



HAL
open science

Numerical methods for the 2nd moment of stochastic ODEs

Roman Andreev, Kristin Kirchner

► **To cite this version:**

Roman Andreev, Kristin Kirchner. Numerical methods for the 2nd moment of stochastic ODEs. 2016. hal-01394195

HAL Id: hal-01394195

<https://hal.science/hal-01394195>

Preprint submitted on 8 Nov 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Numerical methods for the 2nd moment of stochastic ODEs

R. ANDREEV AND K. KIRCHNER

ABSTRACT. Numerical methods for stochastic ordinary differential equations typically estimate moments of the solution from sampled paths. Instead, in this paper we directly target the deterministic equation satisfied by the first and second moments. For the canonical examples with additive noise (Ornstein–Uhlenbeck process) or multiplicative noise (geometric Brownian motion) we derive these deterministic equations in variational form and discuss their well-posedness in detail. Notably, the second moment equation in the multiplicative case is naturally posed on projective–injective tensor products as trial–test spaces. We propose Petrov–Galerkin discretizations based on tensor product piecewise polynomials and analyze their stability and convergence in the natural norms.

1. INTRODUCTION

Ordinary and partial differential equations are pervasive in financial, biological, engineering and social sciences, to name a few. Often, randomness is introduced to model uncertainties in the coefficients, in the geometry of the physical domain, in the boundary or initial conditions, or in the sources (right-hand sides). In this paper we aim at the latter scenario, specifically ordinary or partial differential evolution equations driven by Brownian noise. The random solution is then a continuous-time stochastic process with values in a certain state space. When the state space is of finite dimension (≤ 3 , say), it may be possible to approximate numerically the temporal evolution of the probability density function of the stochastic process, but in most applications, only the first few statistical moments of the random solution may be of interest or even feasible to compute.

The computation of moments of the random solution is typically based on sampling methods such as Monte Carlo. In general, Monte Carlo methods are, however, computationally expensive due to the convergence order $1/2$ of the Monte Carlo estimation and the high cost for computing sample paths of solutions to stochastic differential equations. Recent developments aiming at reducing the computational cost include multilevel Monte Carlo methods, e.g., [6, 7] and quasi-Monte Carlo integration [10] or combinations thereof [8, 11].

An alternative to sampling for the covariance of a parabolic stochastic PDE driven by additive Brownian noise was proposed in [12]. It is based on the insight that the second moment satisfies a deterministic equation that can be formulated as a well-posed linear space-time variational formulation on Hilbert tensor products of Bochner spaces. The main promise of space-time variational formulations is in potential savings in computing time and memory through space-time compressive schemes, e.g., using adaptive wavelet methods or low rank tensor approximations. Multiplicative noise requires a more careful analysis because firstly, an extra term in the space-time variational formulation constrains it to projective–injective tensor products of those Bochner spaces for the trial–test spaces [9]. Secondly, the well-posedness is self-evident only as long as the volatility of the multiplicative noise is sufficiently small. Consequently, while it is relatively straightforward to derive numerical methods in the case of additive noise (by tensorizing existing space-time discretizations of deterministic parabolic evolution equations), new techniques are necessary in the case of multiplicative noise. To fully explain and address those issues, in this paper we focus entirely on canonical examples of stochastic ODEs driven by additive or

Date: November 8, 2016.

2010 Mathematics Subject Classification. 65C30, 60H10, 65L60.

Key words and phrases. Stochastic ordinary differential equations, Additive noise, Multiplicative noise, Variational problems, Hilbert tensor product, Projective and injective tensor product, Petrov–Galerkin discretization.

multiplicative Brownian noise. However, to facilitate the transition to parabolic stochastic PDEs, our estimates are explicit and sharp in the relevant parameters.

The paper is structured as follows. In Section 2 we introduce the model stochastic ODEs and the necessary definitions, derive the deterministic equations for the first and second moments and discuss their well-posedness. In Section 3 we present conforming Petrov–Galerkin discretizations of those equations and discuss their stability, concluding with a numerical example. Section 4 summarizes the paper.

A comment on notation. If X is a Banach space then $S(X)$ denotes its unit sphere. We write $s \wedge t := \min\{s, t\}$. The symbol δ (δ_s) denotes the Dirac measure (at s). The closure of an interval J is \bar{J} . The symbol \otimes variously denotes the tensor product of two functions or the algebraic tensor product of function spaces, depending on the context.

2. DERIVATION OF THE DETERMINISTIC MOMENT EQUATIONS

2.1. Model stochastic ODEs. Let $T > 0$, set $J := (0, T)$. The focus of this paper are the model real-valued stochastic ODEs with additive noise

$$(1) \quad dX(t) + \lambda X(t) dt = \mu dW(t), \quad t \in \bar{J}, \quad \text{with } X(0) = X_0,$$

or with multiplicative noise

$$(2) \quad dX(t) + \lambda X(t) dt = \rho X(t) dW(t), \quad t \in \bar{J}, \quad \text{with } X(0) = X_0.$$

Here,

- $\lambda > 0$ is a fixed positive number that models the action of an elliptic operator,
- W is a real-valued Brownian motion defined on a complete probability space $(\Omega, \mathcal{A}, \mathbb{P})$,
- $\mu, \rho > 0$ are parameters specifying the volatility of the noise,
- the initial value $X_0 \in L_2(\Omega)$ is a random variable independent of the Brownian motion with known first and second moments (but not necessarily with a known distribution).

We call \mathcal{F}_t the σ -algebra generated by the initial value X_0 and the Brownian motion $\{W(s) : 0 \leq s \leq t\}$, and \mathcal{F} the resulting filtration. The expectation operator is denoted by \mathbb{E} . We refer to [13] for basic notions of stochastic integration and Itô calculus.

A real-valued stochastic process X is said to be a (strong continuous) solution of the stochastic differential equation “ $dX + \lambda X = \sigma(X) dW$ on \bar{J} with $X(0) = X_0$ ” if **a)** X is progressively measurable with respect to \mathcal{F} , **b)** the expectation of $\|\lambda X\|_{L_1(J)} + \|\sigma(X)\|_{L_2(J)}^2$ is finite, **c)** the integral equation

$$X(t) = X_0 - \lambda \int_0^t X(s) ds + \int_0^t \sigma(X(s)) dW(s) \quad \forall t \in \bar{J}$$

holds (\mathbb{P} -a.s.), and **d)** $t \mapsto X(t)$ is continuous (\mathbb{P} -a.s.). By standard theory [13, Theorem 5.2.1], a Lipschitz condition on σ implies existence and uniqueness of such a solution. Moreover, it has finite second moments. For future reference, we state here the integral equations for (1)–(2):

$$(3) \quad X(t) = X_0 - \lambda \int_0^t X(s) ds + \mu \int_0^t dW(s) \quad \forall t \in \bar{J} \quad (\mathbb{P}\text{-a.s.}),$$

$$(4) \quad X(t) = X_0 - \lambda \int_0^t X(s) ds + \rho \int_0^t X(s) dW(s) \quad \forall t \in \bar{J} \quad (\mathbb{P}\text{-a.s.}).$$

The solution processes and their first/second moments are known explicitly:

| | Additive (3) (Ornstein–Uhlenbeck process) | Multiplicative (4) (Geometric Brownian motion) |
|-------------------------------------|--------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------|
| (5a) $X(t)$ | $e^{-\lambda t}X_0 + \mu \int_0^t e^{-\lambda(t-s)} dW(s)$ | $X_0 e^{-(\lambda+\rho^2/2)t+\rho W(t)}$ |
| (5b) $\mathbb{E}[X(t)]$ | $e^{-\lambda t}\mathbb{E}[X_0]$ | $e^{-\lambda t}\mathbb{E}[X_0]$ |
| (5c) $\mathbb{E}[X(s)X(t)]$ | $e^{-\lambda(t+s)}\mathbb{E}[X_0^2] + \frac{\mu^2}{2\lambda}(e^{-\lambda t-s } - e^{-\lambda(t+s)})$ | $e^{-\lambda(t+s)+\rho^2(s\wedge t)}\mathbb{E}[X_0^2]$ |
| (5d) $\mathbb{E}[\ X\ _{L_2(J)}^2]$ | $\frac{1-e^{-2\lambda T}}{2\lambda}\mathbb{E}[X_0^2] + \frac{\mu^2}{4\lambda^2}(e^{-2\lambda T} + 2\lambda T - 1)$ | $\frac{e^{(\rho^2-2\lambda)T}-1}{\rho^2-2\lambda}\mathbb{E}[X_0^2]$ |

The square integrability (5d) in conjunction with Fubini's theorem will be used to interchange the order of integration over J and Ω without further mention. Square integrability also implies the useful martingale property (see [13, Corollary 3.2.6] and [13, Definition 3.1.4])

$$(6) \quad \mathbb{E} \left[\int_0^t X(r) dW(r) \middle| \mathcal{F}_s \right] = \int_0^s X(r) dW(r), \quad 0 \leq s \leq t.$$

Choosing $s = 0$ shows that the stochastic integral $\int_0^t X(r) dW(r)$ has expectation zero. If Y_1 and Y_2 are two square integrable processes adapted to \mathcal{F} , the Itô isometry [13, Corollary 3.1.7], along with (6) and the polarization identity yield the equality

$$(7) \quad \mathbb{E} \left[\int_0^s Y_1(r) dW(r) \int_0^t Y_2(r) dW(r) \right] = \int_0^{s\wedge t} \mathbb{E}[Y_1(r)Y_2(r)] dr.$$

These are the main tools in the derivation of (5). We will write $X \otimes X$ for the real-valued stochastic process $(s, t) \mapsto X(s)X(t)$ on $(\Omega, \mathcal{A}, \mathbb{P})$ indexed by the parameter space $J \times J$.

A function $w \in L_2(J \times J)$ is called symmetric if $w(s, t) = w(t, s)$ for (a.e.) $s, t \in J$. It is said to be positive semi-definite if

$$(8) \quad \int_J \int_J w(s, t) \varphi(s) \varphi(t) ds dt \geq 0 \quad \forall \varphi \in L_2(J).$$

Our first aim will be to derive deterministic equations for the first and the second moments

$$m(t) := \mathbb{E}[X(t)] \quad \text{and} \quad M(s, t) := \mathbb{E}[X(s)X(t)], \quad s, t \in J,$$

as well as for the covariance function

$$(9) \quad \text{Cov}(X) := \mathbb{E}[(X - m) \otimes (X - m)] = M - (m \otimes m)$$

of the stochastic process X . The second moment and the covariance are symmetric positive semi-definite.

