Combined Error Estimates for Local Fluctuations of SPDEs

Christian Kuehn[1] and Patrick Kürschner[2,3]

[1] *TU Munich, Faculty of Mathematics Boltzmannstr. 3 85748 Garching, Germany*
[2] *KU Leuven Kulak, Group Science, Engineering and Technology, E. Sabbelaan 53, 8500 Kortrijk, Belgium and KU Leuven, Department of Electrical Engineering ESAT/STADIUS, Kasteelpark Arenberg 10, 3001 Leuven, Belgium*
[3] *Corresponding author:* patrick.kurschner@kuleuven.be*, tel: +3216194224*

February 28, 2020

Abstract

In this work, we study the numerical approximation of local fluctuations of certain classes of parabolic stochastic partial differential equations (SPDEs). Our focus is on effects for small spatially-correlated noise on a time scale before large deviation effects have occurred. In particular, we are interested in the local directions of the noise described by a covariance operator. We introduce a new strategy and prove a Combined ERror EStimate (CERES) for the five main errors: the spatial discretization error, the local linearization error, the noise truncation error, the local relaxation error to steady state, and the approximation error via an iterative low-rank matrix algorithm. In summary, we obtain one CERES describing, apart from modelling of the original equations and standard round-off, all sources of error for a local fluctuation analysis of an SPDE in one estimate. To prove our results, we rely on a combination of methods from optimal Galerkin approximation of SPDEs, covariance moment estimates, analytical techniques for Lyapunov equations, iterative numerical schemes for low-rank solution of Lyapunov equations, and working with related spectral norms for different classes of operators.

Keywords: stochastic partial differential equation, stochastic dynamics, combined error estimates, optimal regularity, Lyapunov equation, low-rank approximation, local fluctuations.

1 Introduction

This work has two main goals. The first - more abstract - goal is to establish a general strategy to find and prove Combined ERror EStimates (CERES) for dynamical systems involving several sources of error. The second - more specific - goal is to demonstrate CERES

for a concrete challenge of an infinite-dimensional stochastic problem. The technically precise formulation of our work starts in Section 2. In this introduction we provide a basic overview of our strategy and our main results. We focus on the evolution equation

$$\mathrm{d}u = [Au + f(u)]\ \mathrm{d}t + g(u)\ \mathrm{d}W, \tag{1}$$

where $W = W(x,t)$ is a stochastic Wiener process, A is a suitable linear operator, f, g are given maps and $u = u(x,t) \in \mathbb{R}$ is the unknown function [49]. As a paradigmatic example, one may think of the Laplacian $A = \Delta$, W as a Q-Wiener process with trace-class operator Q and f, g as sufficiently smooth Lipschitz functions. Suppose the deterministic problem, i.e. $g(u) \equiv 0$, has a steady state solution $u = u^*$, which solves $0 = Au^* + f(u^*)$. Suppose the steady state is linearly stable, which just means that the spectrum of the linear operator

$$\tilde{A} := A + \mathrm{D}_u f(u^*) \tag{2}$$

is properly contained in the left-half of the complex plane [32, 51]. Note that since f is scalar-valued, we may also write $\mathrm{D}_u f(u^*) = f'(u^*)\mathrm{Id}$ so that if A is symmetric then so is $\tilde{A}$. Assume that the initial condition $u(x,0)$ is very close to u^* and the noise in (1) is sufficiently small, $0 < \|g(u)\| =: \psi \leq 1$, in comparison to the spectral gap of $\tilde{A}$ to the imaginary axis. Then the probability is extremely large that one observes only local fluctuations of sample paths of the SPDE (1) near u^* on very long time scales. Only when the time scale reaches roughly order $\mathcal{O}(\mathrm{e}^{c/\psi^2})$ for some constant $c > 0$ as $\psi \to 0$, large deviation effects [59, 49] occur. Here we focus on the initial scale of fluctuations, which we refer to as sub-exponential scale, on which large deviation effects do not play a role. However, the noise does still play an important role near the steady state. Its interplay with the operator $\tilde{A}$ determines the directions, in which we are going to find the process with higher probability locally near u^*. This raises the question, how to numerically compute these directions. A recent practical strategy suggested in the context of a numerical continuation [33] framework for stochastic ordinary differential equations (SODEs) [36] and then extended for SPDEs [37] is to:

(S1) spatially discretize u with approximation level h and consider the resulting SODEs for $u_h \in \mathbb{R}^N$. Note that we are going to view u_h as a vector but also use the notation for the associated function expressed via the basis of a finite-dimensional spatial approximation space;

(S2) locally linearize the SODEs around $u_h^* \approx u^*$ and consider SODEs for the linear approximation $\tilde{U}_h \in \mathbb{R}^N$, which form an Ornstein-Uhlenbeck (OU) process [19];

(S3) truncate the noise term based upon the decay of the Q-Wiener process and consider a reduced linear OU process $U_h \in \mathbb{R}^N$;

(S4) take the covariance matrix $V_h = V_h(t)$ of the OU process U_h, note that V_h satisfies a time-dependent Lyapunov equation [1, 54], and show that V_h converges quickly in time to a stationary Lyapunov equation for a matrix $V_* \in \mathbb{R}^{N\times N}$;

(S5) compute a low-rank approximation $V_* \approx \mathcal{Z}\mathcal{Z}^\top$ using a specialized iterative method to generate $\mathcal{Z} \in \mathbb{R}^{N\times\mathfrak{r}}$ with $\mathfrak{r} \ll N$; j computation steps of an iterative algorithm yield a matrix $V_j \approx V_*$.

Every step (S1)-(S5) produces an error, i.e., the final matrix V_j only provides an approximation to the infinite-dimensional covariance operator $\mathrm{Cov}(u)$ [21], which is precisely the operator describing the different fluctuation directions. In summary, one should aim for a result of the form

$$\sup_{t\in[0,T]} \|\mathrm{Cov}(u) - V_j\| \leq \overset{\text{error in}}{\underset{\text{Step (S1)}}{}} + \overset{\text{error in}}{\underset{\text{Step (S2)}}{}} + \overset{\text{error in}}{\underset{\text{Step (S3)}}{}} + \overset{\text{error in}}{\underset{\text{Step (S4)}}{}} + \overset{\text{error in}}{\underset{\text{Step (S5)}}{}} \tag{3}$$

to really judge the quality from the viewpoint of numerical analysis. In (3) and similar comparison problems we are always going to view finite-dimensional operators such as V_j as infinite-dimensional operators by using an embedding via a basis of the function space on which $\mathrm{Cov}(u)$ is defined. Equation (3) is just a prototypical example, i.e., a chain of different error terms does occur in most challenging high-dimensional problems, particularly those involving stochastic aspects.

Remark: *We do not consider in* (3) *the modelling error of SPDEs of the form* (1)*. On the one hand, it is partially included in a stochastic formulation anyway, and on the other hand, it is always possible to argue in an application, whether other terms or effects matter. The second error we do not include in the CERES is the standard numerical round-off error, which is universal for a given precision.*

From a technical viewpoint, each step demands different techniques and then a combination of the different estimates. For this CERES, we decided to consider the spectral or 2-norm $\|\cdot\|_2$ on the infinite- as well as finite-dimensional levels as well as the associated derived operator norm. This simplifies computing a CERES considerably as one may use the standard triangle inequality and suitable embeddings for the expression

$$\| \underbrace{\mathrm{Cov}(u) - \mathrm{Cov}(u_h)}_{\text{Step (S1)}} + \underbrace{\mathrm{Cov}(u_h) - \mathrm{Cov}(\tilde{U}_h)}_{\text{Step (S2)}} + \underbrace{\mathrm{Cov}(\tilde{U}_h) - V_h}_{\text{Step (S3)}} + \underbrace{V_h - V_*}_{\text{Step (S4)}} + \underbrace{V_* - V_j}_{\text{Step (S5)}} \|. \tag{4}$$

For (S1), we rely on an extension to covariance operators of optimal error estimates for Galerkin finite elements methods for SPDEs [34, 35]. (S2) is treated via small noise approximation in combination with moment equations [55, 38]. (S3) is covered by standard growth estimates for Q-Wiener processes over a finite time scale [49]. (S4) is tackled by results on spectra for Lyapunov equations [8] and decay of the time-dependent problem [27]. (S5) requires a careful tracing of error estimates for low-rank versions of iterative algorithms, such as alternating direction implicit (ADI) [44] and rational Krylov methods [14].

Our final result for (3) is summarized in Theorem 9.1. It illustrates that many factors can influence the error. For example, the spatial resolution h, the final time T, the Lipschitz constants of f, g, the structure of the operator Q, the spectrum of $\tilde{A}$, the noise truncation level R, and the low-rank $\mathfrak{r}$ all appear in some form in the final error. Therefore, *balancing* a CERES is the key practical message of our work. Just making a spatial resolution h small or a dynamical error small by taking higher-order terms into account may not be enough in practice, i.e., one has to be aware, which error term dominates a CERES.

We highlight that extensive numerical continuation calculations, practical convergence tests, as well as large-scale examples already exist for the method to approximate local fluctuations numerically as proposed here. We refer the interested reader to [36] for an

introduction and the SODE case, to [37] for the SPDE case including precise numerical comparisons to theoretical scaling laws, and to [3] for a successful large-scale application in geoscience. In this work, we focus on the theoretical aspects, thereby finishing/completing the previous work on scientific computing side of the method.

The paper is structured as follows: In Section 2 we provide the foundational technical setup for the SPDE (1) using mild solutions and in Section 3 we cover the spatial discretization. Section 4 develops the error estimates for covariance matrices while Section 5 contains the relevant moment estimates. Then we discuss the noise truncation in Section 6. In Section 7, we transition to the Lyapunov equation and its reduction to steady state. The last technical step is carried out in Section 8 tracing the error results for low-rank iterative schemes for Lyapunov equations. In Section 9, we present the full CERES, which draws upon all the previous results. An outlook to open problems and further applications of our methodology is given in Section 10.

2 SPDE - Mild Solutions

Consider a fixed compact time interval $\mathcal{I} = [0, T]$, $t \in \mathcal{I}$ and a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in\mathcal{I}}, \mathbb{P})$. Furthermore, we fix two Hilbert spaces $\mathcal{H}$ and $\mathcal{U}$ with inner products $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and $\langle\cdot,\cdot\rangle_{\mathcal{U}}$ respectively. As concrete cases/examples, we always want to think of the Hilbert spaces being spatial function spaces over a bounded domain $\mathcal{D} \subset \mathbb{R}^d$ with smooth boundary $\partial\mathcal{D}$ and with spatial variable by $x \in \mathcal{D}$ but we phrase our results in a more abstract evolution equation setting. Let $Q \in \mathcal{L}(\mathcal{U},\mathcal{U})$ be a symmetric non-negative linear operator and let $(W(t))_{t\in\mathcal{I}}$ denote the associated Q-Wiener process; see [49] for the construction of $(W(t))_{t\in\mathcal{I}}$. Let $\mathcal{U}_0 := Q^{1/2}\mathcal{U}$ be a Hilbert pace [49, 50] with the inner product $\langle\cdot,\cdot\rangle_{\mathcal{U}_0} := \langle Q^{-1/2}\cdot, Q^{-1/2}\cdot\rangle_{\mathcal{U}}$, where $Q^{-1/2}$ is the Moore-Penrose pseudoinverse of Q. An operator $M : \mathcal{U}_0 \to \mathcal{H}$ is a Hilbert-Schmidt operator if the norm

$$\|M\|_{\mathcal{L}_2^0} := \left(\sum_{k=1}^{\infty} \|M\zeta_k\|_{\mathcal{H}}^2\right)^{1/2}$$

is finite, where the choice of orthonormal basis $\{\zeta_k\}_{k=1}^{\infty}$ for $\mathcal{U}_0$ turns out to be arbitrary. The space of these Hilbert-Schmidt operators will be denoted accordingly by $\mathcal{L}_2^0$. If we consider Hilbert-Schmidt operators on $\mathcal{H}$, then they will be denoted by $\mathcal{L}_2 = \mathcal{L}_2(\mathcal{H},\mathcal{H})$.

