on the right-hand side from above by replacing the upper integral boundary $s \wedge \tau$ by its deterministic upper bound T:

$$\begin{aligned}\frac{1}{2}|e(s\wedge\tau)|^2 + \frac{1}{2}(1-4\delta)\cdot \int_0^{s\wedge\tau} |e(t)|^2 \cdot \tilde{T}(v,\bar{v})\mathrm{d}t &\le C\delta h^2 \cdot M^8 T \\ &+ C\delta M^6 \cdot \int_0^T |W(t) - W(t_n)|^2 \mathrm{d}t + C_\delta M^2 \cdot \int_0^T |v_\eta - \bar{v}|^2 \mathrm{d}t \\ &+ \int_0^{s\wedge\tau} (\sigma(v+e) - \sigma(v)) \cdot e \,\mathrm{d}W + \int_0^{s\wedge\tau} (\sigma(v_\eta) - \sigma(v+e)) \cdot e \,\mathrm{d}W \,.\end{aligned}$$

We take the expectation which removes the last two integrals and use (30) to obtain

$$\begin{aligned}\frac{1}{2}\mathbb{E}|e(s\wedge\tau)|^2 + \frac{1}{2}(1-4\delta)\cdot \mathbb{E}\int_0^{s\wedge\tau} |e(t)|^2 \cdot \tilde{T}(v,\bar{v})\mathrm{d}t \\ \le C\delta h^2 \cdot M^8 T + hC\delta M^6 T + hC_\delta M^8 T \,,\end{aligned}$$

i.e.,

$$\frac{1}{2}\mathbb{E}|e(s\wedge\tau)|^2 + \frac{1}{2}(1-4\delta)\cdot \mathbb{E}\int_0^{s\wedge\tau} |e(t)|^2 \cdot \tilde{T}(v,\bar{v})\mathrm{d}t \le C_\delta h M^8 T \,. \tag{33}$$

We go back to (32), drop the second term on the left-hand side, and apply the supremum to get

$$\begin{aligned}\sup_{s\le\tau} \frac{1}{2}|e(s)|^2 &\le C\delta h^2 M^8 T + C\delta M^6 \int_0^T |W(t) - W(t_n)|^2 \mathrm{d}t + C_\delta M^2 \int_0^T |v_\eta - \bar{v}|^2 \mathrm{d}t \\ &+ \sup_{s\le\tau} \int_0^s (\sigma(v+e) - \sigma(v)) \cdot e \,\mathrm{d}W + \sup_{s\le\tau} \int_0^s (\sigma(v_\eta) - \sigma(v+e)) \cdot e \,\mathrm{d}W\end{aligned}$$

and thus

$$\begin{aligned}\mathbb{E}\sup_{s\le\tau}\frac{1}{2}|e(s)|^2 \le{}& C\delta h^2 M^8 T + C\delta M^6 hT + C_\delta M^8 h\\ &+ \mathbb{E}\sup_{s\le\tau}\left|\int_0^s (\sigma(v+e)-\sigma(v))\cdot e\,\mathrm{d}W\right|\\ &+ \mathbb{E}\sup_{s\le\tau}\left|\int_0^s (\sigma(v_\eta)-\sigma(v+e))\cdot e\,\mathrm{d}W\right|.\end{aligned}$$

Using the Burkholder–Davis–Gundy inequality,

$$\begin{aligned}\mathbb{E}\sup_{s\le\tau}\left|\int_0^s [\sigma(\bar v)-\sigma(v)]\cdot e\,\mathrm{d}W\right| &\le \mathbb{E}\left[\int_0^\tau |\sigma(\bar v)-\sigma(v)|^2|e|^2\mathrm{d}t\right]^{\frac12}\\ &\le \mathbb{E}\left[\int_0^\tau (|\bar v|^2+|v|^2)|e|^4\mathrm{d}t\right]^{\frac12}\\ &\le \sqrt{\mathbb{E}\sup_{t\le\tau}|e|^2}\cdot\sqrt{\mathbb{E}\int_0^T (|\bar v|^2+|v|^2)|e|^2\mathrm{d}t}\\ &\le \sqrt{\mathbb{E}\sup_{t\le\tau}|e|^2}\cdot\sqrt{C_\delta M^8 Th},\end{aligned}$$