2.2. Deterministic first moment equations. We first introduce the spaces

$$E := L_2(J) \quad \text{and} \quad F := H_{0, \{T\}}^1(J),$$

where the latter denotes the closed subspace of the Sobolev space $H^1(J)$ of functions vanishing at $t = T$. Thanks to the embedding $F \hookrightarrow C^0(\bar{J})$, elements of F will be identified by their continuous representant. These spaces are equipped with the λ -dependent norms

$$(10) \quad \|w\|_E^2 := \lambda \|w\|_{L_2(J)}^2 \quad \text{and} \quad \|v\|_F^2 := \lambda^{-1} \|v'\|_{L_2(J)}^2 + \lambda \|v\|_{L_2(J)}^2 + |v(0)|^2,$$

and the obvious corresponding inner products $(\cdot, \cdot)_E$ and $(\cdot, \cdot)_F$. The norm on F is motivated by the fact that

$$(11) \quad \|v\|_F^2 = \lambda^{-1} \| -v' + \lambda v \|_{L_2(J)}^2 \quad \forall v \in F.$$

Lemma 2.1. *Let $v \in F$. Then*

$$(12) \quad |v(t)| \leq \frac{1}{\sqrt{2}} \|v\|_F \quad \forall t \in \bar{J}.$$

Proof. Suppose that the supremum of $|v(t)|$ is attained at some $0 \leq t \leq T$. Integrating $(v^2)' = 2vv'$ over $(0, t)$, applying the Cauchy–Schwarz and the Young inequalities leads to the estimate $|v(t)|^2 \leq \lambda^{-1}\|v'\|^2 + \lambda\|v\|^2 + |v(0)|^2$ in terms of the $L_2(0, t)$ norms. In a similar way, observing that $v(T) = 0$, we obtain $|v(t)|^2 \leq \lambda^{-1}\|v'\|^2 + \lambda\|v\|^2$ in terms of the $L_2(t, T)$ norms. Adding the two inequalities gives (12). \square

The inequality (12) is sharp in general as the function $\psi(t) := \sinh(\lambda(T-t))/\sinh(\lambda T)$ attests:

$$(13) \quad 1 = \psi(0) = \sup_{t \in J} |\psi(t)| \quad \text{and} \quad \|\psi\|_F = \sqrt{\coth(\lambda T) + 1} \rightarrow \sqrt{2} \quad \text{as} \quad \lambda T \rightarrow \infty.$$

The deterministic moment equations will be expressed in terms of the continuous bilinear form

$$(14) \quad b: E \times F \rightarrow \mathbb{R}, \quad b(w, v) := \int_J w(t)(-v'(t) + \lambda v(t)) dt.$$

We employ the same notation for the induced bounded linear operator

$$b: E \rightarrow F', \quad \langle bw, v \rangle := b(w, v),$$

and use whichever is more convenient, as should be evident from the context. The operator b arises in the weak formulation of the ordinary differential equation $u' + \lambda u = f$. With the definition of the norms (10), it is an isometric isomorphism,

$$\|bw\|_{F'} = \|w\|_E \quad \forall w \in E.$$

Indeed, $\|bw\|_{F'} \leq \|w\|_E$ is obvious from (10)–(11). To verify $\|bw\|_{F'} \geq \|w\|_E$, let $w \in E$ be arbitrary. Taking v as the solution to the ODE $-v' + \lambda v = \lambda w$ with $v(T) = 0$, it follows using (10)–(11) that $\langle bw, v \rangle = \|w\|_E^2 = \|v\|_F^2$. Therefore, $\langle bw, v \rangle = \|w\|_E \|v\|_F$, and in particular $\|bw\|_{F'} \geq \|w\|_E$. This shows the isometry property. By a similar argument, $\sup_w \langle bw, v \rangle \neq 0$ for all nonzero $v \in F$. By [3, Theorem 2.1], b is an isomorphism.

If a functional $\ell \in F'$ can be expressed as $\ell(v) = \int_J g v$ for some $g \in L_1(J)$, then $u = b^{-1}\ell$ enjoys the representation

$$(15) \quad u(t) = (b^{-1}\ell)(t) = \int_0^t e^{-\lambda(t-s)} g(s) ds.$$

Despite this integral representation, b^{-1} is not a compact operator (it is an isomorphism).

Applying the expectation operator to (3)–(4) shows that the first moment m of the solution satisfies the integral equation

$$m(t) = \mathbb{E}[X_0] - \lambda \int_0^t m(s) ds.$$

Testing this equation with the derivative of an arbitrary $v \in F$ and integrating by parts in time shows that the first moment of (3)–(4) solves the deterministic variational problem

$$(16) \quad \boxed{\text{Find } m \in E \quad \text{s.t.} \quad b(m, v) = \mathbb{E}[X_0]v(0) \quad \forall v \in F.}$$

2.3. Second moment equations: additive noise. The Hilbert tensor product spaces

$$(17) \quad E_2 := E \otimes_2 E \quad \text{and} \quad F_2 := F \otimes_2 F$$

are obtained as the closure of the algebraic tensor product $E \otimes E$ and $F \otimes F$ under the norm $\|\cdot\|_2$ induced by the tensorized inner product,

$$(u_1 \otimes u_2, w_1 \otimes w_2)_2 := (u_1, w_1)_E (u_2, w_2)_E, \quad u_i, w_i \in E,$$

and similarly for F . We write $\|\cdot\|_2$ also for the norm of F_2 and $\|\cdot\|_{-2}$ for the norm of the dual space F_2' of F_2 . We recall the canonical isometry [15, Theorem II.10]

$$(18) \quad E_2 = L_2(J) \otimes_2 L_2(J) \cong L_2(J \times J).$$

By virtue of square integrability (5d), the second moment M is an element of E_2 . We define the bilinear form

$$B: E_2 \times F_2 \rightarrow \mathbb{R}, \quad B := b \otimes b,$$

or explicitly as

$$(19) \quad B(w, v) := \int_J \int_J w(s, t) (-\partial_s + \lambda) (-\partial_t + \lambda) v(s, t) ds dt.$$

More precisely, B is the unique continuous extension of $b \otimes b$ by bilinearity from the algebraic tensor products to $E_2 \times F_2$. Boundedness and injectivity of the operator $B: E_2 \rightarrow F'_2$ induced by the bilinear form B follow readily from the corresponding properties of b , so that the operator B is an isometry and its inverse is the due continuous extension of $b^{-1} \otimes b^{-1}$. A representation of the inverse analogous to (15) also holds. For example, the integral kernel of the functional $\ell(v) := v(0)$ is $\delta_0 \otimes \delta_0$, which gives $(B^{-1}\ell)(t, t') = e^{-\lambda(t+t')}$. As a further illustration, we give a lemma that will be used below.

A functional $\ell \in F'_2$ is called positive semi-definite if

$$(20) \quad \ell(\psi \otimes \psi) \geq 0 \quad \forall \psi \in F.$$

Lemma 2.2. *The function $U := B^{-1}\ell \in E_2$ is positive semi-definite in the sense of (8) if and only if the functional $\ell \in F'_2$ is positive semi-definite.*

Proof. Identifying $\varphi \in L_2(J)$ with $\psi \in F$ via $(w, \varphi)_{L_2(J)} = b(w, \psi)$ for all $w \in E$, we observe that $(U, \varphi \otimes \varphi)_{L_2(J \times J)} = B(U, \psi \otimes \psi) = \ell(\psi \otimes \psi)$. Thus U is positive semi-definite iff ℓ is. \square

Finally, we introduce the bounded linear functional

$$(21) \quad \delta: F_2 \rightarrow \mathbb{R}, \quad \delta(v) := \int_J v(t, t) dt.$$

As in [12, Lemma 4.1], [19, Lemma 5.1] could be used to show boundedness of δ . We give here an elementary quantitative argument. Writing $\delta(v)$ as the integral of $\delta(s - s')v(s, s')$ over $J \times J$ and exploiting the representation (15) of b^{-1} we find $(B^{-1}\delta)(t, t') = (e^{-\lambda|t-t'|} - e^{-\lambda(t+t')})/(2\lambda)$. Since B is an isometry, the operator norm of δ is

$$\|\delta\|_{-2} = \lambda \|B^{-1}\delta\|_{L_2(J \times J)} = \frac{1}{4\lambda} \left(4\lambda T - 5 + (8\lambda T + 4)e^{-2\lambda T} + e^{-4\lambda T} \right)^{1/2}.$$

In particular, this yields the asymptotics $\|\delta\|_{-2} \sim T^2\lambda/\sqrt{6}$ for small λ and $\|\delta\|_{-2} \sim \sqrt{T/(4\lambda)}$ for large λ . In addition, the uniform bound $\|\delta\|_{-2} \leq \frac{1}{2}T$ holds, see Remark 2.9.

We are now ready to state the following result (derived for stochastic PDEs in [12]).

Proposition 2.3. *The second moment $M = \mathbb{E}[X \otimes X]$ of the solution X to the stochastic ODE (1) with additive noise solves the deterministic variational problem*

$$(22) \quad \boxed{\text{Find } M \in E_2 \text{ s.t. } B(M, v) = \mathbb{E}[X_0^2]v(0) + \mu^2\delta(v) \quad \forall v \in F_2.}$$

Proof. Inserting the solution (3) in the first term of $b(X, v) = \int_J \{-Xv' + \lambda Xv\}$ and integrating it by parts one finds

$$b(X, v) = X_0v(0) - \mu \int_J W(t)v'(t) dt = X_0v(0) + \mu \int_J v(t) dW(t) \quad \forall v \in F,$$

where the stochastic integration by parts formula [13, Theorem 4.1.5] was used in the second equality. Employing this in $B(M, v_1 \otimes v_2) = \mathbb{E}[b(X, v_1)b(X, v_2)]$ with (7) for the μ^2 term leads to the desired conclusion. \square

From the equations for the first and second moments, an equation for the covariance function $\text{Cov}(X) \in E_2$ follows:

$$B(\text{Cov}(X), v) = \text{Cov}(X_0)v(0) + \mu^2 \delta(v) \quad \forall v \in F_2.$$

The proof is straightforward and is omitted.

2.4. Second moment equations: multiplicative noise. Before proceeding with the second moment equation for the multiplicative case we formulate a lemma, which repeats the derivation of the first moment equation (16) without taking the expectation first.

Lemma 2.4. *Let X be the solution (4) to the stochastic ODE (2). Then*

$$(23) \quad b(X, v) = X_0 v(0) - \rho \int_J \left(\int_0^t X(r) dW(r) \right) v'(t) dt \quad \forall v \in F \quad (\mathbb{P}\text{-a.s.}).$$

Proof. Let $v \in F$. We employ the definition (4) of the solution in the first term of $b(X, v)$ and integration by parts on the first two summands of the integrand to obtain (observing that the terms at $t = T$ vanish due to $v(T) = 0$)

$$\begin{aligned} \int_J X(t)v'(t) dt &= \int_J \left(X_0 - \int_0^t \lambda X(r) dr + \int_0^t \rho X(r) dW(r) \right) v'(t) dt \\ &= -X_0 v(0) + \lambda \int_J X(t)v(t) dt + \rho \int_J \left(\int_0^t X(r) dW(r) \right) v'(t) dt \quad (\mathbb{P}\text{-a.s.}). \end{aligned}$$

Inserting this expression in the definition (14) of $b(X, v)$ yields the claimed formula. \square

The next ingredient in the second moment equation for the multiplicative noise, which appears due to the integral term in (23), is the bilinear form

$$(24) \quad \Delta(w, v) := \int_J w(t, t)v(t, t) dt, \quad w \in E \otimes E, \quad v \in F \otimes F,$$

referred to as the trace product. Again, we use the same symbol for the induced operator, where convenient. Here, \otimes denotes the algebraic tensor product. The expression (24) is meaningful because functions in $F \subset H^1(J)$ are bounded. As we will see in Lemma 2.8, this bilinear form extends continuously to a form

$$(25) \quad \Delta: E_\pi \times F_\epsilon \rightarrow \mathbb{R}$$

on the projective and the injective tensor product spaces

$$(26) \quad E_\pi := E \otimes_\pi E \quad \text{and} \quad F_\epsilon := F \otimes_\epsilon F.$$