The unknown function is $u(t) \in \mathcal{H}$. As mentioned above, it is helpful to always think of a more concrete setting when $\mathcal{H}$ is a spatial function space. Then $u(x, t; \omega)$ for $\omega \in \Omega$ denotes the unknown family of random variables $u : \mathcal{D}\times\mathcal{I}\times\Omega \to \mathbb{R}$. In the notation we shall always suppress ω from now on and assume that all maps we define are measurable with respect to ω, which will also imply measurability for u below. Setting $u(t) = u(\cdot, t) \in \mathcal{H}$ and $W(t) = W(\cdot, t)$ includes this case in the abstract setup, where we want to study the SPDE

$$\mathrm{d}u(t) = [Au(t) + f(u(t))]\ \mathrm{d}t + g(u(t))\ \mathrm{d}W(t), \qquad u(0) = u_0 \in \mathcal{H}, \tag{5}$$

as an evolution equation on the Hilbert space $\mathcal{H}$, i.e., $u(t)$ is an $\mathcal{H}$-valued random variable.

(A0) We assume that Q is of trace class. Furthermore, we assume the operator $A : \mathrm{dom}(A) \subset \mathcal{H} \to \mathcal{H}$ is linear, self-adjoint negative definite operator with compact inverse and generates an analytic semigroup $t \mapsto \mathrm{e}^{tA}$ [26].

(A0) implies that there exists an orthonormal basis of eigenvectors of A for $\mathcal{H}$. The maps f, g are going to be specified more precisely below and the initial condition $u_0 \in \mathcal{H}$ is a random variable. A mild solution $u(t)$ to (5) satisfies

$$u(t) = \mathrm{e}^{tA}u_0 + \int_0^t \mathrm{e}^{(t-s)A} f(u(s))\ \mathrm{d}s + \int_0^t \mathrm{e}^{(t-s)A} g(u(s))\ \mathrm{d}W(s), \tag{6}$$

i.e., the integral equation (6) holds $\mathbb{P}$-almost surely ($\mathbb{P}$-a.s.) for $t \in \mathcal{I}$ and

$$\mathbb{P}\left(\int_0^t |u(s)|^2\ \mathrm{d}s < +\infty\right) = 1, \qquad \mathbb{P}\text{-a.s.};$$

see also [49, Chapter 4] or [13] for the construction of the stochastic integral. It is well-known that under certain Lipschitz assumptions [49, Section 7.1] or dissipativity assumptions [49, Section 7.4.2] on f, g, there exists a unique mild solution. However, since we are interested in numerical error estimates, it is important to have optimal regularity results for mild solutions so we follow [34, 35, 30]. Denote by $\{a_k\}_{k=1}^{\infty}$, $0 > a_1 \geq a_2 \geq \cdots$ the eigenvalues of A and by e_k the associated eigenfunctions with $Ae_k = a_k e_k$. For $r \in \mathbb{R}$, define the fractional operator $A^{r/2} : \mathrm{dom}(A^{r/2}) \to \mathcal{H}$ by

$$A^{r/2}v := -(-A)^{r/2}v = -\sum_{k=1}^{\infty}(-a_k)^{r/2}\ \langle v, e_k\rangle_{\mathcal{H}}\ e_k.$$

Set $\dot{\mathcal{H}}^r := \mathrm{dom}(A^{r/2})$ and consider the norm $\|v\|_{\dot{\mathcal{H}}^r} := \|A^{r/2}v\|_{\mathcal{H}}$, which turns $\dot{\mathcal{H}}^r$ into a Hilbert space. Let $\mathcal{L}_{2,r}^0 \subset \mathcal{L}_2^0$ denote the subspace of Hilbert-Schmidt operators, which have finite norm $\|\cdot\|_{\mathcal{L}_{2,r}^0} := \|A^{r/2}\cdot\|_{\mathcal{L}_2^0}$. The following assumptions are assumed to hold from now on (although we are still going to emphasize this several times in statements of theorems below):

(A1) Fix two constants $r \in [0, 1)$ and $p \in [2, \infty)$ for all assumptions. Fix an initial condition $u_0 : \Omega \to \dot{\mathcal{H}}^{r+1}$ and assume it is $\mathcal{F}_0$-measurable with

$$\left[\mathbb{E}\left(\|u_0\|_{\dot{\mathcal{H}}^{r+1}}^p\right)\right]^{1/p} \leq C_{\mathrm{ini},r} < +\infty.$$

(A2) $f : \mathcal{H} \to \dot{\mathcal{H}}^{r-1}$ and there exists a constant $C_f > 0$ such that

$$\|f(u) - f(v)\|_{\dot{\mathcal{H}}^{r-1}} \leq C_f\|u - v\|_{\mathcal{H}}, \quad \text{for all } u, v \in \mathcal{H}. \tag{7}$$

(A3) $g : \mathcal{H} \to \mathcal{L}_2^0$ and there exists a constant $C_{g,1} > 0$ such that

$$\|g(u) - g(v)\|_{\mathcal{L}_2^0} \leq C_{g,1}\|u - v\|_{\mathcal{H}}, \quad \text{for all } u, v \in \mathcal{H}, \tag{8}$$

and furthermore $g(\dot{\mathcal{H}}^r) \subset \mathcal{L}_{2,r}^0$ holds with the estimate

$$\|g(u)\|_{\mathcal{L}_{2,r}^0} \leq C_{g,2}(1 + \|u\|_{\dot{\mathcal{H}}^r}), \quad \text{for all } u \in \mathcal{H}, \tag{9}$$

and some constant $C_{g,2} > 0$.

Essentially (A2)-(A3) are modifications/extensions of the classical Lipschitz assumptions in [49]. Therefore, one immediately gets:

Theorem 2.1 ([49, Theorem 7.4],[29, Theorem 1])**.** *Suppose (A0)-(A3) hold, then there exists a unique mild solution for the SPDE* (5)*.*

In addition, one may provide (optimal) regularity estimates for the mild solution:

Theorem 2.2 ([35, Theorem 3.1 & Theorem 4.1])**.** *Suppose (A0)-(A3) hold then the unique mild solution is almost surely in $\dot{\mathcal{H}}^{r+1}$. There exists a constant $C_{\mathrm{spa}} > 0$ such that*

$$\sup_{t\in\mathcal{I}} \left(\mathbb{E}\left[\|u(t)\|^p_{\dot{\mathcal{H}}^{r+1}}\right]\right)^{1/p} \leq C_{\mathrm{ini},r} + C_{\mathrm{spa}} \left(1 + \sup_{t\in\mathcal{I}} \left(\mathbb{E}\left[\|u(t)\|^p_{\dot{\mathcal{H}}^{r}}\right]\right)^{1/p}\right) \tag{10}$$

holds as a spatial regularity estimate and for temporal regularity we have

$$\sup_{t_1,t_2\in\mathcal{I},t_1\neq t_2} \left(\mathbb{E}\left[\|u(t_1)-u(t_2)\|^p_{\dot{\mathcal{H}}^{s}}\right]\right)^{1/p} \leq C_{\mathrm{time}}|t_1-t_2|^{\min(\frac{1}{2},\frac{1+r-s}{2})} \tag{11}$$

for some constant $C_{\mathrm{time}} > 0$ and for every $s \in [0, r+1)$.

As for the deterministic counterpart ($g \equiv 0$) of the SPDE (5), one may use regularity estimates to obtain convergence rates and error estimates for associated numerical schemes [56, 18]. In this paper, as already discussed in Section 1, we are interested in the spatial approximation error only. We stated the temporal regularity results only for completeness. It is noted that the constant $C_{\mathrm{spa}} > 0$ in Theorem 2.2 does depend upon p, r, A, f, g and T.

As a typical example for the abstract framework, one should always keep in mind classical one-component reaction-diffusion systems, where

$$\mathcal{H} = L^2(\mathcal{D}) \qquad \text{and} \qquad A = \Delta := \sum_{k=1}^{d} \frac{\partial^2}{\partial x_k^2},$$

with Dirichlet boundary conditions. We point out that for rectangular domains and A being the Laplacian, one can easily calculate the eigenfunctions of A explicitly. If one works with the standard stochastic heat equation, i.e., sets $f = 0$ and takes additive noise, then all of our estimates below easily become explicit in a spectral Galerkin setting. However, we shall not restrict to this special case in this paper and continue to work with the more general setup. In fact, it is generally quite difficult to give any *explicit* expressions regarding numerical error estimates for nonlinear SPDEs, even on rectangular domains.

3 SPDE - Spatial Discretization

Let $\mathcal{S}_h$ for $h \in (0, 1]$ denote a continuous family of finite-dimensional subspaces of the Hilbert space $\dot{\mathcal{H}}^1$ with

$$\dim(\mathcal{S}_h) =: N \in \mathbb{N}, \tag{12}$$

which are spanned by N basis elements of a basis of $\dot{\mathcal{H}}^1$. We always assume that in the limit $h \to 0$ we indeed obtain a full basis of $\dot{\mathcal{H}}^1$. The spaces $\mathcal{S}_h$ should be thought of as the spatial discretization spaces; see [56, 18] for a detailed overview. Particularly important examples are the span of a finite set of basis functions of A leading to a spectral Galerkin method, or the span of piecewise polynomial functions on $\mathcal{D}$, where h is the diameter of the largest element of a suitable mesh on $\mathcal{D}$, leading to a Galerkin finite element method. Let $R_h : \dot{\mathcal{H}}^1 \to \mathcal{S}_h$ be the orthogonal projection onto $\mathcal{S}_h$ with respect to the inner product $a(\cdot,\cdot) := \langle A^{1/2}\cdot, A^{1/2}\cdot\rangle_{\mathcal{H}}$ so that we have

$$a(R_h u, v_h) = a(u, v_h), \quad \text{for all } u \in \dot{\mathcal{H}}^1,\ v_h \in \mathcal{S}_h.$$

The discretized version A_h of A is defined by the requirement that for a given $u_h \in \mathcal{S}_h$, the image $A_h u_h$ is the unique element satisfying

$$a(u_h, v_h) = \langle A_h u_h, v_h\rangle_{\mathcal{H}}, \quad \text{for all } v_h \in \mathcal{S}_h.$$

Furthermore, let $P_h : \dot{\mathcal{H}}^{-1} \to \mathcal{S}_h$ be the generalized orthogonal projection onto $\mathcal{S}_h$ defined analogously to R_h, i.e., by requiring

$$\langle P_h u, v_h\rangle_{\mathcal{H}} = a(A^{-1}u, v_h), \quad \text{for all } u \in \dot{\mathcal{H}}^{-1},\ v_h \in \mathcal{S}_h.$$