where the last step is due to (33). Similarly,

$$\begin{aligned}\mathbb{E}\sup_{s\le\tau}\left|\int_0^s [\sigma(v_\eta)-\sigma(\bar v)]\cdot e\,\mathrm{d}W\right| &\le \sqrt{\mathbb{E}\sup_{t\le\tau}|e|^2}\cdot\sqrt{\mathbb{E}\int_0^\tau (|v_\eta|^2+|\bar v|^2)|v_\eta-\bar v|^2\mathrm{d}t}\\ &\le \sqrt{\mathbb{E}\sup_{t\le\tau}|e|^2}\sqrt{\mathbb{E}\int_0^T CM^2|v_\eta-\bar v|^2\mathrm{d}t}\\ &\le \sqrt{\mathbb{E}\sup_{t\le\tau}|e|^2}\sqrt{CTM^6h}\,.\end{aligned}$$

So, all in all,

$$\mathbb{E}\sup_{s\le\tau}\frac{1}{2}|e(s)|^2 \le C_\delta hM^8T + \sqrt{\mathbb{E}\sup_{s\le\tau}|e|^2}\cdot C\cdot M^4h^{\frac12}T^{\frac12}$$

which yields

$$\mathbb{E}\sup_{s\le\tau}\frac{1}{2}|e(s)|^2 \le C_\delta hM^8T\,. \tag{34}$$

Now we set $M = h^{-\gamma}$ for some $\gamma > 0$ small (at least $\gamma < \frac18$).

This amounts to our previous results being

$$\frac12\mathbb{E}|e(s\wedge\tau)|^2 + \frac12(1-4\delta)\cdot\mathbb{E}\int_0^{s\wedge\tau}|e(t)|^2\cdot\tilde T(v,\bar v)\mathrm{d}t \le C_\delta h^{1-8\gamma}T \tag{35}$$

and

$$\mathbb{E}\sup_{s\le\tau}\frac{1}{2}|e(s)|^2 \le C_\delta h^{1-8\gamma}T\,. \tag{36}$$

6.3. Conclusion of the proof. Now we can rely on the a priori estimates of section 7 and use Corollary 6 with $v_1 = v$, $v_2 = \bar{v}$, $p = 1$, and $r = 8$. Thus we get

$$\left(\mathbb{E} \sup_{t\in[0,T]} |e(t)|^\eta \right)^{\frac{1}{\eta}} \leq C h^{\frac{1}{2}\cdot\frac{1}{1+3\eta}} \quad \text{for } \eta < 3/2.$$

Also, application of Lemma 7 with $s = 3$ additionally yields

$$\left(\sup_{t\in[0,T]} \mathbb{E}|e(t)|^\alpha \right)^{\frac{1}{\alpha}} \leq C h^{\frac{1}{2}\cdot\frac{3\alpha}{3+25/3\cdot\alpha}} \qquad \text{for } \alpha < 2.$$

7. A priori estimates. We start by giving a priori estimates for various versions of momenta of the processes u, v, and $\bar{v}$. Recall (we have exchanged ε for h)

$$\mathrm{d}u = -u^3 \mathrm{d}t + u^2 \mathrm{d}W, \tag{37}$$

$$\mathrm{d}v = -\frac{v^3}{1 + h\cdot v^2}\mathrm{d}t + \frac{v^2}{1 + h\cdot v^2}\mathrm{d}W, \tag{38}$$

$$\mathrm{d}\bar{v} = -\frac{v_n^3}{1 + hv_n^2}\mathrm{d}t + \frac{v_n^2}{1 + hv_n^2}\mathrm{d}W \quad \text{for } t \in [nh, (n+1)h). \tag{39}$$

We start with moment bounds on u.

LEMMA 12. *For the solution u of the SDE* (37) *and $T > 0$*

$$\sup_{t\in[0,T]} \mathbb{E}|u(t)|^\alpha + \mathbb{E}\int_0^T |u(s)|^{\alpha+2}\mathrm{d}s < C \tag{40}$$

for every $0 < \alpha < 3$. Also,

$$\mathbb{E}\sup_{t\in[0,T]} |u(t)|^\beta < C \tag{41}$$

for every $0 < \beta < 3/2$.

Note that the constant $C > 0$ is allowed to depend on α, β, and T.