These spaces are defined as the closure of the algebraic tensor product under the projective norm

$$(27) \quad \|w\|_\pi := \inf \left\{ \sum_i \|w_i^1\|_E \|w_i^2\|_E : w = \sum_i w_i^1 \otimes w_i^2 \right\},$$

and the injective norm

$$(28) \quad \|v\|_\epsilon := \sup \{ |(g_1 \otimes g_2)(v)| : g_1, g_2 \in S(F') \},$$

respectively. Note that, initially, these norms are defined on the algebraic tensor product space. In particular, the sums in (27) are finite and the action of $g_1 \otimes g_2$ in (28) is well-defined. The spaces in (26) are separable Banach spaces. They are reflexive if and only if their dimension is finite [16, Theorem 4.21]. By [16, Proposition 6.1(a)], these tensor norms satisfy

$$(29) \quad \|w_1 \otimes w_2\|_\pi = \|w_1\|_E \|w_2\|_E \quad \text{and} \quad \|v_1 \otimes v_2\|_\epsilon = \|v_1\|_F \|v_2\|_F,$$

as well as

$$(30) \quad \|\cdot\|_2 \leq \|\cdot\|_\pi \quad \text{on} \quad E \otimes E \quad \text{and} \quad \|\cdot\|_\epsilon \leq \|\cdot\|_2 \quad \text{on} \quad F \otimes F.$$

We write $\|\cdot\|_{-\epsilon}$ for the norm of the continuous dual $F'_\epsilon := (F_\epsilon)'$.

Example 2.5. Consider $V := \mathbb{R}^N$ with the Euclidean norm. Elements $A \in V \otimes V$ can be identified with $N \times N$ real matrices. Let $\sigma(A)$ denote the singular values of A . The projective, the Hilbert, and the injective norms on $V \otimes V$ are the nuclear norm $\|A\|_\pi = \sum_{s \in \sigma(A)} s$, the Frobenius norm $\|A\|_2 = (\sum_{s \in \sigma(A)} s^2)^{1/2}$, and the operator norm $\|A\|_\epsilon = \max \sigma(A)$, respectively. They are also known as the Schatten p -norms with $p = 1, 2$, and ∞ . Note that $\|\cdot\|_\pi \geq \|\cdot\|_2 \geq \|\cdot\|_\epsilon$.

For a symmetric and positive semi-definite function $w \in E_2$ the operator defined by $S_w : E \rightarrow E$, $S_w \varphi := \int_J w(s, \cdot) \varphi(s) ds$ is self-adjoint and positive semi-definite. Let $\{s_n\}_n \subset [0, \infty)$ denote its eigenvalues. If $\sum_n s_n$ is finite then the operator is trace-class and $\|w\|_\pi = \sum_n s_n$, see [14, Theorem 9.1.38 and comments]. The following specialization will be useful.

Lemma 2.6. *If $w \in E_\pi$ is symmetric positive semi-definite then $\|w\|_\pi = \lambda \delta(w)$ with δ from (21).*

Proof. Let $\{e_n\}_n$ be an orthonormal basis of E consisting of eigenvectors of S_w corresponding to the eigenvalues $\{s_n\}_n$. By symmetry, $w = \sum_n s_n (e_n \otimes e_n)$. Since $\lambda \delta(e_n \otimes e_n) = 1$, we have $\lambda \delta(w) = \sum_n s_n = \|w\|_\pi$. \square

An arbitrary $w \in E_\pi$ can be decomposed (via the corresponding integral operator) as $w = w^+ - w^- + w^a$ with symmetric positive semi-definite $w^\pm \in E_\pi$ and an antisymmetric $w^a \in E_\pi$. This decomposition is stable in the sense that

$$(31) \quad \|w^a\|_\pi \leq \|w\|_\pi \quad \text{and} \quad \|w^+ - w^-\|_\pi = \|w^+\|_\pi + \|w^-\|_\pi \leq \|w\|_\pi.$$

The tensor product spaces E_π and F_ϵ will be necessary because the trace product Δ is *not* continuous on the Hilbert tensor product spaces $E_2 \times F_2$ as the following example illustrates.

Example 2.7. To simplify the notation, suppose $T = 1$, so that $J = (0, 1)$. Define $v \in F_2$ by $v(s, t) := (1-s)(1-t)$ for $s, t \in J$. Consider the sequence u_1, u_2, \dots of indicator functions

$$u_n(s, t) := \chi_{A_n}(s, t), \quad \text{where} \quad A_n := \left(0, \frac{1}{n}\right)^2 \cup \left(\frac{1}{n}, \frac{2}{n}\right)^2 \cup \dots \cup \left(\frac{n-1}{n}, 1\right)^2 \subset J \times J.$$

In view of the isometry (18), this sequence is a null sequence in E_2 . However, $\Delta(u_n, v) = \int_J u_n(t, t) v(t, t) dt = \frac{1}{3}$ for all $n \geq 1$. Therefore, $\Delta(\cdot, v)$ is not continuous on E_2 .

The example additionally shows that Δ is not continuous on $E_\epsilon \times F_\pi$ either, since by (29)–(30) we have $\|v\|_\pi = \|v\|_2$, while $\|u_n\|_\epsilon \leq \|u_n\|_2 \rightarrow 0$ as $n \rightarrow \infty$.

By contrast, $\{u_n\}_{n \geq 1}$ is not a null sequence in E_π . Indeed, Lemma 2.6 gives $\|u_n\|_\pi = \lambda$ for all $n \geq 1$.

Lemma 2.8. *The trace product Δ in (25) is continuous on $E_\pi \times F_\epsilon$ with $\|\Delta\| \leq 1/(2\lambda)$.*

Proof. By density it suffices to bound $\Delta(w, v)$ for arbitrary $w \in E \otimes E$ and $v \in F \otimes F$. By [17, Theorem 2.4] we may assume that $w = w^1 \otimes w^2$. We note first that the point evaluation functionals $\delta_t : v \mapsto v(t)$ have norm $1/\sqrt{2}$ on F by (12). Therefore, if $v = \sum_j v_j^1 \otimes v_j^2$ then

$$(32) \quad |v(s, t)| = \left| \sum_j \delta_s(v_j^1) \delta_t(v_j^2) \right| \leq \sup \left\{ \frac{1}{2} \left| \sum_j g_1(v_j^1) g_2(v_j^2) \right| : g_1, g_2 \in \mathcal{S}(F') \right\} = \frac{1}{2} \|v\|_\epsilon$$

and the continuity of Δ follows:

$$|\Delta(w, v)| = \left| \int_J w(t, t) v(t, t) dt \right| \leq \frac{1}{2} \|v\|_\epsilon \int_J |w(t, t)| dt \leq \frac{1}{2\lambda} \|v\|_\epsilon \|w\|_\pi,$$

where the integral Cauchy–Schwarz inequality on $w(t, t) = w^1(t)w^2(t)$ was used in the last step, together with the fact that $\lambda \|w^1\|_{L_2(J)} \|w^2\|_{L_2(J)} = \|w^1\|_E \|w^2\|_E = \|w^1 \otimes w^2\|_\pi$. \square

We point out that the bound $\|\Delta\| \leq 1/(2\lambda)$ is sharp in general. For $\eta > 0$ take $w = \varphi \otimes \varphi$ with $\varphi := \chi_{(0, \eta)}/\sqrt{\eta}$ and $v = \psi \otimes \psi$ with $\psi(t) := \sinh(\lambda(T-t))/\sinh(\lambda T)$ as in (13). Then $\lim_{\eta \rightarrow 0} \Delta(w, v) = 1$ and $\lim_{\lambda T \rightarrow \infty} \|v\|_\epsilon \|w\|_\pi = 2$, and the bound is tight when applying both limits.

Remark 2.9. *Consider the functional δ from (21). Since $\delta = \Delta(1 \otimes 1)$ and $\|1 \otimes 1\|_\pi = \lambda T$, we have $\|\delta : F_\epsilon \rightarrow \mathbb{R}\|_{-\epsilon} \leq T/2$. In view of $\|\cdot\|_\epsilon \leq \|\cdot\|_2$, see (30), we find $\|\delta : F_2 \rightarrow \mathbb{R}\|_{-2} \leq T/2$. Finally, $\|\delta : E_\pi \rightarrow \mathbb{R}\|_{-\pi} = 1/\lambda$ by the integral Cauchy–Schwarz inequality and Lemma 2.6.*

A crucial observation is that the second moment M lies not only in the Hilbert tensor product space E_2 but in the smaller projective tensor product space, $M \in E_\pi$. This follows by passing the norm under the expectation $\|\mathbb{E}[X \otimes X]\|_\pi \leq \mathbb{E}[\|X \otimes X\|_\pi]$, then using (29) and the square integrability (5d) of X .

We recall here from [17, Theorems 2.5 and 5.13] the fact that

$$F'_\epsilon = (F \otimes_\epsilon F)' \cong F' \otimes_\pi F' \quad \text{isometrically,}$$

(whereas the space $(F')_\epsilon$ is isometric to a proper subspace of $(F_\pi)'$, see [16, p. 46]). A corollary of this representation is that

$$(33) \quad b \otimes b : E_\pi \rightarrow F'_\epsilon \quad \text{defines an isometric isomorphism,}$$

because $b \otimes b$ extends to an isometric isomorphism from $E \otimes_\pi E$ onto $F' \otimes_\pi F'$. We denote it also by B . This isometry property (33), Lemma 2.2 and Lemma 2.6 produce the useful identity

$$(34) \quad \|\ell\|_{-\epsilon} = \|B^{-1}\ell\|_\pi = \lambda\delta(B^{-1}\ell)$$

for any positive semi-definite $\ell \in F'_\epsilon$ as in (20), if it is also symmetric:

$$(35) \quad \ell(\psi \otimes \phi) = \ell(\phi \otimes \psi) \quad \forall \psi, \phi \in F.$$

Here and below, Lemma 2.2 applies to functionals in F'_ϵ mutatis mutandis. Similarly, using the decomposition from (31) we can decompose any $\ell = \ell^+ - \ell^- + \ell^a$ into symmetric positive semi-definite and antisymmetric parts with

$$(36) \quad \|\ell^a\|_{-\epsilon} \leq \|\ell\|_{-\epsilon} \quad \text{and} \quad \|\ell^+ - \ell^-\|_{-\epsilon} = \|\ell^+\|_{-\epsilon} + \|\ell^-\|_{-\epsilon} \leq \|\ell\|_{-\epsilon}.$$

Now we are in position to introduce the bilinear form

$$(37) \quad \mathcal{B} : E_\pi \times F_\epsilon \rightarrow \mathbb{R}, \quad \mathcal{B} := B - \rho^2 \Delta,$$

or more explicitly,

$$\mathcal{B}(w, v) = \int_J \int_J w(s, t)(-\partial_s + \lambda)(-\partial_t + \lambda)v(s, t) ds dt - \rho^2 \int_J w(t, t)v(t, t) dt.$$

The reason for this definition is the following result from [9, Theorem 4.2] derived there for stochastic PDEs. The simplified proof is given here for completeness.