Then the spatial discretization of the SPDE can be written as

$$\mathrm{d}u_h(t) = [A_h u_h(t) + P_h f(u_h(t))]\ \mathrm{d}t + P_h g(u_h(t))\ \mathrm{d}W(t), \qquad u_h(0) = P_h u_0. \tag{13}$$

It is relatively straightforward to check that (13) must also have a unique mild solution; see also the proof of Lemma 3.2. To fix the role of the discretization parameter h, we are going to assume:

(A4) There exists a constant $C_h > 0$ such that

$$\|R_h v - v\|_{\mathcal{H}} \le C_h\ h^s\ \|v\|_{\dot{\mathcal{H}}^s}, \quad \text{for all } v \in \dot{\mathcal{H}}^s,\ s \in \{1,2\},\ h \in (0,1]. \tag{14}$$

Recently, the regularity result of Theorem 2.2 has been transferred to yield results on strong/pathwise approximation properties of the approximating stochastic evolution equation (13) for the SPDE (5). Consider the norm

$$\|\cdot\|_{L^p(\Omega;\mathcal{H})} := (\mathbb{E}[\|\cdot\|_{\mathcal{H}}^p])^{1/p} \tag{15}$$

for $p \in [2, +\infty)$ as above. Then one can prove the following error estimate:

Theorem 3.1 ([34, Theorem 1.1])**.** *Suppose (A0)-(A4) hold, then there exists a constant* $C_{\mathrm{Gal}} > 0$ *such that*

$$\|u(t) - u_h(t)\|_{L^p(\Omega;\mathcal{H})} \le C_{\mathrm{Gal}}\ h^{1+r}, \qquad \forall t \in \mathcal{I}. \tag{16}$$

In the later development of the error estimates for the covariance operator, we shall need another auxiliary result, which we prove here:

Lemma 3.2. *Suppose (A0)-(A4) hold, then here exists a constant $C_+ > 0$ (independent of h) such that*

$$\sup_{t\in\mathcal{I}} \|u(t) + u_h(t)\|_{L^2(\Omega;\mathcal{H})} \leq C_+.$$

Proof. Indeed, we just have

$$\sup_{t\in\mathcal{I}} \|u(t) + u_h(t)\|_{L^2(\Omega;\mathcal{H})} \leq \sup_{t\in\mathcal{I}} \|u(t)\|_{L^2(\Omega;\mathcal{H})} + \sup_{t\in\mathcal{I}} \|u_h(t)\|_{L^2(\Omega;\mathcal{H})}$$

so the first term is bounded by a direct application of Theorem 2.2 (or in fact, the classical result [49, Theorem 7.4(ii)]). For the discretization, note that we may apply the same results since (A0)-(A3) also hold for the discretized SPDE (13). For example, consider (A1) then we have

$$\|P_h f(u) - P_h f(v)\|_{\dot{\mathcal{H}}^{r-1}} = \|P_h[f(u) - f(v)]\|_{\dot{\mathcal{H}}^{r-1}} \leq \|f(u) - f(v)\|_{\dot{\mathcal{H}}^{r-1}} \tag{17}$$

since P_h is the generalized orthogonal projector. The other assumptions are checked similarly. □

Of course, it should be noted that the constant $C_+ > 0$ will depend on the data of the problem, i.e., on f, g, T, A. However, we will only be interested in the convergence rate in h and we will show later on that we can select T as an order one constant anyhow, while f, g, A are given and satisfy the Lipschitz and semigroup assumptions stated above. For more background on discretization of SPDEs, we also refer to [28, 45].

4 SPDEs - Covariance

The next step is to establish numerical error estimates for the covariance. Let $v \in L^2(\Omega;\mathcal{H})$, then one defines the covariance operator [49, 41] of v as

$$\mathrm{Cov}(v) := \mathbb{E}[(v - \mathbb{E}[v]) \otimes (v - \mathbb{E}[v])]. \tag{18}$$

By definition, $\mathrm{Cov}(v) : \mathcal{H} \to \mathcal{H}$ is a symmetric linear operator. In addition one may check that $\mathrm{Cov}(v)$ is nuclear using the equivalent characterization of nuclear operators M on Hilbert spaces via the condition

$$\mathrm{Tr}(M) := \sum_{k=1}^{\infty} \langle M\xi_k, \xi_k\rangle_{\mathcal{H}} < +\infty$$

for an orthonormal basis $\{\xi_k\}_{k=1}^{\infty}$ of $\mathcal{H}$. The space of nuclear operators $\mathcal{L}_1(\mathcal{H},\mathcal{H}) =: \mathcal{L}_1$ becomes a Banach space under the norm $\|\cdot\|_{\mathcal{L}_1(\mathcal{H},\mathcal{H})} := \mathrm{Tr}(\cdot)$ [49, Appendix C]. Note that we have two well-defined covariance operators

$$\mathrm{Cov}(u(t)) \qquad \text{and} \qquad \mathrm{Cov}(u_h(t)) \tag{19}$$

as the SPDE (5) and the spatially discretized version (13) both have mild solutions in $L^2(\Omega \times \mathcal{I};\mathcal{H})$.

Theorem 4.1. *There exists a constant $C_{\mathrm{d}} > 0$ such that*

$$\sup_{t\in\mathcal{I}} \|\mathrm{Cov}(u(t)) - \mathrm{Cov}(u_h(t))\|_{\mathcal{L}_2(\mathcal{H},\mathcal{H})} \leq C_{\mathrm{d}} h^{1+r}. \tag{20}$$

Proof. The proof is a calculation aiming to use the spatial approximation property of Theorem 3.1. Suppose as a first case that $\mathbb{E}[u] = 0 = \mathbb{E}[u_h]$. Consider an orthonormal basis $\{\xi_k\}_{k=1}^{\infty}$ of $\mathcal{H}$. We want to estimate the error in the Hilbert-Schmidt norm $\|\cdot\|_{\mathcal{L}_2}$. For the following steps we suppress the argument $u = u(t)$ and $u_h = u_h(t)$ and use the definition of the covariance operator

$$\begin{aligned}
\sup_{t\in\mathcal{I}} \|\mathrm{Cov}(u) - \mathrm{Cov}(u_h)\|_{\mathcal{L}_2}^2 &= \sup_{t\in\mathcal{I}} \|\mathbb{E}[u\otimes u] - \mathbb{E}[u_h\otimes u_h]\|_{\mathcal{L}_2}^2, \\
&\leq \sup_{t\in\mathcal{I}} \|\mathbb{E}[u\otimes u] - \mathbb{E}[u_h\otimes u_h]\|_{\mathcal{L}_1}^2,
\end{aligned}$$

where we used that the trace-class norm bounds the Hilbert-Schmidt norm. Hence, we have using that off-diagonal terms cancel in the trace-class norm that

$$\begin{aligned}
\sup_{t\in\mathcal{I}} \|\mathrm{Cov}(u) - \mathrm{Cov}(u_h)\|_{\mathcal{L}_2}^2 &\leq \sup_{t\in\mathcal{I}} \sum_{k=1}^{\infty} \langle \mathbb{E}[(u-u_h)\otimes(u+u_h)]\xi_k, \xi_k\rangle_{\mathcal{H}}^2, \\
&= \sup_{t\in\mathcal{I}} \sum_{k=1}^{\infty} \langle \mathbb{E}[((u-u_h)\otimes(u+u_h))\xi_k], \xi_k\rangle_{\mathcal{H}}^2, \\
&= \sup_{t\in\mathcal{I}} \sum_{k=1}^{\infty} \mathbb{E}[\langle u-u_h, \xi_k\rangle_{\mathcal{H}} \langle u+u_h, \xi_k\rangle_{\mathcal{H}}]^2,
\end{aligned}$$

where we use in the last step the definition of the tensor product $(v_1\otimes v_2)v_3 := \langle v_2, v_3\rangle_{\mathcal{H}} v_1$ for $v_k \in \mathcal{H}$ for $k \in \{1,2,3\}$. Then a direct application of Cauchy-Schwarz yields

$$\begin{aligned}
\sup_{t\in\mathcal{I}} \|\mathrm{Cov}(u) - \mathrm{Cov}(u_h)\|_{\mathcal{L}_2}^2 &\leq \sup_{t\in\mathcal{I}} \sum_{k=1}^{\infty} \mathbb{E}[\langle u-u_h, \xi_k\rangle_{\mathcal{H}}^2] \mathbb{E}[\langle u+u_h, \xi_k\rangle_{\mathcal{H}}^2], \\
&\leq \sup_{t\in\mathcal{I}} \|u-u_h\|_{L^2(\Omega;\mathcal{H})}^2 \ \|u+u_h\|_{L^2(\Omega;\mathcal{H})}^2. \qquad (21)
\end{aligned}$$

In the expression (21), we may estimate the two terms separately by estimating the supremum by the product of suprema. Furthermore, a direct application of Theorem 3.1 to the first term and Lemma 3.2 to the second term give

$$\sup_{t\in\mathcal{I}} \|\mathrm{Cov}(u) - \mathrm{Cov}(u_h)\|_{\mathcal{L}_2}^2 \leq C_{\mathrm{Gal}}\ C_+\ h^{2(r+1)},$$

which yields the result in the basic case of zero means. If $\mathbb{E}[u] \neq 0$ and $\mathbb{E}[u_h] \neq 0$, we observe that

$$\begin{aligned}
\sup_{t\in\mathcal{I}} \|u - u_h - \mathbb{E}[u-u_h]\|_{L^2(\Omega;\mathcal{H})}^2 &\leq \sup_{t\in\mathcal{I}} \|u-u_h\|_{L^2(\Omega;\mathcal{H})}^2 + \sup_{t\in\mathcal{I}} \|\mathbb{E}[u-u_h]\|_{L^2(\Omega;\mathcal{H})}^2, \\
&\leq C_{\mathrm{Gal}}\ h^{2(r+1)} + \sup_{t\in\mathcal{I}} \mathbb{E}[\|u-u_h\|_{L^2(\Omega;\mathcal{H})}]^2, \\
&\leq 2C_{\mathrm{Gal}}\ h^{2(r+1)},
\end{aligned}$$

using again Theorem 3.1 twice. Furthermore, it is easy to see that

$$\mathbb{E}[\|u + u_h - \mathbb{E}[u + u_h]\|_{L^2(\Omega;\mathcal{H})}]$$

remains bounded for nonzero means. The result now follows repeating the same steps shown for the zero means case also for the nonzero means case. □

Obviously the constant C_{d} also depends upon the data of the problem as do C_+ and C_{Gal} but all constants are independent of h, which is the key discretization parameter in the step (S1).

5 SODEs - Linearization

Having estimated the error of the spatial discretization, we are now dealing with

$$\mathrm{d}u_h(t) = [A_h u_h(t) + P_h f(u_h(t))]\ \mathrm{d}t + P_h g(u_h(t))\ \mathrm{d}W(t), \qquad u_h(0) = P_h u_0, \tag{22}$$

where $u_h \in \mathbb{R}^N$ is a finite-dimensional approximation of u. We assume that the original SPDE (5) and its projection have a locally asymptotically stable homogeneous steady state for zero noise and that we only study the small noise regime with sufficiently fast decay for the eigenvalues of Q and a noise with $\mathcal{U} = \mathcal{H}$.