Proof. Ito's formula for u yields

$$\mathrm{d}u^\gamma = -\gamma\left(1 - \frac{\gamma-1}{2}\right)\cdot u^{\gamma+2}\mathrm{d}t + \gamma u^{\gamma+1}\mathrm{d}W. \tag{42}$$

This means that for the finite stopping time $\tau_n = \inf\{t > 0 : u(t) > n\} \wedge T$,

$$\mathbb{E}u(t\wedge\tau_n)^\gamma = -\gamma\left(1 - \frac{\gamma-1}{2}\right)\cdot\mathbb{E}\int_0^{t\wedge\tau_n} u^{\gamma+2}(s)\mathrm{d}s + u_0^\gamma$$

and

$$\limsup_{n\to\infty} \mathbb{E}u(t\wedge\tau_n)^\gamma = -\gamma\left(1 - \frac{\gamma-1}{2}\right)\cdot\mathbb{E}\int_0^{t} u^{\gamma+2}(s)\mathrm{d}s + u_0^\gamma$$

by monotone convergence (we know that $u \geq 0$ and hence $\int_0^{t\wedge\tau_{n+1}} u^{\gamma+2}(s)\mathrm{d}s \geq$

$\int_0^{t\wedge\tau_n} u^{\gamma+2}(s)\mathrm{d}s$). As the left-hand side is positive and the first term on the right-hand side is negative for $\gamma \in (0,3)$, we can take the supremum to obtain the first claim

$$\sup_{t\in[0,T]} \mathbb{E}u(t)^\gamma + \mathbb{E}\int_0^t u^{\gamma+2}(s)\mathrm{d}s < C\,.$$

In particular,

$$\sup_{t\in[0,T]} \mathbb{E}\int_0^t u^{p}(s)\mathrm{d}s < C$$

exists and is bounded for all $p \in (2,5)$ with a constant depending on p and T. With this, we obtain the following result for the supremum:

$$\begin{aligned}\mathbb{E}\sup_{[0,\tau_n]} u^\gamma(t) &\le u_0^\gamma + \mathbb{E}\sup_{[0,\tau_n]} \left[-\gamma\left(1-\frac{\gamma-1}{2}\right)\cdot\int_0^t u^{\gamma+2}(s)\mathrm{d}s + \int_0^t \gamma u^{\gamma+1}\mathrm{d}W(s)\right]\\ &\le u_0^\gamma + \mathbb{E}\sup_{[0,\tau_n]}\int_0^t \gamma u^{\gamma+1}\mathrm{d}W(s)\end{aligned}$$

because the first term in the square brackets is negative. Now we continue with the Burkholder–Davis–Gundy inequality

$$\begin{aligned}&\le u_0^\gamma + \mathbb{E}\left[\int_0^{\tau_n}\gamma^2 u^{2\gamma+2}(s)\mathrm{d}s\right]^{\frac12}\\ &\le u_0^\gamma + \left[\mathbb{E}\int_0^{\tau_n}\gamma^2 u^{2\gamma+2}(s)\mathrm{d}s\right]^{\frac12}\end{aligned}$$

and the last integral exists for $2\gamma+2 \in (2,5)$, i.e., $\gamma \in (0,3/2)$. ∎

LEMMA 13. *For $T>0$*

$$\sup_{t\in[0,T]} \mathbb{E}|v(t)|^\alpha + \mathbb{E}\int_0^T \frac{v^{\alpha+2}}{(1+hv^2)^2}\mathrm{d}s < C \tag{43}$$

for all $\alpha < 3$ and

$$\mathbb{E}\sup_{t\in[0,T]} |v(t)|^\beta < C \tag{44}$$

for all $\beta < 3/2$.

Proof. Ito's formula yields

$$\mathrm{d}v^\alpha = -\alpha\cdot\left(1-\frac{\alpha-1}{2}\right)\cdot\frac{v^{\alpha+2}}{(1+hv^2)^2}\mathrm{d}t - \alpha h\cdot\frac{v^{\alpha+4}}{(1+hv^2)^2}\mathrm{d}t + \alpha\frac{v^{\alpha+1}}{1+hv^2}\mathrm{d}W\,.$$