Proposition 2.10. *The second moment $M = \mathbb{E}[X \otimes X]$ of the solution X to the stochastic ODE (2) with multiplicative noise solves the deterministic variational problem*

$$(38) \quad \boxed{\text{Find } M \in E_\pi \text{ s.t. } \mathcal{B}(M, v) = \mathbb{E}[X_0^2]v(0) \quad \forall v \in F_\epsilon.}$$

Proof. It suffices to verify the claim for v of the form $v = v_1 \otimes v_2$ with $v_1, v_2 \in F$. The more general statement follows by linearity and continuity of both sides in $v \in F_\epsilon$. We first observe with Fubini's theorem on $\Omega \times J$ that $B(M, v_1 \otimes v_2) = B(\mathbb{E}[X \otimes X], v_1 \otimes v_2) = \mathbb{E}[b(X, v_1)b(X, v_2)]$. Next, we insert the expression (23) for both $b(X, v_j)$ and expand the product. The cross-terms vanish because the terms of the form $X_0 \int_0^t X(r) dW(r)$ vanish in expectation; this is seen by conditioning this term on \mathcal{F}_0 and employing the martingale property (6). With the identity (7) and $\mathbb{E}[X(r)^2] = M(r, r)$ we arrive at

$$B(M, v_1 \otimes v_2) = \mathbb{E}[X_0^2]v(0) + \rho^2 \int_J \int_J v'_1(s)v'_2(t) \int_0^{s \wedge t} M(r, r) dr ds dt.$$

It remains to verify that $\rho^2 \Delta(M, v)$ coincides with the last term on the right-hand side. Let us distinguish the two cases $s = s \wedge t$ and $t = s \wedge t$ and write that triple integral as

$$(39) \quad \int_J v'_1(s) \int_s^T v'_2(t) dt \int_0^s M(r, r) dr ds + \int_J v'_2(t) \int_t^T v'_1(s) ds \int_0^t M(r, r) dr dt.$$

Evaluating the dt integral in the first summand and the ds integral in the second summand, we see that $((39) - \Delta(M, v)) = \int_J \frac{d}{dt} \{-v_1(t)v_2(t) \int_0^t M(r, r) dr\} dt = 0$. Hence, $(39) = \Delta(M, v)$. This completes the proof. \square

Using the equations for the first and second moments we obtain an equation for the covariance function $\text{Cov}(X) \in E_\pi$ from (9):

$$(40) \quad \mathcal{B}(\text{Cov}(X), v) = \text{Cov}(X_0)v(0) + \rho^2 \Delta(m \otimes m, v) \quad \forall v \in F_\epsilon.$$

Identity (34) yields $\|v \mapsto v(0)\|_{-\epsilon} = \|\delta_0 \otimes \delta_0\|_{-\epsilon} = \frac{1}{2}(1 - e^{-2\lambda T})$ for the functional appearing on the right-hand side of (38) and (40). Similarly, $\|\Delta(m \otimes m)\|_{-\epsilon} = \frac{1}{2} \int_J (1 - e^{-2\lambda(T-t)}) |m(t)|^2 dt \leq \frac{1}{2\lambda} \|m\|_E^2$, providing some details on the estimate $\|\Delta\| \leq 1/(2\lambda)$ from Lemma 2.8.

We emphasize that it is not possible to replace in the present case of multiplicative noise the pair of trial and test spaces $E_\pi \times F_\epsilon$ by either pair $E_2 \times F_2$ or $E_\epsilon \times F_\pi$, because by Example 2.7 the operator Δ is not continuous there. We note, however, that in the case of additive noise (Section 2.3) the pair $E_\pi \times F_\epsilon$ could be used instead of $E_2 \times F_2$. Then $\|\delta\|_{-\epsilon} = \lambda \delta(B^{-1}\delta) = \frac{1}{4\lambda}(e^{-2T\lambda} - 1 + 2T\lambda)$ with the asymptotics $\frac{1}{2}T^2\lambda$ (small λ) and $\frac{1}{2}T$ (large λ).

In order to discuss the well-posedness of the variational problem (38), given a functional $\ell \in F'_\epsilon$, we consider the more general problem:

$$(41) \quad \text{Find } U \in E_\pi \text{ s.t. } \mathcal{B}(U, v) = \ell(v) \quad \forall v \in F_\epsilon.$$

Owing to $\|Bw\|_{-\epsilon} = \|w\|_\pi$ and $\|\Delta\| \leq 1/(2\lambda)$ we have $\|\mathcal{B}w\|_{-\epsilon} \geq (1 - \rho^2/(2\lambda))\|w\|_\pi$. Thus, injectivity of \mathcal{B} holds under the condition $\rho^2 < 2\lambda$ of small ‘‘volatility’’. A similar condition was imposed in [9, Theorem 5.5]. This is exactly the threshold for the second moment (5c) to diverge as $s = t \rightarrow \infty$, but it stays nevertheless finite for all finite $s = t$. We discuss here what happens in the variational formulation (41) for larger volatilities ρ , and summarize in Theorem 2.11 below.

Since B is an isomorphism, problem (41) is equivalent to $U = \rho^2 B^{-1} \Delta U + B^{-1} \ell$. Using the representation of $\Delta(U, v)$ as the double integral of $\delta(s - s')U(s, s')v(s, s')$, and the integral representation of B^{-1} through (15), we obtain the integral equation

$$(42) \quad U(t, t') = \rho^2 \int_0^{t \wedge t'} e^{-\lambda(t+t'-2s)} U(s, s) ds + (B^{-1} \ell)(t, t').$$

Defining $f(t) := (B^{-1} \Delta U)(t, t) = \int_0^t e^{-2\lambda(t-s)} U(s, s) ds$ and $g(t) := (B^{-1} \ell)(t, t)$ we find from (42) the ODE $f'(t) + 2\lambda f(t) = \rho^2 f(t) + g(t)$ with the initial condition $f(0) = 0$. The solution is

$$(43) \quad f(t) = (B^{-1} \Delta U)(t, t) = \int_0^t e^{-(2\lambda - \rho^2)(t-r)} g(r) dr.$$

Inserting

$$(44) \quad U(s, s) = \rho^2 f(s) + g(s) = \rho^2 \int_0^s e^{-(2\lambda - \rho^2)(s-r)} g(r) dr + g(s)$$

under the integral of (42) provides a unique candidate for U . Moreover, $U \in E_2$. We now estimate $\|U\|_\pi$ in terms of the norm of ℓ .

Clearly, not all functionals ℓ lead to solutions that are potential second moments. Let us therefore assume that ℓ is symmetric positive semi-definite (35)/(20). Then $B^{-1} \ell$ is positive semi-definite (8) by Lemma 2.2. In particular, $f \geq 0$ and $g \geq 0$. With $w := (\rho^2 f + g)^{1/2}$, the functional $v \mapsto \Delta(w \otimes w, v) = \int_J (\rho^2 f(t) + g(t)) v(t, t) dt$ is symmetric positive semi-definite. The function $U = \rho^2 B^{-1} \Delta(w \otimes w) + B^{-1} \ell$ inherits the definiteness (Lemma 2.2) as well as the symmetry from $\Delta(w \otimes w)$ and ℓ . Under those assumptions, Lemma 2.6 gives

$$(45) \quad \|U\|_\pi = \lambda \delta(U) = \rho^2 \lambda \delta(B^{-1} \Delta U) + \lambda \delta(B^{-1} \ell).$$

For the first term on the right-hand side of (45) we employ (43) as follows:

$$(46) \quad \delta(B^{-1}\Delta U) = \int_J g(r) \int_r^T e^{-(2\lambda-\rho^2)(s-r)} ds dr \leq \delta(B^{-1}\ell) \frac{e^{(\rho^2-2\lambda)T}-1}{\rho^2-2\lambda},$$

where we have exchanged the order of integration in the first step, evaluated the inner integral and used $g \geq 0$ with $\|g\|_{L_1(J)} = \delta(B^{-1}\ell)$ in the last step. The fraction evaluates to T in the limit $\rho^2 = 2\lambda$. Combining (45)–(46) and (34), we arrive at the following theorem.

Theorem 2.11. *Suppose that $\ell \in F'_\epsilon$ is symmetric and positive semi-definite (20). Then, for any $\rho \geq 0$ and $\lambda > 0$, the variational problem (41) has a unique solution $U \in E_\pi$. This solution is symmetric and positive semi-definite (8) and admits the bound*

$$(47) \quad \|U\|_\pi \leq C \|\ell\|_{-\epsilon} \quad \text{with} \quad C := \frac{\rho^2 e^{(\rho^2-2\lambda)T} - 2\lambda}{\rho^2 - 2\lambda},$$

where $C = \rho^2 T + 1$ for $\rho^2 = 2\lambda$.

The bound in (47) is sharp in general because for $\eta > 0$ and $\ell := \eta^{-1}B(\chi_{(0,\eta)} \otimes \chi_{(0,\eta)})$ we have $g = \eta^{-1}\chi_{(0,\eta)}$ in (46), and the inequality in (46) approaches an equality as $\eta \searrow 0$.

For a general functional $\ell \in F'_\epsilon$, we decompose $\ell = \ell^+ - \ell^- + \ell^a$ as in (36). The corresponding solutions $U^\pm := \mathcal{B}^{-1}\ell^\pm$ and $U^a := \mathcal{B}^{-1}\ell^a = B^{-1}\ell^a$ (noting that $\Delta U^a = 0$ by antisymmetry) satisfy the bounds $\|U^\pm\|_\pi \leq C \|\ell^\pm\|_{-\epsilon}$ and $\|U^a\|_\pi = \|\ell^a\|_{-\epsilon}$. By linearity, $U := U^+ - U^- + U^a$ is the solution to (41), and the estimate $\|U\|_\pi \leq C(\|\ell^+\|_{-\epsilon} + \|\ell^-\|_{-\epsilon}) + \|\ell^a\|_{-\epsilon} \leq (C+1)\|\ell\|_{-\epsilon}$ follows by triangle inequality in the first step and by (36) in the last step.

In contrast to Lemma 2.2, the solution U to (41) may be symmetric and positive semi-definite even though the right-hand side ℓ is not. Indeed, for any $(w, v) \in E \times F$ with $\Delta(w \otimes w, v \otimes v) = \int_J |w(t)v(t)|^2 dt \neq 0$, the expression $\mathcal{B}(w \otimes w, v \otimes v) = |b(w, v)|^2 - \rho^2 \Delta(w \otimes w, v \otimes v)$ is negative for sufficiently large ρ .

The variational formulation (40) for the covariance function is of the form (41) for the functional $\ell := \text{Cov}(X_0)(\delta_0 \otimes \delta_0) + \rho^2 \Delta(m \otimes m)$.

The proof of the above theorem highlights the special status of the diagonal $t \mapsto U(t, t)$. First, it is determined by an integral equation. Second, the projective norm (45) only “looks” at the diagonal when U is symmetric and positive semi-definite. These insights will guide the development of the numerical methods below.