(A5) Assume u^* satisfies $Au^* + f(u^*) = 0$. Furthermore, suppose f is Fréchet differentiable, $\mathrm{D}_u f(u^*) = f'(u^*)\mathrm{Id}$ and

$$\mathrm{spec}(A + f'(u^*)\mathrm{Id}) \subset \{\rho \in \mathbb{R} : \rho < 0\}. \tag{23}$$

Furthermore, assume $u_h^* := P_h u^*$ satisfies $A_h u_h^* + P_h f(u_h^*) = 0$ for all $h \in (0, 1]$.

(A6) Suppose there exists a constant $\psi > 0$ such that

$$\|P_h g(u_h(t))\|_2 \leq \psi \tag{24}$$

for all $h \in (0, 1]$, where we recall that we use $\|\cdot\|_2$ to also denote the usual Euclidean norm.

Below we are also going to assume that ψ is chosen sufficiently small to get a good approximation of the linearized system. The goal is to provide a finite-time estimate for the difference between the covariance matrix $\mathrm{Cov}(u_h(t))$ of (22) and covariance matrix $\mathrm{Cov}(U_h(t))$ of the linearized OU process

$$\mathrm{d}\tilde{U}_h(t) = [A_h + P_h[\mathrm{D}_u f](u^*)]\,\tilde{U}_h(t)\ \mathrm{d}t + P_h g(u^*)\ \mathrm{d}W(t), \qquad \tilde{U}_h(0) = P_h u_0, \tag{25}$$

where $\mathrm{D}_u f$ denotes the usual Fréchet derivative as introduced already above. Note that the linear operator $P_h g(u^*)$ only acts nontrivially on the first N basis elements of $\mathcal{S}_h$. In the case

of spectral Galerkin, these basis elements are $\{e_i\}_{i=1}^N$ and so $P_h g(u^*)$ is zero on $\{e_i\}_{i=N+1}^\infty$. In this case, we can replace $W(t)$ by

$$W^N(t) = \sum_{i=1}^N \sqrt{\lambda_{Q,i}}\beta_i(t)e_i \tag{26}$$

where we assume that $\{e_i\}_{i=1}^\infty$ are also eigenfunctions of Q, $\{\lambda_{Q,i}\}_{i=1}^\infty$ are eigenvalues of Q, and $\{\beta_i(t)\}_{i=1}^\infty$ are independent Brownian motions. For a more general space $\mathcal{S}_h$, let $\{\kappa_k\}_{k=1}^N$ be its basis and note that

$$\begin{aligned} P_h g(u^*)W(t) &= P_h g(u^*)\sum_{i=1}^\infty \sqrt{\lambda_{Q,i}}\beta_i(t)e_i \\ &= P_h g(u^*)\sum_{i=1}^\infty \sqrt{\lambda_{Q,i}}\beta_i(t)\sum_{k=1}^N \langle e_i,\kappa_k\rangle\kappa_k \\ &= P_h g(u^*)\left[\sum_{i=1}^{N_0} \sqrt{\lambda_{Q,i}}\beta_i(t)\sum_{k=1}^N \langle e_i,\kappa_k\rangle\kappa_k + \sum_{i=N_0}^\infty \sqrt{\lambda_{Q,i}}\beta_i(t)\sum_{k=1}^N \langle e_i,\kappa_k\rangle\kappa_k\right]. \end{aligned}$$

Hence, up to a change of basis, the first summand in the last expression is of the same form as (26) up to a change of basis, while the second term can be made arbitrarily small with a suitable choice on N_0 due to the trace class assumption; see Section 6, where we estimate the noise truncation error explicitly. To not overload the notation, we shall work from now on with the spectral formulation (26).

Recall that without additional assumptions on the nonlinearity of f, it is usually not possible to estimate the error between a linearized and a nonlinear system, even on a finite time scale. To simplify the notation we let

$$u_h =: z, \qquad A_h u_h(t) + P_h f(u_h(t)) =: F(z), \qquad P_h g(u_h(t)) =: G(z),$$

as well as

$$\tilde{U}_h =: \tilde{Z}, \qquad [A_h + P_h[\mathrm{D}_u f](u^*)] =: \mathcal{A}, \qquad P_h g(u^*) =: \tilde{B},$$

where $G(z)$ and $\tilde{B}$ are operators projected/restricted onto the first N basis functions. Hence, we have to compare the SODE

$$\mathrm{d}z = F(z)\ \mathrm{d}t + G(z)\ \mathrm{d}W^N, \qquad z(0) = z_0, \tag{27}$$

near a steady state $z^* := u_h^*$ to the SODE

$$\mathrm{d}\tilde{Z} = \mathcal{A}\tilde{Z}\ \mathrm{d}t + \tilde{B}\ \mathrm{d}W^N, \qquad \tilde{Z}(0) = \tilde{Z}_0. \tag{28}$$

Without loss of generality we may assume that $z^* \equiv (0,\dots,0)^\top$ since we can always translate the steady state if necessary. Let $\mathbf{p} = (p_1, p_2, \dots, p_N) \in (\mathbb{N}_0)^N$ be a multi-index, define the mean values of z by $\mu := \mathbb{E}[z] \in \mathbb{R}^N$ and the centered moments as

$$\mathbb{E}[(z-\mu)^{\mathbf{p}}] := \mathbb{E}[(z_1-\mu_1)^{p_1}(z_2-\mu_2)^{p_2}\cdots(z_N-\mu_N)^{p_N}]. \tag{29}$$

To make the notation more compact, we also introduce 'altered' multi-indices as follows:

$$\mathbf{p}(k:\zeta) = (p_1, p_2, \dots, p_{k-1}, p_k + \zeta, p_{k+1}, \dots, p_N)$$

for $\zeta \in \mathbb{Z}$ and multiple arguments pertain to changes in the respective components. Furthermore, we consider the diffusion operators

$$\mathcal{G}(z) := G(z)G(z)^\top, \qquad \tilde{\mathcal{B}} := \tilde{B}\tilde{B}^\top. \tag{30}$$

Lemma 5.1. *The evolution equations for the centered moments* $\mathbb{E}[(z-\mu)^{\mathbf{p}}]$ *of* (27) *are given by*

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[(z-\mu)^{\mathbf{p}}] &= \sum_{k=1}^{N} p_k \mathbb{E}[F_k(z)(z-\mu)^{\mathbf{p}(k:-1)}] + \frac{1}{2}\sum_{k=1}^{N} p_k(p_k-1)\mathbb{E}[\mathcal{G}_{kk}(z)(z-\mu)^{\mathbf{p}(k:-2)}] \\ &\quad + \sum_{l=2}^{N}\sum_{k=1}^{l-1} p_l p_k \mathbb{E}[\mathcal{G}_{kl}(z)(z-\mu)^{\mathbf{p}(k:-1,l:-1)}]\end{aligned}$$

Proof. The calculation of the SODEs for the moments follows from first applying Itô's formula to monomials of the form $z^{\mathbf{p}}$ as shown in [55, Section 4.1]. However, we then need to average via $\mathbb{E}[\cdot]$, and terms of the form $\int(\cdot)\,\mathrm{d}W^N$ average to zero as they satisfy the martingale property [31] by the Lipschitz assumption on G. □

Consider the linear approximation of (27) given by (28). We use the notation $\nu := \mathbb{E}[\tilde{Z}]$ and we may assume without loss of generality that $\mathcal{A}$ is already diagonal; indeed, by assumptions (A0) and (A5) we already have that $\mathcal{A}$ is symmetric with real spectrum so we can apply a coordinate change to make $\mathcal{A}$ diagonal, which will just change constants in the estimates so we do not display this explicitly here. Now we have two processes and we would like to compare $\mathrm{Cov}(z)$ with $\mathrm{Cov}(\tilde{Z})$. From Lemma 5.1 it follows that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mu = \mathbb{E}[F(z)], \qquad \frac{\mathrm{d}}{\mathrm{d}t}\nu = \mathcal{A}\nu. \tag{31}$$

It is relatively easy to write down the formal evolution equations for the covariances. For the diagonal entries we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Cov}(z)_{ii} = 2\mathbb{E}[F_i(z)(z_i-\mu_i)] + \mathbb{E}[\mathcal{G}_{ii}(z)] \tag{32}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Cov}(Z)_{ii} = 2\mathbb{E}[(\mathcal{A}\tilde{Z})_i(\tilde{Z}_i-\nu_i)] + \mathcal{B}_{ii} \tag{33}$$

Consider the difference $\mathrm{Cov}(z) - \mathrm{Cov}(\tilde{Z}) =: \mathrm{Cov}^\Delta$ and also define the remainders

$$R^F(z) := F(z) - \mathcal{A}z, \qquad R^G(z) := G(z) - \tilde{\mathcal{B}}.$$

Observe that the remainder R^F is Lipschitz

$$\|R^F(z_1) - R^F(z_2)\| \le (C_F + \|\mathcal{A}\|)\|z_1 - z_2\|, \tag{34}$$

by assumptions (A2) and (A5) for some constant $C_F > 0$, where $\|\cdot\|$ always denotes the 2-norm. Regarding the remainder R^G, recall that we assumed a uniform noise bound in (A6) so that

$$\|R^G(z_1) - R^G(z_2)\| \leq C_G \psi^2 \tag{35}$$

for some constant $C_G > 0$. Then one finds, using that $\mathcal{A}$ is diagonal and via assumption (A5), that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Cov}_{ii}^{\Delta} &= 2\mathbb{E}[F_i(z)(z_i - \mu_i) - (\mathcal{A}\tilde{Z})_i(\tilde{Z}_i - \nu_i)] + \mathbb{E}[G_{ii}(z) - \mathcal{B}_{ii}], \\
&= 2\mathcal{A}_{ii}\mathbb{E}[(z_i + \mu_i - \mu_i - R_i^F(z)/\mathcal{A}_{ii})(z_i - \mu_i) - (\tilde{Z}_i + \nu_i - \nu_i)(\tilde{Z}_i - \nu_i)] + \mathbb{E}[R_{ii}^G(z)], \\
&= 2\mathcal{A}_{ii}\left(\mathrm{Cov}_{ii}^{\Delta} + \mathbb{E}[(\mu_i - R_i^F(z)/\mathcal{A}_{ii})(z_i - \mu_i) - \nu_i(\tilde{Z}_i - \nu_i)]\right) + \mathbb{E}[R_{ii}^G(z)], \\
&= 2\mathcal{A}_{ii}\left(\mathrm{Cov}_{ii}^{\Delta} + \mathbb{E}[R_i^F(z)/\mathcal{A}_{ii}(\mu_i - z_i)]\right) + \mathbb{E}[R_{ii}^G(z)].
\end{aligned}$$

Using the Lipschitz conditions (34) and the bound (35) one has

$$\begin{aligned}
|\mathbb{E}[R_i^F(z)/\mathcal{A}_{ii}(\mu_i - z_i) + R_{ii}^G(z)]| \;\leq\; & \frac{(\|\mathcal{A}\| + C_F)|\mu_i|}{|\mathcal{A}_{ii}|}\mathbb{E}[\|z\|] \\
& + \frac{(\|\mathcal{A}\| + C_F)}{|\mathcal{A}_{ii}|}\mathbb{E}[\|z\|^2] + C_G\psi^2.
\end{aligned} \tag{36}$$

We denote the right-hand side of the last inequality by $\eta_{ii}(t)$, which implies the final estimate

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Cov}_{ii}^{\Delta}(t) \leq \eta_{ii}^* + 2\mathcal{A}_{ii}\mathrm{Cov}_{ii}^{\Delta}(t), \qquad \eta_{ii}^* := \max_{t\in\mathcal{I}} \eta_{ii}(t). \tag{37}$$

Considering the coordinate shift $\mathrm{Cov}_{ii}^{\Delta}(t) = C_{ii}^{\Delta}(t) - \eta_{ii}^*/2\mathcal{A}_{ii}$ yields

$$\frac{\mathrm{d}}{\mathrm{d}t}C_{ii}^{\Delta}(t) \leq 2\mathcal{A}_{ii}C_{ii}^{\Delta}(t). \tag{38}$$

Applying Gronwall's inequality to (38), transforming back into original coordinate frame, and using $\eta_{ii}^*/(2\mathcal{A}_{ii}) < 0$ yields the following result:

Lemma 5.2. *Suppose the assumptions (A0)-(A6) hold for all $t \in \mathcal{I}$, then*

$$\mathrm{Cov}_{ii}^{\Delta}(t) \leq -\frac{\eta_{ii}^*}{2\mathcal{A}_{ii}} + \left[\mathrm{Cov}_{ii}^{\Delta}(0) + \frac{\eta_{ii}^*}{2\mathcal{A}_{ii}}\right]\mathrm{e}^{2\mathcal{A}_{ii}t}. \tag{39}$$

holds for all $t \in \mathcal{I}$.