When we define a stopping time $\tau_n = \inf\{t>0 : |v|>n\}$, we obtain

$$\begin{aligned}\mathbb{E}v(t\wedge\tau_n)^\alpha = -\alpha\cdot\left(1-\frac{\alpha-1}{2}\right)\cdot\mathbb{E}\int_0^{t\wedge\tau_n}\frac{v^{\alpha+2}}{(1+hv^2)^2}\mathrm{d}s\\ -\alpha h\cdot\mathbb{E}\int_0^{t\wedge\tau_n}\frac{v^{\alpha+4}}{(1+hv^2)^2}\mathrm{d}s + v_0^\alpha\end{aligned}$$

and, for $\alpha \in (0,3)$ by monotone convergence ($n \to \infty$),

$$\mathbb{E}v(t)^\alpha + \alpha \cdot \left(1 - \frac{\alpha-1}{2}\right) \cdot \int_0^t \frac{v^{\alpha+2}}{(1+hv^2)^2} \mathrm{d}s + \alpha h \cdot \int_0^t \frac{v^{\alpha+4}}{(1+hv^2)^2} \mathrm{d}s = v_0^\alpha,$$

and thus bounded uniformly in h, also after taking the supremum over $[0,T]$ on both sides.

For the second claim, note that for $\beta < 3$ (see the first line of the proof of the first claim)

$$\begin{aligned}\mathbb{E} \sup_{t\in[0,T]} v^\beta &\le \mathbb{E} \sup_{t\in[0,T]} \int_0^t \frac{v^{\beta+1}}{1+hv^2} \mathrm{d}W \\ &\le \mathbb{E} \int_0^T \frac{v^{2\beta+2}}{(1+hv^2)^2} \mathrm{d}t \le C\end{aligned}$$

after an application of the Burkholder–Davis–Gundy inequality and using the bound on the integral obtained above, which holds for $2\beta + 2 \le 5$, i.e., $\beta < 3/2$. ∎

For the discretization we can achieve the same bound on the "sup $\mathbb{E}$" moment, but only order 1 (instead of 3/2 for v) for the $\mathbb{E}$ sup moment.

LEMMA 14. *For the solution $\bar{v}$ of (39), we have for $T > 0$*

$$\sup_{t\in[0,T]} \mathbb{E}|\bar{v}(t)|^2 + \mathbb{E} \int_0^t \frac{\bar{v}^4([s]_h)}{(1+h\bar{v}^2([s]_h))^2} ds \le C \tag{45}$$

and

$$\mathbb{E} \sup_{[0,T]} |\bar{v}| \le C. \tag{46}$$

Proof. Define $[t]_h = t_n$ such that $t \in [t_n, t_{n+1})$. Note first, for $p \in (2,3)$,

$$(|x|^p)' = p|x|^{p-2}x \qquad \text{and} \qquad (|x|^p)'' = p(p-1)|x|^{p-2}.$$

By Ito's formula

$$\begin{aligned}d|\bar{v}|^p(t) = p\bar{v}(t)|\bar{v}(t)|^{p-2} \frac{-\bar{v}^3([t]_h)}{1+h\bar{v}^2([t]_h)} dt + p\bar{v}(t)|\bar{v}(t)|^{p-2} \frac{\bar{v}^2([t]_h)}{1+h\bar{v}^2([t]_h)} dW \\ + \frac{1}{2}p(p-1)|\bar{v}(t)|^{p-2} \frac{\bar{v}^4([t]_h)}{(1+h\bar{v}^2([t]_h))^2} dt\,.\end{aligned}$$

Thus for $t \in [t_n, t_{n+1})$ we derive

$$d|\bar{v}|^p(t) = p|\bar{v}(t)|^{p-2} \left[\bar{v}(t) \frac{-v_n^3}{1+hv_n^2} dt + \bar{v}(t) \frac{v_n^2}{1+hv_n^2} dW + \frac{(p-1)v_n^4}{2(1+hv_n^2)^2} dt \right].$$

In particular, for $p = 2$,

$$\partial_t \mathbb{E}\bar{v}^2(t) = 2\mathbb{E}\bar{v}(t) \frac{-v_n^3}{1+hv_n^2} + \mathbb{E} \frac{v_n^4}{(1+hv_n^2)^2}\,.$$

Using the definition of $\bar{v}$ and the independence of the stochastic increment from the filtration at time t_n yields

$$\begin{aligned}\partial_t \mathbb{E}\bar{v}^2(t) &= -2\mathbb{E}\frac{v_n^4}{1+hv_n^2} + 2(t-t_n)\mathbb{E}\frac{v_n^6}{(1+hv_n^2)^2} + \mathbb{E}\frac{v_n^4}{(1+hv_n^2)^2} \\ &= \mathbb{E}\frac{-2v_n^4(1+hv_n^2) + 2(t-t_n)v_n^6 + v_n^4}{(1+hv_n^2)^2} \\ &= -\mathbb{E}\frac{v_n^4}{(1+hv_n^2)^2} + 2\underbrace{(t-t_n-h)}_{<0}\mathbb{E}\frac{v_n^6}{(1+hv_n^2)^2}\,.\end{aligned}$$