3. CONFORMING DISCRETIZATIONS OF THE DETERMINISTIC EQUATIONS

3.1. Orientation. In Section 2 we have derived deterministic variational formulations for the first and second moments of the stochastic processes (3) and (4). In particular, the first moment satisfies a known “weak” variational formulation of an ODE. To our knowledge, [4, 5] were the first to discuss the numerical analysis of conforming finite element discretizations of a space-time variational formulation for linear parabolic PDEs. The problem was first reduced to the underlying family of ODEs parameterized by the spectral parameter λ . With the notation from Section 2.2 for the bilinear form b and the spaces E and F , the solution u to such an ODE is characterized by a well-posed variational problem of the above form (16), with a general right-hand side ℓ . The temporal discretization analyzed in [5] was of the conforming type, employing discontinuous piecewise polynomials as the discrete trial space for u and continuous piecewise polynomials of one degree higher as the discrete test space for v . The analysis in essence revealed that the discretization is *not* uniformly stable (in the Petrov–Galerkin sense, as discussed below) in the choice of the discretization parameters such as the polynomial degree and the location of the temporal nodes [5, Theorem 2.2.1].

The same question of stability of was taken up in [2] for a “strong” space-time variational formulation of linear parabolic PDEs and for the two classes of discretizations, of Gauss–Legendre (e.g., Crank–Nicolson, CN) or Gauss–Radau (e.g., implicit Euler, iE) type. It was confirmed that both types are in general only *conditionally* space-time stable, but the Gauss–Radau type can be

made *unconditionally* stable under mild restrictions on the temporal mesh. We will first revisit the simplest representative of each group adapted to the present variational formulation. The adaptation consists in switching the roles of the discrete trial and test spaces and by reversing the temporal direction, the latter due to the integration by parts that was used in the derivation of the variational formulation (16). The resulting adjoint discretizations will therefore be denoted by CN^* and iE^* , respectively. The CN^* discretization is thus a special case of the discretizations analyzed in [5].

In summary, in Section 3.2 we will discuss two conforming discretizations for the deterministic first moment equation (16): CN^* which is only conditionally stable (depending on the spectral parameter λ) and iE^* which is stable under a mild condition on the temporal mesh (comparable size of neighboring temporal elements). Both employ discontinuous trial spaces but iE^* requires additional discussion due to the somewhat unusual shape functions, whereby the discrete trial spaces are not nested and therefore do not generate a dense subspace in the usual sense. The situation transfers with no surprises to the second moment equations with additive noise (22) by tensorizing the discrete trial/test spaces. The case of multiplicative noise (38), however, presents a significant twist due to:

- (1) the presence of the Δ term in the definition (37) of the bilinear form \mathcal{B} . We will see that CN^* interacts naturally with the Δ operator while iE^* requires a modification to restore the expected convergence order.
- (2) the non-Hilbertian nature of the trial and test spaces in (38).

We will then provide a common framework for both discretizations, generalizing to arbitrary polynomial degrees. This will allow us to use the unconditionally stable Gauss–Radau discretization family without resorting to the modification of the lowest-order iE^* discretization because the discrete trial spaces with higher polynomial degree do generate a dense subspace.

In Section 3.4 we construct discretizations on tensor product spaces and comment on their stability. In Section 3.5, they are applied to the variational problem (22) for the second moment in the additive case.

In the multiplicative case we obtained existence and stability of the exact solution for arbitrary $\rho \geq 0$ in Theorem 2.11, even beyond the trivial range $0 \leq \rho^2 < 2\lambda$. The situation is similar in the discrete setting, where this trivial range is reduced by the discrete inf-sup constant γ_k to $0 \leq \rho^2 < 2\lambda\gamma_k^2$. In Section 3.6 we will therefore investigate, for the low order CN^* and iE^* schemes and some of their variants, whether stability holds for all $\rho \geq 0$. The behavior of the high order discretizations beyond the trivial stability range remains an open question.

3.2. First moment discretization. We are using the notation from Section 2.2. Let us consider the general formulation of (16) as the variational problem

$$(48) \quad \text{Find } u \in E \quad \text{s.t.} \quad b(u, v) = \ell(v) \quad \forall v \in F$$

with some bounded linear functional $\ell \in F'$. Recall that the E and F carry the λ -dependent norms (10) that render $b : E \rightarrow F'$ an isometric isomorphism. This variational problem is formally obtained by testing the real-valued ODE

$$(49) \quad u'(t) + \lambda u(t) = f \quad \text{on } J = (0, T), \quad u(0) = g,$$

with a test function v , integrating over J , moving the derivative from u' to v via integration by parts and then replacing the exposed $u(0)$ by the given initial datum g . The corresponding right-hand side then reads as $\ell(v) := \int_J \langle f, v \rangle dt + \langle g, v(0) \rangle$. We write $\langle \cdot, \cdot \rangle$ for the simple multiplication to emphasize the structure of the problem and to facilitate the transition to vector-valued ODEs.

For the discretization of the variational problem (48) we need to define subspaces

$$E^k \subset E \quad \text{and} \quad F^k \subset F$$

of the same (nontrivial) finite dimension. We then consider the discrete variational problem

$$(50) \quad \text{Find } u^k \in E^k \quad \text{s.t.} \quad b(u^k, v) = \ell(v) \quad \forall v \in F^k.$$

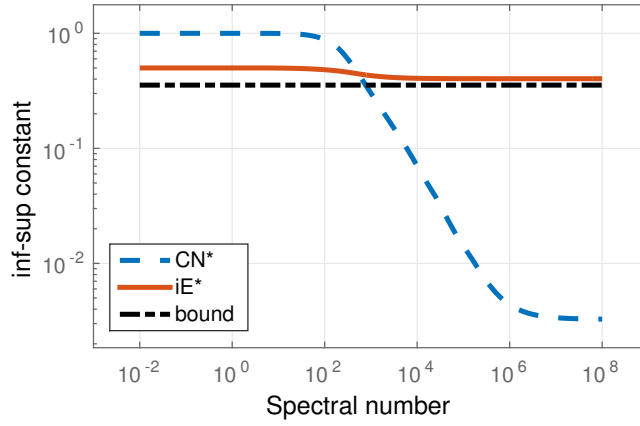


FIGURE 1. The inf-sup constant (51) for the CN^* and the iE^* discretizations on the same “random” temporal mesh of the interval $(0, 1)$ with 210 nodes and backward successive temporal element ratio $\sigma \leq 3$ in (57). The bound shown is the estimate from (58).

The well-posedness of this discrete problem is quantified by the discrete inf-sup constant

$$(51) \quad \gamma_k := \inf_{w \in S(E^k)} \sup_{v \in S(F^k)} b(w, v) > 0,$$

since the norm of the discrete data-to-solution mapping $\ell|_{F^k} \mapsto u_k$ equals $1/\gamma_k$. Moreover, the quasi-optimality estimate

$$(52) \quad \|u - u^k\|_E \leq (\|b\|/\gamma_k) \inf_{w \in E^k} \|u - w\|_E$$

holds [20, Theorem 2], where in fact $\|b\| = 1$. We call a family $\{E^k \times F^k\}_{k \geq 0}$, of discretization pairs uniformly stable if $\inf_{k \geq 0} \gamma_k > 0$. To construct $E^k \times F^k$ we introduce a temporal mesh

$$(53) \quad \mathcal{T} := \{0 =: t_0 < t_1 < \dots < t_N := T\}$$

subdividing $J = (0, T)$ into N temporal elements. Below, the dependence on \mathcal{T} is implicit in the notation. We write

$$J_n := (t_{n-1}, t_n) \quad \text{and} \quad k_n := |t_n - t_{n-1}|, \quad n = 1, \dots, N.$$

As announced above, we first discuss the simplest representatives of the Gauss–Legendre and Gauss–Radau discretizations in §3.2.1–§3.2.2, which are the CN^* and the iE^* schemes. For both methods, the discrete test space $F^k \subset F$ is defined as the spline space of continuous piecewise affine functions v with respect to the temporal mesh \mathcal{T} such that $v(T) = 0$. A common framework is the subject of §3.2.3.

3.2.1. The CN^* discretization. For the discrete trial space $E^k \subset E$, the space of piecewise constant functions with respect to \mathcal{T} seems a natural choice. We call this discretization CN^* in reference to the reversal of the roles of the trial and test spaces compared to the usual Crank–Nicolson time-stepping scheme. Unfortunately, if we keep the temporal mesh \mathcal{T} fixed, the discrete inf-sup constant (51) of the couple $E^k \times F^k$ depends on the spectral parameter λ , see Figure 1. This was already observed in [5, Equation (2.3.10)]. It can be shown along the lines of [2] that $\gamma_k \gtrsim (1 + \min\{\sqrt{\lambda T}, \text{CFL}\})^{-1}$, where $\text{CFL} := \max_n k_n \lambda$ is the parabolic CFL number. The three-phase behavior of the CN^* scheme in Figure 1 can be intuitively understood as follows: Consider $b(w, v) = \int_J (-v' + \lambda v)w$ from (51). For any $w \in E^k$ we can find a $v \in F^k$ such that $-v' = w$, so that at sufficiently low spectral numbers λ , the estimate $\gamma_k \geq 1 - \epsilon$ is evident. For large λ , the function $-v' + \lambda v$ is, up to relatively small jumps, a piecewise linear continuous one. Such functions approximate a general piecewise constant w poorly, see [5, Equation (2.3.10)].

This behavior renders the method less useful for parabolic PDEs because following a spatial semi-discretization a low parabolic CFL number has to be maintained for uniform stability.

3.2.2. *The iE^{*} discretization.* To obtain stability under only mild restrictions we adapt an idea from [2]; for the sake of a self-contained exposition and sharp results we confine the discussion first to the lowest order case. We take E^k as the space of functions $w \in L_2(J)$ for which each $w|_{J_n}$ is a dilated translate of the shape function $\phi : s \mapsto (4 - 6s)$ from the reference temporal element $(0, 1)$ to the temporal element $J_n = (t_{n-1}, t_n)$. We refer to this combination of $E^k \times F^k$ as iE^{*} (adjoint implicit Euler). The explanation for this definition is the following. Consider the adjoint (backward) ODE

$$(54) \quad -v' + \lambda v = f, \quad v(T) = 0,$$

with a given f that for the sake of argument is piecewise affine with respect to \mathcal{T} . Define the approximate continuous piecewise affine solution $v \in F^k$ (hence, $v(T) = 0$) through the implicit Euler time-stepping scheme *backward in time*:

$$(55) \quad -\frac{1}{k_n}(v(t_n) - v(t_{n-1})) + \lambda v(t_{n-1}) = f(t_{n-1}^+), \quad n = N, \dots, 1,$$

where t_{n-1}^+ denotes the limit from above. We shall use the obvious abbreviations v_n and f_{n-1}^+ when referring to (55). The definition of the discrete trial space E^k implies that the time-step condition (55) is equivalent to the variational requirement

$$(56) \quad \int_{J_n} \langle w, -v' + \lambda v - f \rangle dt = 0 \quad \forall w \in E^k \quad \forall n = N, \dots, 1.$$

The equivalence is due to the identity $\int_0^1 \phi(s)(as + b) ds = b$ for all real a and b , which implies that the integral in (56) is a multiple of $(-v' + \lambda v - f)(t_{n-1}^+)$.