Note that an estimate of the form (39) is fully expected to hold since it states that the growth of the difference between the covariances in the case of Lipschitz F and sufficiently bounded noise is controlled by the first- and second-moments of the nonlinear process. In particular, if we are close to a linear SODE or the spectral gap is very large, then we have an excellent finite-time approximation on $\mathcal{I}$, while large noise and a strong nonlinearity make

the approximation worse. The next step is to look at the off-diagonal terms and consider the case $i > j$ (the case $i < j$ is similar). The evolution equations are

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Cov}(z)_{ij} &= \mathbb{E}[F_i(z)(z_j-\mu_j)] + \mathbb{E}[F_j(z)(z_i-\mu_i)] + \mathbb{E}[G_{ij}(z)] \\ \frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Cov}(\tilde{Z})_{ij} &= \mathbb{E}[(\mathcal{A}\tilde{Z})_i(\tilde{Z}_j-\nu_j)] + \mathbb{E}[(\mathcal{A}\tilde{Z})_j(\tilde{Z}_i-\nu_i)] + \mathcal{B}_{ij}\end{aligned}$$

A similar calculation as above leads one to define

$$\begin{aligned}\eta_{ij}(t) &= \left(\frac{(\|\mathcal{A}\|+C_F)|\mu_i|}{|\mathcal{A}_{ii}|} + \frac{(\|\mathcal{A}\|+C_F)|\mu_j|}{|\mathcal{A}_{jj}|}C_G\psi^2\right)\mathbb{E}[\|z\|] \\ &\quad + \left[\frac{(\|\mathcal{A}\|+C_F)}{|\mathcal{A}_{ii}|} + \frac{(\|\mathcal{A}\|+C_F)}{|\mathcal{A}_{jj}|}\right]\mathbb{E}[\|z\|^2].\end{aligned}$$

As before, we use the notation $\eta^*_{ij} := \max_{t\in\mathcal{I}}\eta_{ij}(t)$ and use Gronwall's inequality to obtain the result:

Lemma 5.3. *Suppose the assumptions (A0)-(A6) hold for all $t \in \mathcal{I}$, then*

$$\mathrm{Cov}^{\Delta}_{ij}(t) \le -\frac{\eta^*_{ij}}{\mathcal{A}_{ii}+\mathcal{A}_{jj}} + \left[\mathrm{Cov}^{\Delta}_{ij}(0) + \frac{\eta^*_{ij}}{\mathcal{A}_{ii}+\mathcal{A}_{jj}}\right]\mathrm{e}^{(\mathcal{A}_{ii}+\mathcal{A}_{jj})t}. \tag{40}$$

holds for all $t \in \mathcal{I}$.

Of course, the estimates (39)-(40) may blow-up as $t \to +\infty$ if the stationary distribution cannot be approximated well by an OU process. Indeed, if the noise is small, this is exactly the effect of large deviations [59, 49], which are going to occur on an asymptotic time scale $\mathcal{O}(\mathrm{e}^{c/\psi^2})$ as $\psi \to 0$. However, the estimate is rather explicit, i.e., if we know the Lipschitz constant, the spectral gap, the noise level and/or have some a-priori knowledge of the norms $\|z\|$ and/or $\|z\|^2$, then $\mathcal{A}_{ij} < 0$ is going to give decay for the exponential terms so that the only remaining term is $-\eta^*_{ij}/(\mathcal{A}_{ii}+\mathcal{A}_{jj})$. The linear approximation (28) will be a good approximation for a certain initial time-scale. The worst-case bound is the following:

Theorem 5.4. *Suppose the assumptions (A0)-(A6) hold for all $t \in \mathcal{I}$, then there exists a constant $C_1 > 0$ and a constant $\eta^*_{\mathcal{A}} := \max_{ij} -\eta^*_{ij}/(\mathcal{A}_{ii}+\mathcal{A}_{jj})$ such that*

$$\|\mathrm{Cov}(z(t)) - \mathrm{Cov}(\tilde{Z}(t))\|_2 \le \eta^*_{\mathcal{A}} + C_1[\|\mathrm{Cov}(z_0) - \mathrm{Cov}(\tilde{Z}_0)\|_2]\mathrm{e}^{-t\min_i|\mathcal{A}_{ii}|}. \tag{41}$$

Remark: Note that taking the maximum $\eta^*_{\mathcal{A}}$ implies that the constants in (41) can be chosen independently of N as the matrix $\mathcal{A}$ has spectrum in the left half of the complex plane due to (A5).

Proof. Using Lemma 5.2 and Lemma 5.3 the result easily follows. □

In summary, Theorem 5.4 just states that one has to be extremely careful to trust a local linear approximation in a numerical context for problems with large noise and/or a very strong nonlinearity, which both lead to a very quick pathwise sampling of a non-Gaussian stationary distribution. However, for small noise, sub-exponential time scales and/or a weak nonlinearity, the local approximation via linearization and covariance operators of an OU process is going to correctly expose the relevant directions of fluctuations.

6 SODEs - Noise Truncation

In many practical applications, one tends to make one further approximation which we also discuss here. The main observation is that for sufficiently fast decay of the eigenvalues $\lambda_{Q,i}$, it is highly beneficial in practical computations to consider a further approximation

$$W^N(t) \approx \sum_{i=1}^{R} \sqrt{\lambda_{Q,i}}\beta_i(t)e_i =: W^R(t)$$

for some $R \leq N$, i.e., one truncates the noise. This means we now have to compare two OU processes given by our previous linear SODE

$$\mathrm{d}\tilde{Z} = \mathcal{A}\tilde{Z}\ \mathrm{d}t + \tilde{B}\ \mathrm{d}W^N, \qquad \tilde{Z}(0) = \tilde{Z}_0. \tag{42}$$

as well as the noise-truncated SODE

$$\mathrm{d}U = \mathcal{A}U\ \mathrm{d}t + B\ \mathrm{d}W^R, \qquad U(0) = \tilde{Z}_0. \tag{43}$$

Here B denotes the reduction of the matrix $\tilde{B} \in \mathbb{R}^{N\times N}$ to the first $N \times R$ block. In particular, we have to compare $\mathrm{Cov}(\tilde{Z})$ and $\mathrm{Cov}(U)$.

Theorem 6.1. *Considering the SODEs* (42) *and* (43) *we have*

$$\sup_{t\in\mathcal{I}} \|\mathrm{Cov}(\tilde{Z}) - \mathrm{Cov}(U)\|_2 \leq C_{tr} \sum_{i=R+1}^{N} \lambda_{Q,i}$$

where $C_{tr} > 0$ is a constant.

Proof. As in Section 4, we find that

$$\sup_{t\in\mathcal{I}} \|\mathrm{Cov}(\tilde{Z}) - \mathrm{Cov}(U)\|_2 \leq C\ \sup_{t\in\mathcal{I}} \|\mathrm{Cov}(\tilde{Z} - U)\|_2$$

where $C > 0$ is some constant. Then we observe that the process $\tilde{Z} - U$ satisfies the SODE

$$\mathrm{d}(\tilde{Z} - U) = \tilde{B}\ \mathrm{d}W^N - B\ \mathrm{d}W^R = B_{N-R}\mathrm{d}W^{N-R}$$

for a matrix $B_{N-R} \in \mathbb{R}^{(N-R)\times N}$ since the drift term and the first R noise components are identical. Therefore, the result follow from the fact that for a Q-Wiener process the covariance matrix is given by tQ. □

The main conclusion from Theorem 6.1 is that we can decrease the noise truncation error by making R larger but that this yields a matrix B of higher rank if R is closer to N; we shall see that the rank of B is actually crucial in low-rank approximation calculations later on. Furthermore, recall that we always assume that the eigenvalues $\lambda_{Q,k}$ decay sufficiently fast as $k \to +\infty$, which is the case in many commonly encountered practical problems, so Theorem 6.1 also states that we do not expect the noise truncation to be often not a significant practical problem.

7 Lyapunov Equation - Algebraic Reduction

Recall that we have now considered a spatial discretization of the initial SPDE and we have localized the problem near a locally deterministically stable steady state via linearization. We calculated upper bounds on the discretization error and on the linearization error for a finite time scale, including the noise truncation error. However, although we now can work with the linear SODE problem

$$\mathrm{d}U = \mathcal{A}U \ \mathrm{d}t + B \ \mathrm{d}W^R, \qquad U(0) = U_0, \tag{44}$$

we are still surprisingly *far from a practical computable problem for many applications!* It is well-known [11, 36] that $V(t) := \mathrm{Cov}(U(t))$ satisfies the matrix ordinary differential equation (ODE) given by

$$\frac{\mathrm{d}}{\mathrm{d}t}V = \mathcal{A}V + V\mathcal{A}^\top + BB^\top =: L_{\mathcal{A}}V + \mathcal{B}. \tag{45}$$

The stationary problem is given by

$$0 = \mathcal{A}V + V\mathcal{A}^\top + \mathcal{B}. \tag{46}$$

It is well-known [1], that under the assumption (A5), there exists a unique stationary solution V_* to (46), which is stable for the time-dependent problem (45). However, we work on a finite-time interval $\mathcal{I}$ so we need a convergence rate.

Theorem 7.1. *Suppose (A5) holds so that $\mathcal{A}$ also has a spectral gap then there exists a constant $C_\tau > 0$ such that*

$$\|V(t) - V_*\|_2 \leq C_\tau(\|V(0) - V_*\|_2)\mathrm{e}^{-2t\min_i(|\mathrm{Re}(\lambda_i)|)} \tag{47}$$

or, alternatively

$$\|V(t) - V_*\|_2 \leq C_\tau(H) \ (\|V(0) - V_*\|_2)\mathrm{e}^{-2t/\|H\|_2}, \tag{48}$$

where H solves $\mathcal{A}H + H\mathcal{A}^\top + 2\mathrm{Id} = 0$ and $\{\lambda_i\}_{i=1}^N$ are the eigenvalues of $\mathcal{A}$.