In particular, which proves the first claim for $p=2$,

$$\mathbb{E}\bar{v}^2(t) + \int_0^t \mathbb{E}\frac{\bar{v}^4([s]_h)}{(1+h\bar{v}^2([s]_h))^2} + 2\int_0^t ([s]_h + h - s)\mathbb{E}\frac{v^6([s]_h)}{(1+hv^2([s]_h))^2}ds \le \mathbb{E}\bar{v}^2(0)\,.$$

Regarding the second claim, from the definition we know

$$\begin{aligned}|\bar{v}(t)| &\le |\bar{v}(0)| + \int_0^t \frac{|\bar{v}|^3([s]_h)}{1+h\bar{v}^2([s]_h)}ds + \left|\int_0^t \frac{\bar{v}^2([s]_h)}{1+h\bar{v}^2([s]_h)}dW(s)\right| \\ &\le |\bar{v}(0)| + \left(\int_0^t |\bar{v}(s)|^2 ds\right)^{1/2}\left(\int_0^t \frac{|\bar{v}|^4([s]_h)}{(1+h\bar{v}^2([s]_h))^2}ds\right)^{1/2} \\ &\quad + \left|\int_0^t \frac{\bar{v}^2([s]_h)}{1+h\bar{v}^2([s]_h)}dW(s)\right|.\end{aligned}$$

Using the Burkholder–Davis–Gundy inequality, we obtain

$$\mathbb{E}\sup_{[0,T]}|\bar{v}| \le \mathbb{E}|\bar{v}(0)| + \sqrt{T}\mathbb{E}\bar{v}(0)^2 + \mathbb{E}\left(\int_0^T \frac{\bar{v}^4([s]_h)}{(1+h\bar{v}^2([s]_h))^2}ds\right)^{1/2},$$

where the boundedness of the last integral holds by setting $p=2$ in the first claim. ∎

8. Conclusion. We were able to prove that a simplified version of the numerical scheme arising from EnKF continuum analysis is a strongly converging explicit scheme, which is a rare and desirable property. It is remarkable that the Ensemble Kalman methodology canonically yields a strongly converging numerical scheme for an SDE with nonglobally Lipschitz continuous nonlinearities. The analysis presented here was conducted for a simplified model (one-dimensional state and observation space, two particle setting, identity operator as forward response operator). However, the analysis suggests that the results can be carried over to a much more general setting, i.e., (20) seems to be a strongly convergent iteration for a much broader class of SDE (21). The generalization of the theory will be the subject of future work.

The weak "taming" (i.e., the fact that the numerical scheme divides by $(1+\varepsilon v^2)$) was necessary to prove the a priori moment bounds. This means that the method of proof presented here could in principle point to a method of constructing strongly converging numerical schemes for nonglobally Lipschitz SDEs: By constructing a "Lipschitzified" version (this would be v in our example) of the SDE (here, u) and defining the vanilla Euler–Maruyama discretization (v_n) for this regularized SDE, we might sometimes obtain a strongly converging numerical scheme for the original SDE. Note that this is both similar and contrary to the spirit of Hutzenthaler and Jentzen's

construction of tamed schemes: They "tame" the numerical scheme (or rather its increment term), where in the setting presented here we tame (or "Lipschitzify") the SDE itself. It seems that this can only work for SDEs exhibiting some kind of stability (in this case all solutions tend to stay close to the origin) but further analysis is needed to corroborate this claim.

Appendix. A recurring function is

$$\tilde{T}(a,b) = \min\{\varepsilon^{-1}, a^2 + b^2\}.$$

Calculation 15.