The role of the adjoint ODE (54) is elucidated in the proof of the following proposition concerning the inf-sup condition (51) for the iE^{*} discretization. The result is formulated in terms of the backward successive temporal element ratio

$$(57) \quad \sigma := \max_{n=1, \dots, N-1} k_n/k_{n+1}.$$

Proposition 3.1. *The inf-sup condition (51) holds for the iE^{*} discretization with*

$$(58) \quad \gamma_k \geq \gamma_\sigma := 1/\sqrt{2(1 + \max\{1, \sigma\})},$$

uniformly in $\lambda > 0$.

Thus, in order to obtain uniform stability of the iE^{*} discretization it suffices to ensure that the backward successive temporal element ratio (57) stays bounded. This is verified numerically in Figure 1. We generated an initial temporal mesh for $T = 1$ with 129 nodes by distributing the inner nodes in interval $(0, 1)$ uniformly at random. New nodes were inserted by subdividing large temporal elements into two equal ones until $\sigma \leq 3$, leading to a temporal mesh with 210 nodes. On this new temporal mesh, we observe that the inf-sup constant of the iE^{*} discretization is controlled as in (58), while that of CN^{*} depends strongly on the spectral parameter λ , as already explained in Section 3.2.1.

Proof of Proposition 3.1. Let $w \in E^k$ be arbitrary nonzero. We will find a discrete $v \in F^k$ such that $b(w, v) \geq \gamma_\sigma \|w\|_E \|v\|_F$. To this end, consider the adjoint ODE (54) with $f := \lambda w$. If we took v as the exact solution we would obtain $b(w, v) = \|w\|_E^2 = \lambda^{-1} \| -v' + \lambda v \|_{L_2(J)}^2 = \|v\|_F^2$. However, the exact solution is not necessarily an element of the discrete test space F^k , so we take $v \in F^k$ according to the implicit Euler scheme (55) instead. By the equivalence of (55)–(56) we see that $b(w, v) = \int_J \langle w, -v' + \lambda v \rangle dt = \int_J \langle w, \lambda w \rangle dt = \|w\|_E^2$ still holds.

To conclude, it is enough to establish $\|w\|_E \geq \gamma_\sigma \|v\|_F$. For this purpose, we square (55) with $f := \lambda w$ on both sides and rearrange to obtain

$$(59) \quad \lambda^{-1} k_n^{-1} |v_n - v_{n-1}|^2 + \lambda k_n |v_{n-1}|^2 + |v_n - v_{n-1}|^2 + |v_{n-1}|^2 - |v_n|^2 = \lambda k_n |w_{n-1}^+|^2.$$

Let Iv denote the piecewise constant function with $Iv(t_{n-1}^+) = v(t_{n-1})$ for all $n = 1, \dots, N$. We introduce the mesh-dependent norm

$$(60) \quad \|\|v\|\|_F^2 := \|v'\|_{E'}^2 + \|Iv\|_E^2 + |v(0)|^2 + \sum_{n=1}^N |v_n - v_{n-1}|^2$$

and sum up (59) over n . This yields the equality $\|w\|_E = \|\|v\|\|_F$, since $\int_0^1 |\phi(s)|^2 ds = 4 = \frac{1}{4}|\phi(0)|^2$. With σ from (57) we obtain the estimate (the last term is omitted for $n = N$)

$$(61) \quad \|v\|_{L_2(J_n)}^2 \leq \frac{1}{2}k_n(|v_{n-1}|^2 + |v_n|^2) \leq \frac{1}{2}\|Iv\|_{L_2(J_n)}^2 + \frac{1}{2}\sigma\|Iv\|_{L_2(J_{n+1})}^2.$$

Summation over n yields $\|v\|_F^2 \leq 2(1 + \max\{1, \sigma\})\|\|v\|\|_F^2$. In concatenation, $\|w\|_E = \|\|v\|\|_F \geq \gamma_\sigma\|v\|_F$, as anticipated. \square

The choice of the shape function $\phi : s \mapsto (4 - 6s)$ in the trial space E^k defining the iE^* discretization leads to uniform stability as discussed above. In view of the quasi-optimality estimate (52) we need to address the approximation properties of this trial space E^k . Unfortunately, we do not have nestedness $E^k \subset E^{k+1}$. Moreover, no matter how fine the temporal mesh, E^k does not approximate the constant function. To be precise, let L_d denote the L_2 -orthonormal Legendre polynomial (normalized to $L_d(1) = \sqrt{1 + 2d}$) of degree $d \geq 0$ on the reference interval $(0, 1)$. For real a, b , set $u := aL_0 + bL_1 + r$, where r is E -orthogonal to L_0 and L_1 . The E -orthogonal projection of u onto the span of the shape function $\phi = L_0 - \sqrt{3}L_1$ is $w := c\phi$ with $c = \frac{1}{4}(a - \sqrt{3}b)$. The error $\|u - w\|_E^2 = \lambda\frac{1}{4}|\sqrt{3}a + b|^2 + \|r\|_E^2$ may be large, for example, if u is constant.

3.2.3. Common framework. On each element of the temporal mesh \mathcal{T} in (53) let $\mathcal{N}_n \subset [t_{n-1}, t_n]$ be a set of $p \geq 1$ collocation nodes (we choose the same p for all n for simplicity). The compound element-wise interpolation operator based on these collocation nodes \mathcal{N}_n is denoted by I . As the discrete test space $F^k \subset F$, we take the subspace of piecewise polynomials of degree p with respect to \mathcal{T} . We introduce $I^* : IF^k \rightarrow F^k$ by $(I^*\cdot)_{L_2(J)} = (\cdot, I^*\cdot)_{L_2(J)}$ on $F^k \times IF^k$. The discrete trial space is then defined as $E^k := I^*IF^k$. Note that $\dim E^k = \dim F^k$ holds.

We are interested in two types of nodes: Gauss–Legendre nodes and (left) Gauss–Radau nodes, to which we refer as GL_p and GR_p^\leftarrow , respectively. All temporal elements host the same type of nodes. The lowest-order examples are $\mathcal{N}_n = \{\frac{1}{2}(t_{n-1} + t_n)\}$ for GL_1 and $\mathcal{N}_n = \{t_{n-1}\}$ for GR_1^\leftarrow , corresponding to the CN^* and iE^* schemes. The shape functions on the reference element $(0, 1)$ for the space $E^k = I^*IF^k$ are (cf. [2, Section 2.3])

- (1) the Legendre polynomials L_0, \dots, L_{p-1} for GL_p , and
- (2) the Legendre polynomials L_0, \dots, L_{p-2} together with $L_{p-1} - \frac{L_p(1)}{L_{p-1}(1)}L_p$ for GR_p^\leftarrow .

In particular, for $p \geq 2$, the GR_p^\leftarrow family contains the piecewise constant functions, which means that any function in E can be approximated to arbitrary accuracy upon mesh refinement.

Define the mesh-dependent norm $\|\| \cdot \|\|_F$ by

$$\|\|v\|\|_F^2 := \|v'\|_{E'}^2 + \|Iv\|_E^2 + |v(0)|^2 + \begin{cases} 0 & \text{for } GL_p, \\ \sum_{n=1}^N [v - Iv]_{\rightarrow n}^2 & \text{for } GR_p^\leftarrow, \end{cases}$$

where $[f]_{\rightarrow n}$ denotes $\lim_{t \rightarrow t_n^-} f(t)$. This is the generalization of (60).

Following [2, Proof of Theorem 3.3], we can now show:

Lemma 3.2. *For any $v \in F^k$ there exists a nonzero $w \in E^k = I^*IF^k$ such that*

$$(62) \quad b(w, v) \geq \|(I^*)^{-1}w\|_E \|\|v\|\|_F.$$

Proof. The space $IF^k \subset E$ carries the norm of E . Let $v \in F^k$. We first show that $\|\Gamma v\|_E = \|\|v\|\|_F$, where $\Gamma : F^k \rightarrow IF^k$ is defined by

$$(\Gamma v, \tilde{w})_E = b(I^*\tilde{w}, v) \quad \forall (v, \tilde{w}) \in F^k \times IF^k.$$

To this end, we expand $\|\Gamma v\|_E^2 = \|\Gamma v - Iv\|_E^2 + 2(\Gamma v, Iv)_E - \|Iv\|_E^2$. For the first term we have

$$\|\Gamma v - Iv\|_E = \sup_{\tilde{w} \in S(IF^k)} (\Gamma v - Iv, \tilde{w})_E = \sup_{\tilde{w} \in S(IF^k)} \{b(I^* \tilde{w}, v) - (Iv, \tilde{w})_E\} = \|v'\|_{E'}.$$

For the second term, we use the definition of Γ , followed by [2, Lemma 3.1]:

$$(\Gamma v, Iv)_E = \|Iv\|_E^2 - (Iv, v')_{L_2(J)} = \|Iv\|_E^2 + \frac{1}{2}|v(0)|^2 + \begin{cases} 0 & (\text{GL}_p), \\ \frac{1}{2} \sum_{n=1}^N [v - Iv]_{\rightarrow n}^2 & (\text{GR}_p^{\leftarrow}). \end{cases}$$

Hence, $\|\Gamma v\|_E = \|v\|_F$. Now take $\tilde{w} := \Gamma v$. Then $b(I^* \tilde{w}, v) = (\Gamma v, \tilde{w})_E = \|\Gamma v\|_E^2 = \|\tilde{w}\|_E \|v\|_F$. The claim (62) follows for $w := I^* \tilde{w}$. \square

In order to convert (62) to a statement with the original norms, we need to compare those norms. First, it can be shown as in [2, Section 3.2.2] that $\|w\|_E \leq \|I^*\| \| (I^*)^{-1} w \|_E \leq 2 \| (I^*)^{-1} w \|_E$.

Second, we need to quantify $\|v\|_F \lesssim \|v\|_F$. For the Gauss–Radau family GR_p^{\leftarrow} we can, for example, use the estimate (akin to (61)); see [2, Section 3.4])

$$\|v - Iv\|_{L_2(t_{n-1}, t_n)}^2 \leq \frac{2p^2}{4p-1/p} \left(\|Iv\|_{L_2(t_{n-1}, t_n)}^2 + \frac{k_n}{k_{n+1}} \|Iv\|_{L_2(t_n, t_{n+1})}^2 \right)$$

to derive $\|v\|_F \leq C \sqrt{p(1+\sigma)} \|v\|_F$ with the backward successive temporal element ratio σ from (57) and a universal constant $C > 0$. Therefore, the discrete inf-sup condition (51) holds for the GR_p^{\leftarrow} family with

$$(63) \quad \gamma_k \geq \gamma_0 / \sqrt{p(1+\sigma)},$$

where $\gamma_0 > 0$ is a constant independent of all parameters. The Gauss–Legendre family GL_p suffers from the same potential instability as the CN^* scheme, see §3.2.1.

Consider now the solution u^k to (50). From the ODE (49), the reconstruction

$$\hat{u}^k := g + \int_0^t \{f(s) - \lambda u^k(s)\} ds$$

can be expected to provide a better approximation of the exact solution. With (50) we find the orthogonality property $(\hat{u}^k - u^k, v')_E = 0$ for all $v \in F^k$. Let

$$(64) \quad q_k : E \rightarrow \partial_t F^k$$

be the orthogonal projection (in E or in $L_2(J)$). The orthogonality property gives $q_k \hat{u}^k = q_k u^k$. Hence, the postprocessed solution $\bar{u}^k := q_k u^k$ is an approximation of the reconstruction \hat{u}^k . In the case of Gauss–Legendre collocation nodes, I^* is the identity, so that $E^k = IF^k$, and therefore $q_k u^k = u^k$ has no effect. In the Gauss–Radau case, however, the projection is useful to improve the convergence rate upon mesh refinement, as will be seen in §3.6.4.