Proof. For (47), one first uses that the eigenvalues of the linear operator $L_{\mathcal{A}}$ are given by $\lambda_i + \lambda_j$ where $\{\lambda_k\}_{k=1}^N$ are the eigenvalues of $\mathcal{A}$ [8]. For (48), we use that V_* is a steady state of (45) to obtain

$$\begin{aligned} V(t) &= V_* + \exp(t\mathcal{A})(V(0) - V_*)\exp(t\mathcal{A}^\top) \\ \Rightarrow \quad \|V(t) - V_*\|_2 &\leq \|\exp(\mathcal{A}t)\|_2^2\|V(0) - V_*\|_2 \end{aligned}$$

Then by [27, Theorem 3.1]

$$\|\exp(t\mathcal{A})\|_2^2 \leq C_\tau(H) \ \exp\left(\frac{-2t}{\|H\|_2}\right),$$

which leads to (48). □

Basically, Theorem 7.1 gives an estimate, which allows us to reduce the computation to the Lyapunov equation (46) to the stationary case as long we do not start far away from the locally linearized approximate solution. Unfortunately, direct solution methods, such as the Bartels-Stewart algorithm [5] are unlikely to work as the space dimension N grows drastically in practice as we decrease h. Furthermore, N increases due to the curse of dimensionality as d increases. Direct solution methods come with a complexity $\mathcal{O}(N^3)$ and with storage requirements of $\mathcal{O}(N^2)$, which limits their applicability to problems of moderate sizes. Therefore, we have to use special methods for large-scale Lyapunov equations that work with complexities and storage requirements linear in N.

8 Lyapunov Equation - Low-Rank and Computation

In this section, we want to consider and numerically compute a low-rank approximation of V_*. This involves two steps, which can be be accomplished by carefully tracing the literature: (1) understanding error estimates for the low-rank approximation and (2) finding an error estimate for the computation in a low-rank iterative algorithm for solving Lyapunov equations.

8.1 Low-rank Approximations and Singular Value Decay

Consider the singular value decomposition (SVD) of V_*:

$$\begin{aligned} &V_* = Y\Sigma X^\top, \quad Y^\top Y = \mathrm{Id} = X^\top X, \\ &Y = [y_1(V_*), \dots, y_N(V_*)], \quad X = [x_1(V_*), \dots, x_N(V_*)], \\ &\Sigma = \mathrm{diag}(\sigma_1(V_*), \dots, \sigma_N(V_*)), \quad \sigma_1(V_*) \geq \dots \geq \sigma_N(V_*) \geq 0, \end{aligned}$$

where $\sigma_i(V_*)$, $x_i(V_*)$, $y_i(V_*)$ are the singular values, left and, respectively, right singular vectors. The best approximation of V_* of rank $\mathfrak{r}$ is, by the Eckhart-Young theorem [22, Theorem 2.4.8.], obtained by

$$V_* \approx V_*^{\mathrm{lr},\mathfrak{r}} := \sum_{i=1}^{\mathfrak{r}} \sigma_i(V_*) u_i(V_*) x_i(V_*)^\top = Y_\mathfrak{r} \Sigma_\mathfrak{r} X_\mathfrak{r}, \tag{49}$$

where $Y_\mathfrak{r}$, $X_\mathfrak{r}$ contain the first $\mathfrak{r}$ columns of Y, X, and $\Sigma_\mathfrak{r}$ the $\mathfrak{r}$ largest singular values. Note that since $V_* = V_*^\top$, $X_\mathfrak{r} = Y_\mathfrak{r}$ can be chosen. The approximation error is given by

$$\|V_* - V_*^{\mathrm{lr},\mathfrak{r}}\|_2 \leq \sigma_{\mathfrak{r}+1}(V_*). \tag{50}$$

If the singular values of V_* decay rapidly towards zero, a small error can be achieved with small values of $\mathfrak{r}$. In fact, it is possible to show [48, 23, 2, 53, 7] that solutions of large-scale Lyapunov equations with low-rank inhomogeneities often show a fast singular value decay. By assumption (A0) we restrict to the symmetric case $\mathcal{A} = \mathcal{A}^\top$. The following basic estimate on the singular value decay can be found in [48]

$$\sigma_{R\mathfrak{r}+1}(V_*) \leq \sigma_1(V_*) \left(\prod_{i=0}^{\mathfrak{r}-1} \frac{\kappa(\mathcal{A})^{(2i+1)/(2\mathfrak{r})} - 1}{\kappa(\mathcal{A})^{(2i+1)/(2\mathfrak{r})} + 1} \right)^2, \quad 1 \leq R\mathfrak{r} \leq N, \tag{51}$$

where $\kappa(\mathcal{A}) = \|\mathcal{A}\|_2\|\mathcal{A}^{-1}\|_2$. A more precise bound is developed in [53] using $\mathrm{spec}(\mathcal{A}) \subset [a, b]$ and quantities related to elliptic functions and integrals, which we briefly recall in the following definition (see also, e.g., [24, 42]).

Definition 8.1. *The elliptic integral of the first kind defined on* $[0, 1]$ *with respect to the modulus* $0 < k < 1$ *is*

$$s_k(x) := \int_0^x \frac{1}{\sqrt{(1-t^2)(1-k^2t^2)}}\,\mathrm{d}t.$$

Define also the elliptic functions sn, dn *by*

$$\mathrm{sn}(s_k) = x(s_k), \quad \mathrm{dn}(s_k) = \sqrt{1-k^2\mathrm{sn}(s_k)},$$

where $\mathrm{sn}(s_k)$ *resembles the inverse function of* $s_k(x)$*. The associated complete elliptic integral (w.r.t. the modulus* k*) is the value* $K := s_k(1)$*. The complementary modulus is* $k' = \sqrt{1-k^2}$ *and the associated complementary complete elliptic integral by* $K' = s_{k'}(1)$*. The nome* q *is defined by* $q := \exp(-\pi K'/K)$*.*

Lemma 8.2. *([24, p. 430]) The nome* q *and complementary modulus* k' *also satisfy the identity*

$$\sqrt{k'} = \prod_{i=1}^{\infty}\left[\frac{1-q^{2i-1}}{1+q^{2i-1}}\right]^2 = \frac{1-2q+2q^4-2q^9+\ldots}{1+2q+2q^4+2q^9+\ldots}. \tag{52}$$

Using these quantities, the singular values of V_* can be bounded by the following result:

Theorem 8.3. *([53, Theorem 2.1.1.],[17]) Let* $\mathcal{A} = \mathcal{A}^\top$ *and* $\mathrm{spec}(\mathcal{A}) \in [a, b]$ *with* $a := \min\lambda_i < b := \max\lambda_i < 0$*. Set the complementary modulus to* $k' = b/a = 1/\kappa(\mathcal{A})$ *and set* k*, the complete elliptic integrals* K' *and* K*, as well as the nome* q *via the relations in Definition 8.1. Then it holds for the singular values of the solution of* (46)

$$\sigma_{R\mathfrak{r}+1}(V_*) \le \sigma_1(V_*)\left(\frac{1-\sqrt{k'_\mathfrak{r}}}{1+\sqrt{k'_\mathfrak{r}}}\right)^2, \quad 1 \le R\mathfrak{r} \le N, \tag{53}$$

where $k'_\mathfrak{r}$ *relates to* $q^\mathfrak{r}$ *in the same way* k' *is build from* q *via* (52)*:*

$$\sqrt{k'_\mathfrak{r}} = \frac{1-2q^\mathfrak{r}+2q^{4\mathfrak{r}}-2q^{9\mathfrak{r}}+\ldots}{1+2q^\mathfrak{r}+2q^{4\mathfrak{r}}+2q^{9\mathfrak{r}}+\ldots}.$$

We stress that, while better than (51), the bound (53) might not be very sharp in practice, where one often observes an even faster singular value decay. One reason is that Theorem 8.3 only uses $\kappa(\mathcal{A})$, i.e., the extremal eigenvalue a, b of $\mathcal{A}$. More realistic, but also more difficult to compute, bounds can be obtained by using more than these two eigenvalues of $\mathcal{A}$ [53].

Remark: *In case of non-symmetric* $\mathcal{A}$*, the above bounds are not applicable. In this case the singular value decay of* V_* *is more complicated and we refer to e.g., the discussions in [4, 7, 53].*

Judging by (53), the decay of the singular values depends mainly on $\kappa(\mathcal{A})$ and the value $R = \operatorname{coldim}(B)$. For instance for fixed a, the closer b to the origin, i.e. the smaller the spectral gap, the larger $\kappa(\mathcal{A})$ and, consequently, the closer k, k' will be to one and, respectively, zero. Hence, K' and K will tend towards $\frac{\pi}{2}$ and ∞, respectively, leading to q close to one. In this extreme situation, $q^{\mathfrak{r}}$ will also be close to one, leading, in the end, to

$$\frac{1-\sqrt{k'_{\mathfrak{r}}}}{1+\sqrt{k'_{\mathfrak{r}}}} \approx 1$$

such that no singular value decay might be observed. In particular, the main insight is that the spectral gap also plays a *numerical analysis* role at the error in step (S5). Furthermore, the column dimension R of B plays a significant role in the bound (53). Recall the B is the result from evaluating g at u^* and approximating the stochastic process $W(t)$. Recall that in (S3) we truncated the Q-Wiener process [49] to obtain a numerical approximation [45]

$$W(t) \approx \sum_{i=1}^{R} \sqrt{\lambda_{Q,i}}\ \beta_i(t)\ e_i$$

for a basis $\{e_i\}_{i=1}^{\infty}$ of $\mathcal{H}$, eigenvalues $\lambda_{Q,i}$ of Q, and independent Brownian motions $\beta_i(t)$. The scalars $\lambda_{Q,i}$ form a non-increasing sequence. Hence, the slower the $\lambda_{Q,i}$ decrease, the higher the value of R should be chosen and, consequently, the slower the decay of the singular values of V_*. This implies that space-time white noise is more difficult to treat numerically than a Q-trace-class Wiener process in our setting.

8.2 Error Bounds for Numerically Computed Low-Rank Solutions

Using the low-rank approximation (49) is not practical as it requires to first obtain V_* and then compute its singular valued decomposition. Numerical methods for large-scale Lyapunov equations [54, 10] typically directly compute low-rank factors $\mathcal{Z} \in \mathbb{R}^{N\times\mathfrak{r}}$, $\mathfrak{r} \ll N$ that form a low-rank approximate solution $V_*^{\mathrm{lr},\mathfrak{r}} = \mathcal{Z}\mathcal{Z}^\top$ which will, however, not be optimal in the sense of the SVD based approximation (49) but close if the method is properly executed. The advantage of these methods is that they are able to provide accurate low-rank solutions in a very efficient manner by utilizing tools from large-scale numerical linear algebra. By exploiting, e.g., the sparsity of $\mathcal{A}$ and the low-rank R of the inhomogeneity, state-of-the-art methods [54, 10] are able to compute low-rank solution factors at complexities and memory requirements of $\mathcal{O}(N)$.