$$\begin{aligned} T(v,e) &= \frac{(v+e)^2 + v^2 + \varepsilon(v+e)^2 v^2 + (2v+e)^2 \cdot \left(1 - \frac{1}{(1+\varepsilon(v+e)^2)(1+\varepsilon v^2)}\right)}{1 + \varepsilon(v+e)^2 + \varepsilon v^2 + \varepsilon^2 v^2 (v+e)^2} \\ &\geq \frac{(v+e)^2 + v^2 + \varepsilon(v+e)^2 v^2}{1 + \varepsilon(v+e)^2 + \varepsilon v^2 + \varepsilon^2 v^2 (v+e)^2} \\ &\geq \frac{\bar{v}^2 + v^2 + \varepsilon \bar{v}^2 v^2}{1 + \varepsilon \bar{v}^2 + \varepsilon v^2 + \varepsilon^2 v^2 \bar{v}^2} \\ &\geq C \min\left\{\frac{1}{\epsilon};\ \bar{v}^2 + v^2 + \varepsilon v^2 \bar{v}^2\right\} \\ &\geq C \min\left\{\frac{1}{\epsilon};\ \bar{v}^2 + v^2 = \tilde{T}(v,\bar{v})\right\}. \end{aligned}$$

Calculation 16.

$$\begin{aligned} |f(\xi) - f(z)| &= \frac{z^2 + z\xi + \xi^2}{(1+\epsilon z^2)(1+\epsilon \xi^2)} \cdot |z - \xi| \\ &\leq \frac{1}{2} \frac{z^2 + \xi^2}{1 + \epsilon z^2 + \epsilon \xi^2} \cdot |z - \xi| \\ &\leq C \min\left\{\frac{1}{\epsilon};\ z^2 + \xi^2\right\} \cdot |z - \xi|. \end{aligned}$$

Thus

$$|f(v_\eta) - f(\bar{v})| \leq \tilde{T}(v_\eta, \bar{v}) |v_\eta - \bar{v}|.$$

Calculation 17.

$$\begin{aligned} |\sigma(\xi) - \sigma(z)| &= \frac{z + \xi}{(1+\epsilon z^2)(1+\epsilon \xi^2)} \cdot |z - \xi| \\ &\leq \sqrt{2} \frac{\sqrt{z^2 + \xi^2}}{1 + \epsilon z^2 + \epsilon \xi^2} \cdot |z - \xi| \qquad \text{by Cauchy–Schwarz} \\ &\leq C \min\left\{\frac{1}{\sqrt{\epsilon}};\ \sqrt{z^2 + \xi^2}\right\} \cdot |z - \xi| \\ &= \sqrt{\tilde{T}(x,\xi)} \cdot |z - \xi|. \end{aligned}$$

Acknowledgments. P. W. thanks Andrew M. Stuart for excellent working conditions in Warwick, Cusanuswerk for additional funding during his stay, as well as Weijun Xu for pointing out a mistake.

REFERENCES

[1] J. Amezcua, K. Ide, E. Kalnay, and S. Reich, *Ensemble transform Kalman–Bucy filters*, Q. J. R. Meteorol. Soc., 140 (2014), pp. 995–1004.

[2] K. Bergemann and S. Reich, *An ensemble Kalman-Bucy filter for continuous data assimilation*, Meteorol. Z., 21 (2012), pp. 213–219.

[3] J. de Wiljes, S. Reich, and W. Stannat, *Long-Time Stability and Accuracy of the Ensemble Kalman-Bucy Filter for Fully Observed Processes and Small Measurement Noise*, preprint, https://arxiv.org/abs/1612.06065, 2016.

[4] P. Del Moral and J. Tugaut, *On the Stability and the Uniform Propagation of Chaos Properties of Ensemble Kalman-Bucy Filters*, Ann. Appl. Probab., 28 (2018), pp. 790–850.

[5] O. G. Ernst, B. Sprungk, and H.-J. Starkloff, *Analysis of the ensemble and polynomial chaos Kalman filters in Bayesian inverse problems*, SIAM/ASA J. Uncertain. Quantif., 3 (2015), pp. 823–851.

[6] G. Evensen, *Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics*, J. Geophy. Res. Oceans, 99 (1994), pp. 10143–10162.

[7] B. G. Fitzpatrick, *Bayesian analysis in inverse problems*, Inverse Problems, 7 (1991), pp. 675–702.

[8] J. N. Franklin, *Well-posed stochastic extensions of ill-posed linear problems*, J. Math. Anal. Appl., 31 (1970), pp. 682–716.

[9] E.-H. Bergou, S. Gratton, and J. Mandel, *On the Convergence of a Non-Linear Ensemble Kalman Smoother*, preprint, arXiv:1411.4608, 2014.

[10] D. J. Higham, X. Mao, and A. M. Stuart, *Strong convergence of Euler-type methods for nonlinear stochastic differential equations*, SIAM J. Numer. Anal., 40 (2002), pp. 1041–1063.