Note that q_k is injective on E^k in both cases. In the Gauss–Radau case, q_k^{-1} sends the shape function L_{p-1} to $L_{p-1} - \frac{L_p(1)}{L_{p-1}(1)} L_p$. Since $L_d(1) = \sqrt{2d+1}$, this gives

$$(65) \quad \|q_k^{-1}\|^2 = 1 + \frac{2p+1}{2(p-1)+1}.$$

3.3. Petrov–Galerkin approximations. In this subsection we comment on Petrov–Galerkin discretizations of the generic linear variational problem

$$\text{Find } u \in X : \quad \langle Bu, v \rangle = \langle \ell, v \rangle \quad \forall v \in Y,$$

where X and Y are *normed vector spaces*. This generality (that can also be found e.g. in [18]) will allow us to address the variational problem (2.10).

We assume that $X_h \times Y_h \subset X \times Y$ are finite-dimensional subspaces with nonzero $\dim X_h = \dim Y_h$. Here, h refers to the “discrete” nature of those subspaces, and the pair $X_h \times Y_h$ is fixed. We write $\|\cdot\|_{Y'_h} := \sup_{v \in S(Y_h)} |\langle \cdot, v \rangle|$.

In order to admit variational crimes we suppose that we have access to an operator $\bar{B} : X \rightarrow Y'$ that approximates B (although $\bar{B} : X \rightarrow Y'_h$ suffices). For this approximation we assume the

discrete inf-sup condition in the form of a constant $\bar{\gamma}_h > 0$ such that $\|\bar{B}w_h\|_{Y'_h} \geq \bar{\gamma}_h \|w_h\|_X$ for all $w_h \in X_h$. The proof of the following Proposition is obtained by standard arguments (for the discussion of the constant “1+” see [20, 18, 1]).

Proposition 3.3. *Fix $u \in X$. Under the above assumptions there exists a unique $u_h \in X_h$ such that*

$$\langle \bar{B}u_h, v_h \rangle = \langle Bu, v_h \rangle \quad \forall v_h \in Y_h.$$

Then $u \mapsto u_h$ is linear with $\|u_h\|_X \leq \bar{\gamma}_h^{-1} \|Bu\|_{Y'_h}$, and satisfies the quasi-optimality estimate

$$\|u - u_h\|_X \leq (1 + \bar{\gamma}_h^{-1} \|\bar{B}\|) \inf_{w_h \in X_h} \|u - w_h\|_X + \bar{\gamma}_h^{-1} \|(B - \bar{B})u\|_{Y'_h}.$$

3.4. Tensorized discretizations. Recall the definition of the tensor product spaces $E_{2/\pi}$ and $F_{2/\epsilon}$ from (17) and (26). Recall also that we can extend $B := (b \otimes b)$ to an isometric isomorphism $B: E_2 \rightarrow F'_2$ or $B: E_\pi \rightarrow F'_\epsilon$. We discuss here these two viewpoints in parallel. Consider the variational formulation

$$(66) \quad \text{Find } U \in E_{2/\pi} \quad \text{s.t.} \quad B(U, v) = \ell(v) \quad \forall v \in F_{2/\epsilon},$$

where $\ell \in F'_{2/\epsilon}$. If $E^k \times F^k$ is a discretization for (48) then the tensorized discretization

$$(67) \quad E_{2/\pi}^k \times F_{2/\epsilon}^k := (E^k \otimes E^k) \times (F^k \otimes F^k) \subset E_{2/\pi} \times F_{2/\epsilon}$$

is a natural discretization choice for (66). The subscript 2 or π (and 2 or ϵ) indicates which norm the algebraic tensor product $E^k \otimes E^k$ (and $F^k \otimes F^k$) is equipped with; since these spaces are finite-dimensional, no norm-closure is necessary.

We now turn to the discrete variational formulation

$$(68) \quad \text{Find } U^k \in E_{2/\pi}^k \quad \text{s.t.} \quad B(U^k, v) = \ell(v) \quad \forall v \in F_{2/\epsilon}^k.$$

The inf-sup constant required in the analysis is the square γ_k^2 of the discrete inf-sup constant γ_k from (51) in both cases:

$$(69) \quad \inf_{w \in S(E_2^k)} \sup_{v \in S(F_2^k)} B(w, v) = \gamma_k^2 = \inf_{w \in S(E_\pi^k)} \sup_{v \in S(F_\epsilon^k)} B(w, v).$$

Indeed, consider the π/ϵ situation. For $w \in E^k$ let $b_k w$ denote the restriction of bw to F^k . The discrete inf-sup condition (51) says that $b_k: E^k \rightarrow (F^k)'$ is an isomorphism with $\|b_k^{-1}\| = \gamma_k^{-1}$. The mapping $B_k := b_k \otimes b_k: E^k \otimes_\pi E^k \rightarrow (F^k)' \otimes_\pi (F^k)'$ has the inverse $b_k^{-1} \otimes b_k^{-1}$. It is therefore an isomorphism with $\|B_k^{-1}\| = \gamma_k^{-2}$. The identification $(F^k)' \otimes_\pi (F^k)' \cong (F_\epsilon^k)'$ shows that for any $w \in E_\pi^k$, the functional $B_k w$ is the restriction of Bw to F_ϵ^k . This gives (69).

Proposition 3.3 (with $\bar{B} := B$) provides a unique solution $U^k \in E^k \otimes E^k$ to the discrete variational problem (68) that approximates the solution U of (66) as soon as $\gamma_k > 0$ in (51). The solution is, moreover, quasi-optimal (recall that $\|B\| = 1$):

$$(70) \quad \|U - U^k\|_{2/\pi} \leq (1 + \gamma_k^{-2}) \inf_{w \in E^k \otimes E^k} \|U - w\|_{2/\pi}.$$

We will also be interested in the postprocessed solution $\bar{U}^k := (q_k \otimes q_k)U^k$, where $q_k: E \rightarrow \partial_t F^k$ is the orthogonal projection in (64).

Analogously to Lemma 2.2 one proves:

Lemma 3.4. *The discrete solution U^k to (68) is positive semi-definite (8) if and only if ℓ is positive semi-definite on $F^k \otimes F^k$. The same is true for the postprocessed solution.*

3.5. Second moment discretization: additive noise. In view of the previous section, any discretization pair $E^k \times F^k$ satisfying the discrete inf-sup condition (51) induces a valid discretization of the variational problem (22) for the second moment of the solution process to the stochastic ODE with additive noise (1) if we choose the trial space as $E^k \otimes E^k$ and the test space as $F^k \otimes F^k$. The functional on the right-hand side of (66) is then $\ell := \mathbb{E}[X_0^2](\delta_0 \otimes \delta_0) + \mu^2 \delta$. Moreover, the discrete solution satisfies the quasi-optimality estimates in (70) simultaneously with respect to $\|\cdot\|_2$ and $\|\cdot\|_\pi$, because $\ell \in F'_\epsilon \subset F'_2$.

3.6. Second moment discretization: multiplicative noise. As in the continuous case for sufficiently small values of the volatility ρ , namely in the range

$$(71) \quad 0 \leq \rho^2 < 2\lambda\gamma_k^2,$$

we immediately obtain a discrete inf-sup condition for the operator $B - \rho^2\Delta$. The purpose of this section is to address the whole range $\rho > 0$.

We will focus on the CN^* and iE^* discretizations discussed in §3.2.1–§3.2.2, although with some work, our methods may be adapted to higher-order schemes from §3.2.3. Throughout, we assume that the discretization pair $E^k \times F^k \subset E \times F$ satisfies the discrete inf-sup condition (51). The discrete trial and test spaces $E_\pi^k \times F_\epsilon^k \subset E_\pi \times F_\epsilon$ are defined as in (67).

We introduce some more notation. In what follows, the default range of the indices is

$$0 \leq i, j \leq N - 1 \quad \text{and} \quad 1 \leq m, n \leq N.$$

Recall that the discrete test space $F^k \subset F$ consists of continuous piecewise affine functions with respect to the temporal mesh \mathcal{T} in (53) that vanish at the terminal time T . It is equipped with the hat function basis $\{v_i\}_i$, determined by $v_i(t_j) = \delta_{ij}$. The basis functions $\{e_n\}_n$ of the discrete trial space $E^k \subset E$ are supported on $\text{supp}(e_n) = [t_{n-1}, t_n]$ in both schemes. Specifically, e_n is a constant for CN^* and is a dilated translate of the shape function $\phi : s \mapsto (4 - 6s)$ for iE^* . The following statements do not depend on the scaling of the basis functions, if not specified otherwise.

3.6.1. The discrete problem. In the multiplicative case, the trace product Δ from (24) appears in the variational problem (38) for the second moment. The basis functions $\{e_n\}_n \subset E^k$ for the iE^* discretization lead to an inconsistency in the Δ term, see §3.6.5. For this reason, we introduce the approximate trace product

$$(72) \quad \Delta^k : E_\pi \times F_\epsilon \rightarrow \mathbb{R},$$

to be specified below. We require that Δ^k reproduces the following properties of the exact trace product Δ :

(i) *Symmetry and definiteness:* for every symmetric positive semi-definite $w \in E_\pi^k$, the functional $\Delta^k w$ is symmetric and positive semi-definite on $F^k \otimes F^k$, i.e.,

$$\Delta^k(w, \phi \otimes \psi) = \Delta^k(w, \psi \otimes \phi) \quad \text{and} \quad \Delta^k(w, \psi \otimes \psi) \geq 0 \quad \forall \phi, \psi \in F^k.$$

(ii) *Sparsity:*

$$\Delta^k(e_m \otimes e_n, v_i \otimes v_j) \neq 0 \quad \text{only if} \quad m = n \quad \text{and} \quad i, j \in \{n - 1, n\}.$$

(iii) *Bilinearity and continuity on $E_\pi \times F_\epsilon$.*

The corresponding approximation of the operator \mathcal{B} is defined as $\mathcal{B}^k := B - \rho^2\Delta^k$. We are now interested in the solution of the discrete variational problem

$$(73) \quad \text{Find } U^k \in E_\pi^k \quad \text{s.t.} \quad \mathcal{B}^k(U^k, v) = \ell(v) \quad \forall v \in F_\epsilon^k$$

which approximates (41).

3.6.2. Well-posedness of the discrete problem. The solution U^k to (73) can be expanded in terms of the basis $\{e_m \otimes e_n\}_{m,n}$ of E_π^k as

$$(74) \quad U^k = \sum_{m,n} U_{mn}^k (e_m \otimes e_n) \quad \text{with} \quad U_{mn}^k = \frac{(U^k, e_m \otimes e_n)_2}{\|e_m\|_E^2 \|e_n\|_E^2}.$$

We combine its coefficients in the $N \times N$ matrix $\mathbf{U} := (U_{mn}^k)_{m,n}$. Furthermore, we define the values

$$b_{in} := b(e_n, v_i) \quad \text{and} \quad \ell_{ij} := \ell(v_i \otimes v_j).$$

If the discrete inf-sup condition (51) is satisfied then $b_{n-1,n} \neq 0$ follows.