One iterative method for solving (46) is based on the fact that for any $\alpha \in \mathbb{C}_-$, (46) is equivalent to the matrix equation

$$\begin{aligned} X &= \mathcal{C}(\alpha)X\mathcal{C}(\alpha) + \mathcal{B}(\alpha)\mathcal{B}(\alpha)^H \\ \text{with}\quad \mathcal{C}(\alpha) &:= (\mathcal{A}+\alpha\mathrm{Id})^{-1}(\mathcal{A}-\overline{\alpha}\mathrm{Id}), \quad \mathcal{B}(\alpha) := \sqrt{-2\mathrm{Re}(\alpha)}(\mathcal{A}+\alpha\mathrm{Id})^{-1}B, \end{aligned}$$

where $(\cdot)^H$ is the Hermitian conjugate. This motivates the self-evident iteration scheme

$$X_j = \mathcal{C}(\alpha_j)X_{j-1}\mathcal{C}(\alpha_j) + \mathcal{B}(\alpha_j)\mathcal{B}(\alpha_j)^H$$

for varying $\alpha_j \in \mathbb{C}_-$, which is the alternating directions implicit (ADI) iteration for Lyapunov equations [58]. In order to be applicable to large-scale equations, one uses $X_0 = 0$ and, after a series of basic algebraic manipulation [44], one arrives at the low-rank ADI (LR-ADI) iteration [44, 9, 39, 40]

$$H_j = (\mathcal{A} + \alpha_j \mathrm{Id})^{-1} \mathcal{W}_{j-1}, \quad \mathcal{W}_j = \mathcal{W}_{j-1} - 2\mathrm{Re}(\alpha_j) H_j,$$
$$\mathcal{Z}_j = [\mathcal{Z}_{j-1},\ \sqrt{-2\mathrm{Re}(\alpha_j)} H_j] \in \mathbb{C}^{N \times Rj}$$

for $\mathcal{W}_0 := B$, $j \geq 1$. It produces low-rank approximations of the solution of (46) of the form $V_* \approx V_j := \mathcal{Z}_j \mathcal{Z}_j^H$. The numbers $\alpha_j \in \mathbb{C}_-$ are referred to as shift parameters and are crucial for a fast convergence of the LR-ADI iteration. Obviously, the main numerical effort of the LR-ADI iteration comes from the solution of the linear systems of equations $(\mathcal{A} + \alpha_j \mathrm{Id}) H_j = W_{j-1}$ for H_j which can for sparse $\mathcal{A}$ be done efficiently by sparse-direct [15] or iterative solvers [52]. However, the number of right hand sides in each linear system is given by the key quantity R. We see here that the higher R, the larger the numerical effort of the LR-ADI iteration becomes. The error of the constructed low-rank approximation V_j is given by

$$V_j - V_* = \mathcal{J}_j V_* \mathcal{J}_j^H, \quad \mathcal{J}_j := \prod_{i=1}^{j} \mathcal{C}_i, \quad \mathcal{C}_i := \mathcal{C}(\alpha_i) \tag{54}$$

such that

$$\|V_j - V_*\|_2 \leq \kappa(S)^2 \|V_*\|_2 R_j^2, \quad \Theta_j := \prod_{i=1}^{j} \rho_i,$$
$$\rho_i = \rho(\mathcal{C}_i) = \max_{z \in \mathrm{spec}(\mathcal{A})} \left| \frac{z - \overline{\alpha_i}}{z + \alpha_i} \right|,$$

and S is the matrix containing the eigenvectors of $\mathcal{A}$. Since $\mathrm{spec}(\mathcal{A}) \in \mathbb{C}_-$ and $\alpha_i \in \mathbb{C}_-$, it holds $\rho_i < 1$, $\forall i \geq 1$ and, thus, $\Theta_j = \rho_j \Theta_{j-1} < \Theta_{j-1}$, indicating that the sequence of the spectral radii Θ_j is monotonically decreasing and, in the limit, will approach the value zero. The shifts α_i should therefore be chosen such that Θ_j is as small as possible leading to the ADI parameter problem

$$\{\alpha_1^*, \ldots, \alpha_j^*\} = \arg\min \Theta_j = \arg \min_{\alpha_i \in \mathbb{C}_-} \max_{z \in \mathrm{spec}(\mathcal{A})} \left| \prod_{i=1}^{j} \frac{z - \overline{\alpha_i}}{z + \alpha_i} \right|. \tag{55}$$

This problem is in general a formidable task which has been addressed in numerous works, e.g., in [17, 58, 53, 9, 39]. Again, the situation simplifies for the important case $\mathcal{A} = \mathcal{A}^\top$, where real shifts are usually sufficient (and $\mathcal{Z}_j$ will also be real). Note that in this case $\kappa(S) = 1$. In summary, we have the following relevant result for our purposes:

Theorem 8.4. *([57, 53, 58]) With the same assumptions and settings for k, k', K, K', q as in Theorem 8.3, construct real shift parameters $\alpha_1, \ldots, \alpha_j \in \mathbb{R}_-$ by*

$$\alpha_i = a\mathrm{dn}((2i-1)K/2j),\ 1 \leq i \leq j \tag{56}$$

with the elliptic function dn *from Definition 8.1. Using these shifts, the smallest value of the spectral radius* $\Theta_j = \Theta_j(\alpha_1, \ldots, \alpha_j)$ *in* (55) *is* $\frac{1-\sqrt{k'_j}}{1+\sqrt{k'_j}}$*, where the modulus* k'_j *is associated to the nome* q^j *via* (52)*. Hence, carrying out* j *steps of the LR-ADI iteration using* (56) *yields*

$$\|V_j - V_*\|_2 \leq \|V_*\|_2 \left(\frac{1-\sqrt{k'_j}}{1+\sqrt{k'_j}}\right)^2.$$

We can, thus, expect to approximate the solution V_* by the LR-ADI low-rank approximation V_j at a speed similar to the predicted singular value decay by Theorem 8.3. Other low-rank algorithms for solving (46), e.g., rational Krylov subspace methods allow similar error bounds [14, 6].

9 Summary and Main Result

Although we stated and proved all our main error estimates, it is helpful to summarize the results to provide a CERES. Recall from the introduction that our setup considered four steps. To combine the four steps, we have to link operators on $\dot{\mathcal{H}}^1$ with the finite-dimensional approximation spaces $\mathcal{S}_h$. If $L_h : \mathcal{S}_h \to \mathcal{S}_h$ is a finite-dimensional linear operator, then we can always view it as an infinite-dimensional linear operator L on $\dot{\mathcal{H}}^1$ by declaring basis vectors not in $P_h\dot{\mathcal{H}}^1$ to lie in nullspace(L).

Theorem 9.1. *Suppose (A0)-(A6) hold. Let* $\mathrm{Cov}(u)$ *denote the covariance operator of the SPDE* (5)*. Then a low-rank solution* V_j *of the locally linearized discretized problem on* $\mathcal{S}_h$ *near the steady state* u^* *computed after* j *ADI steps satisfies the CERES*

$$\sup_{t\in[0,T]} \|\mathrm{Cov}(u) - V_j\|_{\mathcal{L}_2} \lesssim \sup_{t\in\mathcal{I}} \left[\mathrm{err}_{(\mathrm{S1})} + \mathrm{err}_{(\mathrm{S2})} + \mathrm{err}_{(\mathrm{S3})} + \mathrm{err}_{(\mathrm{S4})} + \mathrm{err}_{(\mathrm{S5})}\right] \tag{57}$$

and the individual error terms are given by

$$\begin{aligned}
\mathrm{err}_{(\mathrm{S1})} &= C_{\mathrm{d}} h^{1+r}, && (58)\\
\mathrm{err}_{(\mathrm{S2})} &= \eta^*_{\mathcal{A}} + C_{\mathrm{l}}[\|\mathrm{Cov}(u_h(0)) - \mathrm{Cov}(\tilde{U}_h(0))\|_2]\mathrm{e}^{-t\min_i|\mathcal{A}_{ii}|}, && (59)\\
\mathrm{err}_{(\mathrm{S3})} &= C_{tr}\sum_{i=R+1}^{N} \lambda_{Q,i} && (60)\\
\mathrm{err}_{(\mathrm{S4})} &= C_\tau(\|\mathrm{Cov}(U_h(0)) - V_*\|_2)\mathrm{e}^{t\max(\mathrm{spec}(\mathcal{A}))} && (61)\\
\mathrm{err}_{(\mathrm{S5})} &= \|V_*\|_2 \left(1-\sqrt{k'_j}\right)^2 / \left(1+\sqrt{k'_j}\right)^2, && (62)
\end{aligned}$$

where $C_{\mathrm{d}}, C_{\mathrm{l}}, C_\tau, k'_j, C_{tr}, \eta^*_{\mathcal{A}} > 0$ *are constants depending upon the data (i.e. on* A, f, g, Q*), the terms* $u_h(0), \tilde{U}_h(0)$ *are initial conditions for the discretized full and linearized problems,* $\lambda_{Q,i}$ *are eigenvalues of* Q*,* $R \in \mathbb{N}$ *is the noise truncation level,* $V_* = \lim_{t\to+\infty} U_h(t)$ *is the finite asymptotic limit for the stationary problem, and* $\mathcal{A}$ *is the leading-order discretized linear operator part near* u^**.*

Proof. Just applying a triangle inequality to the left-hand side of (57), the proof follows from a direct application of Theorems 4.1, 5.4, 6.1, 7.1, and 8.4. □

Theorem 9.1 illustrates very well that it would be *short-sighted* to just look at one source of error. For example, even if the spatial discretization h is extremely small, the actual error could be very large if the spectral gap is small, i.e., the deterministic steady state is only weakly locally stable. We may also summarize the behaviour of the different constants and terms in the CERES into more practical observations, which effects lead to *smaller error*:

- A large gap in the spectrum of the local linearization of the deterministic part exists.
- A small spatial discretization level h is chosen.
- One starts with initial data close to the local approximating OU stationary distribution.
- The nonlinear part of f does not have a strong effect on sub-exponential time scales.
- The noise is small enough to stay for a long time in the sub-exponential regime.
- The Q-trace-class operator has fast decaying eigenvalues.
- The iteration number j in the low-rank Lyapunov solver is chosen large.

For a detailed discussion of these effects, we refer to the individual proofs of the different parts of the CERES in previous sections. However, it is very interesting to note that certain effects, which decrease the error occur in *multiple steps*. Obviously this is true for the spectral gap of $\mathcal{A} = A_h + P_h \mathrm{D}_u f(u^*)$ in (S2) and (S4) but also for the type of noise, which crucially influences the, usually growing, function $\eta^*(t)$ in (S2) as well as the convergence rate of a low-rank approximation in (S5). In addition, note that we have observed that several errors are *linked* and cannot be treated independently! For example, a small column dimension R guarantees a good low rank approximation in (S5) but selecting R small gives a larger error for the noise truncation in (S3).

10 Outlook

We stress again that our results presented here should be viewed as a first key step to introduce a general framework of CERES for high-dimensional stochastic problems, where many different sources of error naturally occur. In this regard, a multitude of problems can, and should, now be tackled from a similar viewpoint. For example, uncertainty quantification of random partial differential equations (RPDE) [20, 60, 46, 47] contains an entire chain of error sources such as truncation error for polynomial chaos, dynamical error if the RPDE is just an approximation, error from the large-scale numerical linear algebra, and/or error due to reduced bases, just to name a few. Hence, to compute a CERES in a single norm, such as the spectral norm we used here, for all the steps would be very worthwhile. Similar issues also appear for problems involving large deviations and transition paths in high-dimensional energy landscapes [12, 16, 43, 25], where developing a CERES would definitely be very helpful.