[11] M. Hutzenthaler and A. Jentzen, *Numerical Approximations of Stochastic Differential Equations with Non-globally Lipschitz Continuous Coefficients*, Mem. Amer. Math. Soc. 236, AMS, Providence, RI, 2015.

[12] M. Hutzenthaler, A. Jentzen, and P. E. Kloeden, *Strong and weak divergence in finite time of Euler's method for stochastic differential equations with non-globally Lipschitz continuous coefficients*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 467 (2011), pp. 1563–1576.

[13] M. Hutzenthaler, A. Jentzen, and P. E. Kloeden, *Strong convergence of an explicit numerical method for SDEs with nonglobally Lipschitz continuous coefficients*, Ann. Appl. Probab., (2012), pp. 1611–1641.

[14] M. A. Iglesias, K. J. Law, and A. M. Stuart, *Ensemble Kalman methods for inverse problems*, Inverse Problems, 29 (2013), 045001.

[15] D. Kelly, K. Law, and A. Stuart, *Well-posedness and accuracy of the ensemble Kalman filter in discrete and continuous time*, Nonlinearity, 27 (2014), pp. 2579–2604.

[16] D. Kelly, A. J. Majda, and X. T. Tong, *Concrete ensemble Kalman filters with rigorous catastrophic filter divergence*, Proc. Natl. Acad. Sci. USA, 112 (2015), pp. 10589–10594, https://doi.org/10.1073/pnas.1511063112.

[17] P. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Springer, Berlin, 1992.

[18] E. Kwiatkowski and J. Mandel, *Convergence of the square root ensemble Kalman filter in the large ensemble limit*, SIAM/ASA J. Uncertain. Quantif., 3 (2015), pp. 1–17.

[19] K. Law, A. Stuart, and K. Zygalakis, *Data Assimilation*, Springer, Cham, Switzerland, 2015.

[20] K. J. H. Law, H. Tembine, and R. Tempone, *Deterministic mean-field ensemble Kalman filtering*, SIAM J. Sci. Comput., 38 (2016), pp. A1251–A1279.

[21] F. Le Gland, V. Monbet, and V.-D. Tran, *Large Sample Asymptotics for the Ensemble Kalman Filter*, PhD thesis, INRIA, Le Chesnay, France, 2009.

[22] G. J. Lord, C. E. Powell, and T. Shardlow, *An Introduction to Computational Stochastic PDEs*, Cambridge Texts Appl. Math. 50, Cambridge University Press, New York, 2014.

[23] J. Mandel, L. Cobb, and J. D. Beezley, *On the convergence of the ensemble Kalman filter*, Appl. Math., 56 (2011), pp. 533–541.
[24] A. Mandelbaum, *Linear estimators and measurable linear transformations on a Hilbert space*, Z. Wahrscheinlichkeitstheorie Verwandte Gebiete, 65 (1984), pp. 385–397.
[25] X. Mao, *Stochastic Differential Equations and Applications*, Horwood, Chichester, England, 2007.
[26] A. Neubauer and H. K. Pikkarainen, *Convergence results for the Bayesian inversion theory*, J. Inverse Ill-posed Probl., 16 (2008), pp. 601–613.
[27] S. Reich and C. Cotter, *Probabilistic Forecasting and Bayesian Data Assimilation*, Cambridge University Press, Cambridge, 2015.
[28] C. Schillings and A. M. Stuart, *Analysis of the ensemble Kalman filter for inverse problems*, SIAM J. Numer. Anal., 55 (2017), pp. 1264–1290, https://doi.org/10.1137/16M105959X.
[29] W. Stannat, *Stability of the optimal filter for nonergodic signals-a variational approach*, in The Oxford Handbook of Nonlinear Filtering, Oxford University Press, Oxford, 2011, pp. 374–399.
[30] A. M. Stuart, *Inverse problems: A Bayesian perspective*, Acta Numer., 19 (2010), pp. 451–559.
[31] X. T. Tong, A. J. Majda, and D. Kelly, *Nonlinear Stability of the Ensemble Kalman Filter with Adaptive Covariance Inflation*, preprint, arXiv:1507.08319, 2015.
[32] X. T. Tong, A. J. Majda, and D. Kelly, *Nonlinear stability and ergodicity of ensemble based Kalman filters*, Nonlinearity, 29 (2016), pp. 657–691, http://stacks.iop.org/0951-7715/29/i=2/a=657.