On the very concrete level of the SPDE (5) studied here, several interesting directions could be pursued. Firstly, we do not claim that all our estimates are sharp and/or the assumptions are the theoretically weakest possible. Already the CERES presented in this work is interesting and difficult to chain together properly from its different components. Nevertheless, improvements might be possible, e.g., if f is a strongly dissipative non-Lipschitz nonlinearity, we expect that the results still hold from a dynamical perspective but essentially the steps (S1)-(S2) would then require a major extension or even a completely new approach.

Furthermore, it could be desirable to specify the constant C_{d} precisely as the techniques in [34, 35] are more explicit than the final statements on optimal regularity. Unfortunately this would entail re-writing the entire optimal regularity proof for the spatial discretization so we refrain from attempting to carry this program in this work. Similar remarks also apply to other constants, which are expected to be of moderate/non-asymptotic relevance only in many practical applications anyhow. Regarding applications, it would also be useful to test the method on a broader range of SPDEs beyond the current available numerical experiments and examples [36, 37, 3].

Acknowledgments: CK would like to thank the VolkswagenStiftung for support via a Lichtenberg Professorship. Furthermore, CK would like to thank Daniele Castellano for interesting discussions regarding moment equations. The work on this article was done while PK was affiliated with the Max Planck Institute for Dynamics of Complex Technical Systems in Magdeburg, Germany. We also thank two anonymous referees for their helpful comments.

References

[1] A.C. Antoulas. *Approximation of Large-Scale Dynamical Systems*. Springer, 2005.

[2] A.C. Antoulas, D.C. Sorensen, and Y. Zhou. On the decay rate of Hankel singular values and related issues. *Syst. Control Lett.*, 46(5):323–342, 2002.

[3] S. Baars, J.P. Viebahn, T.E. Mulder, C. Kuehn, F.W. Wubs and H.A. Dijkstra. Continuation of probability density functions using a generalized Lyapunov approach. *J. Comput. Phys.*, 336(1):627–643, 2017.

[4] J. Baker, M. Embree, and J. Sabino. Fast singular value decay for Lyapunov solutions with nonnormal coefficients. *SIAM J. Matrix Anal. Appl.*, 36(2):656–668, 2015.

[5] R.H. Bartels and G.W. Stewart. A solution of the equation $AX + XB = C$. *Commun. ACM*, 15:820–826, 1972.

[6] B. Beckermann. An error analysis for rational Galerkin projection applied to the Sylvester equation. *SIAM J. Num. Anal.* 49(6):2430–2450, 2012.

[7] B. Beckermann and A. Townsend. On the Singular Values of Matrices with Displacement Structure. *SIAM J. Matrix Anal. Appl.*, 38(4):1227–1248, 2017.

[8] R. Bellman. *Introduction to Matrix Analysis*. McGraw-Hill, 1960.

[9] P. Benner, P. Kürschner, and J. Saak. Self-generating and efficient shift parameters in ADI methods for large Lyapunov and Sylvester equations. *Electron. Trans. Numer. Anal.*, 43:142–162, 2014.

[10] P. Benner and J. Saak. Numerical solution of large and sparse continuous time algebraic matrix Riccati and Lyapunov equations: a state of the art survey. *GAMM Mitteilungen*, 36(1):32–52, 2013.

[11] N. Berglund and B. Gentz. *Noise-Induced Phenomena in Slow-Fast Dynamical Systems*. Springer, 2006.

[12] P.G. Bolhuis, D. Chandler, C. Dellago, and P.L. Geissler. Transition path sampling: throwing ropes over rough mountain passes. *Ann. Rev. Phys. Chem.*, 53:291–318, 2002.

[13] R.C. Dalang and L. Quer-Sardanyons. Stochastic integrals for SPDE's: A comparison. *Expositiones Math.*, 29(1):67–109, 2011.

[14] V. Druskin, L.A. Knizhnerman, and V. Simoncini. Analysis of the rational Krylov subspace and ADI methods for solving the Lyapunov equation. *SIAM J. Numer. Anal.*, 49:1875–1898, 2011.

[15] I.S. Duff, A.M. Erisman, and J.K. Reid. *Direct Methods for Sparse Matrices*. Clarendon Press, Oxford, UK, 1989.

[16] W. E., W. Ren, and E. Vanden-Eijnden. String method for the study of rare events. *Phys. Rev. B*, 66(5):052301, 2002.

[17] N.S. Ellner and E.L. Wachspress. Alternating direction implicit iteration for systems with complex spectra. *SIAM J. Numer. Anal.*, 28(3):859–870, 1991.

[18] A. Ern and J.-L. Guermond. *Theory and Practice of Finite Elements*. Springer, 2004.

[19] C. Gardiner. *Stochastic Methods*. Springer, Berlin Heidelberg, Germany, 4th edition, 2009.

[20] R.G. Ghanem and P.D. Spanos. *Stochastic Finite Elements: A Spectral Approach*. Courier Corporation, 2003.

[21] B. Goldys and J.M.A.M. Van Neerven. Transition semigroups of Banach space-valued Ornstein-Uhlenbeck processes. *Acta Applicandae Mathematica*, 76:283–330, 2003.

[22] G.H. Golub and C. van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, fourth edition, 2013.

[23] L. Grasedyck. Existence of a low rank or H-matrix approximant to the solution of a Sylvester equation. *Numer. Lin. Alg. Appl.*, 11:371–389, 2004.

[24] H. Hancock. *Lectures on the theory of elliptic functions*. Dover Publications, 1958.

[25] G. Henkelman, B.P. Uberuaga, and H. Jónsson. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *J. Chem. Phys.*, 113:9901–9904, 2000.

[26] D. Henry. *Geometric Theory of Semilinear Parabolic Equations*. Springer, Berlin Heidelberg, Germany, 1981.

[27] G.-D. Hu and M. Liu. The weighted logarithmic matrix norm and bounds of the matrix exponential. *Linear Algebra Appl.*, 390(1):145–154, 2004.

[28] A. Jentzen and P.E. Kloeden. The numerical approximation of stochastic partial differential equations. *Milan J. Math.*, 77:205–244, 2009.

[29] A. Jentzen and M. Röckner. A Milstein scheme for SPDEs. *arXiv:1001.2751v4*, pages 1–37, 2012.

[30] A. Jentzen and M. Röckner. Regularity analysis for stochastic partial differential equations with nonlinear multiplicative trace class noise. *J. Differential Equat.*, 252:114–136, 2012.

[31] O. Kallenberg. *Foundations of Modern Probability*. Springer, New York, NY, 2nd edition, 2002.

[32] H. Kielhoefer. *Bifurcation Theory: An Introduction with Applications to PDEs*. Springer, 2004.

[33] B. Krauskopf, H.M. Osinga, and J. Galán-Vique, editors. *Numerical Continuation Methods for Dynamical Systems: Path following and boundary value problems*. Springer, 2007.

[34] R. Kruse. Optimal error estimates for Galerkin finite element methods for stochastic partial differential equations with multiplicative noise. *IMA. J. Numer. Anal.*, 34:217–251, 2014.

[35] R. Kruse and S. Larsson. Optimal regularity for semilinear stochastic partial differential equations with nonlinear multiplicative trace-class noise. *Electron. J. Probab.*, 17:1–19, 2012.

[36] C. Kuehn. Deterministic continuation of stochastic metastable equilibria via Lyapunov equations and ellipsoids. *SIAM J. Sci. Comp.*, 34(3):A1635–A1658, 2012.

[37] C. Kuehn. Numerical continuation and SPDE stability for the 2d cubic-quintic Allen-Cahn equation. *SIAM/ASA J. Uncertain. Quantif.*, 3(1):762–789, 2015.

[38] C. Kuehn. Moment closure - a brief review. In E. Schöll, S. Klapp, and P. Hövel, editors, *Control of Self-Organizing Nonlinear Systems*, pages 253–271. Springer, 2016.

[39] P. Kürschner. *Efficient Low-Rank Solution of Large-Scale Matrix Equations*. Dissertation, OVGU Magdeburg, Shaker Verlag, 2016.

[40] P. Kürschner. *Approximate residual minimizing shift parameters for the low-rank ADI iteration. Electron. Trans. Numer. Anal.*, 51:240–261, 2019.

[41] A. Lang, S. Larsson, and C. Schwab. Covariance structure of parabolic stochastic partial differential equations. *Stoch. PDE: Anal. Comp.*, 1(2):351–364, 2013.

[42] D.F. Lawden. *Elliptic Functions and Applications*. Springer New York, 1989.

[43] T. Lelièvre, G. Stoltz, and M. Rousset. *Free Energy Computations: A Mathematical Perspective*. World Scientific, 2010.

[44] J.-R. Li and J. White. Low rank solution of Lyapunov equations. *SIAM J. Matrix Anal. Appl.*, 24(1):260–280, 2002.

[45] G.J. Lord, C.E. Powell, and T. Shardlow. *An Introduction to Computational Stochastic PDEs*. CUP, 2014.

[46] O. Le Maître and O.M. Knio. *Spectral Methods for Uncertainty Quantification: with applications to computational fluid dynamics*. Springer, 2010.

[47] F. Nobile, R. Tempone, and C.G. Webster. A sparse grid stochastic collocation method for partial differential equations with random input data. *SIAM J. Numer. Anal.*, 46(5):2309–2345, 2008.

[48] T. Penzl. Eigenvalue decay bounds for solutions of Lyapunov equations: the symmetric case. *Syst. Control Lett.*, 40:139–144, 2000.

[49] G. Da Prato and J. Zabczyk. *Stochastic Equations in Infinite Dimensions*. Cambridge University Press, 1992.

[50] C. Prévot and M. Röckner. *A Concise Course on Stochastic Partial Differential Equations*, volume 1905 of *Lecture Notes in Mathematics*. Springer, 2008.

[51] J.C. Robinson. *Infinite-Dimensional Dynamical Systems*. CUP, 2001.

[52] Y. Saad. *Iterative Methods for Sparse Linear Systems*. SIAM, Philadelphia, PA, 2003.

[53] John Sabino. *Solution of Large-Scale Lyapunov Equations via the Block Modified Smith Method*. PhD thesis, Rice University, Houston, Texas, June 2007.

[54] V. Simoncini. Computational methods for linear matrix equations. *SIAM Rev.*, 58(3):377–441, 2016.

[55] L. Socha. *Linearization Methods for Stochastic Dynamic Systems*. Springer, 2008.

[56] V. Thomée. *Galerkin Finite Element Methods for Parabolic Problems*. Springer, 2006.

[57] E.L. Wachspress. Extended application of alternating direction implicit iteration model problem theory. *J. Soc. Ind. Appl. Math.*, 11(4):994–1016, 1963.

[58] E.L. Wachspress. *The ADI model problem*. Springer New York, 2013.

[59] A.D. Wentzell and M.I. Freidlin. On small random perturbations of dynamical systems. *Russ. Marth. Surv.*, 25:1–55, 1970.

[60] D. Xiu and J.S. Hesthaven. High-order collocation methods for differential equations with random inputs. *SIAM J. Sci. Comput.*, 27(3):1118–1139, 2005.