The sparsity assumption on Δ^k together with the fact that the discretization pair $E_\pi^k \times F_\epsilon^k$ is a tensor product discretization allow for an explicit formula for the diagonal entries of \mathbf{U} . This is presented in the lemma below.

For future purpose, we note that $w \in E^k \otimes E^k$ is symmetric positive semi-definite if and only if the matrix of coefficients $\mathbf{w} := (w_{mn})_{m,n}$ with respect to $\{e_m \otimes e_n\}_{m,n}$ is. Indeed, if $\varphi \in L_2(J)$ and $\varphi = ((e_n, \varphi)_{L_2(J)})_n \in \mathbb{R}^N$ then $\varphi^\top \mathbf{w} \varphi = \sum_{m,n} w_{mn} (e_m, \varphi)_{L_2(J)} (e_n, \varphi)_{L_2(J)} = (w, \varphi \otimes \varphi)_{L_2(J \times J)}$.

According to the sparsity assumption (ii), the nonzero values of Δ^k (as acting on the basis functions) can be combined in the 2×2 matrices

$$\Delta^n := \begin{pmatrix} \Delta^k(e_n \otimes e_n, v_{n-1} \otimes v_{n-1}) & \Delta^k(e_n \otimes e_n, v_{n-1} \otimes v_n) \\ \Delta^k(e_n \otimes e_n, v_n \otimes v_{n-1}) & \Delta^k(e_n \otimes e_n, v_n \otimes v_n) \end{pmatrix}, \quad 1 \leq n \leq N-1,$$

and in $\Delta^N := \Delta^k(e_N \otimes e_N, v_{N-1}, v_{N-1})$. The foregoing remark and Assumption (i) on Δ^k imply that each Δ^n is symmetric positive semi-definite.

We define

$$(75) \quad \beta_n := (1 - \rho^2 b_{n-1,n}^{-2} \Delta_{11}^n)^{-1}, \quad n = 1, \dots, N,$$

where Δ_{pq}^n denotes the (p, q) -th entry in the matrix Δ^n , and for $n \geq 2$:

$$(76) \quad \begin{aligned} \theta_n &:= b_{n-1,n}^{-2} b_{n-1,n-1}^2, \\ \Pi_n &:= -b_{n-2,n-1}^{-1} b_{n-1,n-1}, \\ \alpha_n &:= \beta_n \theta_n \left[1 + \rho^2 b_{n-1,n-1}^{-2} \left(\Delta_{22}^{n-1} + 2\Pi_n \Delta_{12}^{n-1} \right) \right]. \end{aligned}$$

We note that

$$(77) \quad \frac{\|e_n\|_E^2}{\|e_{n-1}\|_E^2} \alpha_n, \quad \frac{\|e_n\|_E^2}{\|e_{n-1}\|_E^2} \theta_n, \quad \Pi_n, \quad \text{and} \quad \beta_n$$

do not depend on the scaling of the basis $\{e_n\}_n$.

For technical reasons we also introduce the function $G^k \in E_\pi^k$ as the solution (which is well-defined under the inf-sup condition (51)) to

$$(78) \quad \text{Find } G^k \in E_\pi^k \text{ s.t. } B(G^k, v) = \ell(v) \quad \forall v \in E_\epsilon^k.$$

Let G_{mn}^k denote its coefficients.

Lemma 3.5. *Let $\ell \in F'_\epsilon$. Assume that β_n is finite for all n . Then there exists a unique solution $U^k \in E_\pi^k$ to the discrete variational problem (73). Its diagonal coefficients in (74) are*

$$(79) \quad U_{nn}^k = \beta_n G_{nn}^k + \sum_{m=1}^{n-1} G_{mm}^k (\beta_m \alpha_{m+1} - \beta_{m+1} \theta_{m+1}) \prod_{v=m+2}^n \alpha_v.$$

Proof. By locality of the support of e_n and v_i , the values $b_{in} = b(e_n, v_i)$ are non-zero at most for $i \in \{n-1, n\}$. Therefore, the coefficients $\{w_n\}_n$ of the solution $w \in E^k$ to the problem “ $b(w, v) = f(v)$ for all $v \in F^k$ ” are obtained by recursion,

$$b_{n-1,n} w_n = f(v_{n-1}) - b_{n-1,n-1} w_{n-1} = \sum_{j=0}^{n-1} \Pi_j^{n-1} f(v_j), \quad \text{where} \quad \Pi_j^n := \prod_{i=j+1}^n \frac{-b_{ii}}{b_{i-1,i}}.$$

Hence, the coefficients of the solution G^k to the tensorized problem (78) satisfy

$$(80) \quad b_{m-1,m} b_{n-1,n} G_{mn} = \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} \Pi_i^{m-1} \Pi_j^{n-1} \ell_{ij}.$$

Applying this formula to $BU = \ell + \rho^2 \Delta^k U$ instead of $BG = \ell$ gives

$$(81) \quad b_{m-1,m} b_{n-1,n} U_{mn} = b_{m-1,m} b_{n-1,n} G_{mn} + \rho^2 \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} \Pi_i^{m-1} \Pi_j^{n-1} [\Delta^k U^k]_{ij}.$$

Due to the sparsity assumption (ii) on Δ^k , the double sum contains only the diagonal U_{rr} coefficients with $r \leq \min\{m, n\}$ and no off-diagonal ones; specifically, only the entries

$$[\Delta^k U^k]_{r-1, r-1} = U_{r-1, r-1} \Delta_{22}^{r-1} + U_{rr} \Delta_{11}^r,$$

$$\begin{aligned} [\Delta^k U^k]_{r-2,r-1} &= U_{r-1,r-1} \Delta_{12}^{r-1}, \\ [\Delta^k U^k]_{r-1,r-2} &= U_{r-1,r-1} \Delta_{21}^{r-1}, \end{aligned}$$

occur. In particular, if $m = n$ then the formula gives a recursion for U_{nn} with $\rho^2 \Delta_{11}^n U_{nn}$ on the right-hand side. Therefore, we can solve for U_{nn} if $b_{n-1,n}^2 \neq \rho^2 \Delta_{11}^n$ (which is equivalent to β_n being finite). The formula then provides the remaining off-diagonal coefficients U_{mn} . With this, the existence of the discrete solution is established.

To obtain the representation (79), we subtract from formula (81) for U_{nn} that for $U_{n-1,n-1}$. After some manipulation, this leads to the iteration

$$U_{11} = \beta_1 G_{11}, \quad U_{nn} = \beta_n G_{nn} - \beta_n \theta_n G_{n-1,n-1} + \alpha_n U_{n-1,n-1}, \quad 2 \leq n \leq N,$$

and hence the claim (79). \square

Equation (79) is the discrete version of the identity in (44), which was used to prove (see Theorem 2.11) that a positive semi-definite right-hand side ℓ entails the same property for the solution U . The following Lemma characterizes the conditions on the discretization parameters for which this is true in the discrete.

Lemma 3.6. *The following are equivalent:*

- (i) $\beta_n > 0$ in (75) for all n ;
- (ii) For every symmetric positive semi-definite $\ell \in F'_\epsilon$ the discrete variational problem (73) has a unique symmetric positive semi-definite solution $U^k \in E_\pi^k$.

Proof. Assume (i). Let $\ell \in F'_\epsilon$ be SPSPD. Then $G^k \in E_\pi^k$ defined in (78) is also SPSPD by Lemma 3.4. As remarked above, its matrix of coefficients is therefore also SPSPD, in particular $G_{nn}^k \geq 0$. From this and (79), it follows that also $U_{nn}^k \geq 0$. Indeed, with (i) $\beta_n > 0$, we obtain the equivalence

$$(83) \quad \alpha_{n+1} \geq \beta_n^{-1} \beta_{n+1} \theta_{n+1} \iff (b_{n-1,n}^{-1}, b_{nn}^{-1}) \Delta^n (b_{n-1,n}^{-1}, b_{nn}^{-1})^\top \geq 0.$$

Since the matrices Δ^n are positive semi-definite, $\beta_n \alpha_{n+1} \geq \beta_{n+1} \theta_{n+1}$ holds. In addition, $\alpha_n \geq 0$ for all $n \geq 2$, because $\theta_n \geq 0$ by definition (76). Hence $U_{nn}^k \geq 0$. Set now $\widehat{U}^k := \sum_{n=1}^N U_{nn}^k (e_n \otimes e_n)$. Since the discrete inf-sup condition (51) is assumed, there exists a unique $U^k \in E_\pi^k$ satisfying $B(U^k, v) = \widehat{\ell}(v)$ for all $v \in F_\epsilon^k$, where $\widehat{\ell} := \rho^2 \Delta^k \widehat{U}^k + \ell$. By Assumption (i) on Δ^k , the functional $\widehat{\ell}$ is SPSPD on $F^k \otimes F^k$. The function U^k inherits the symmetry, and by Lemma 3.4, it is also SPSPD. Moreover, the identity (80) applied to the right-hand side $\widehat{\ell}$ yields $b_{n-1,n}^2 U_{nn}^k = \sum_{i,j < n} \Pi_i^{n-1} \Pi_j^{n-1} [\rho^2 \Delta^k \widehat{U}^k + \ell]_{ij} = b_{n-1,n}^2 \widehat{U}_{nn}$, where the last equality follows from the definition of the coefficients $\widehat{U}_{nn} = U_{nn}$ and the sparsity properties (82). Consequently, $\Delta^k \widehat{U}^k = \Delta^k U^k$ on F_ϵ^k , and U^k is the desired solution.

Conversely, assume (ii). For any $g_1, \dots, g_N \geq 0$, the function $G^k := \sum_n g_n (e_n \otimes e_n) \in E_\pi^k \subset E_\pi$ is SPSPD. By Lemma 2.2, the functional $\ell := B G^k \in F'_\epsilon$ inherits this property and, moreover, by assumption also the solution U^k to (73) is positive semi-definite. In particular, $U_{nn}^k \geq 0$. Fix $n \in \{1, \dots, N\}$ and choose $g_n = 1$ and $g_m = 0$ for all $m \neq n$. With this choice, the nonnegativity of U_{nn}^k along with its representation in (79) imply that $\beta_n \geq 0$. Since β_n is a fraction (75), we conclude that (i) β_1, \dots, β_N are positive. \square

3.6.3. Discrete stability and inf-sup. The representation of U_{nn}^k in (79) in combination with the Lemmas 2.6 and 3.6 allow for an explicit representation of the E_π -norm of the discrete solution:

Corollary 3.7. *Suppose $\beta_n > 0$ in (75) for all n . Let $\ell \in F'_\epsilon$ be symmetric positive semi-definite. Then the discrete variational problem (73) admits a unique solution $U^k \in E_\pi^k$. It is symmetric positive semi-definite with norm*

$$(84) \quad \|U^k\|_\pi = \sum_{n=1}^N \left(\beta_n G_{nn}^k + \sum_{m=1}^{n-1} G_{mm}^k (\beta_m \alpha_{m+1} - \beta_{m+1} \theta_{m+1}) \prod_{v=m+2}^n \alpha_v \right) \|e_n\|_E^2.$$

Proof. Lemmas 2.6, 3.5 and 3.6 give $\|U^k\|_\pi = \lambda\delta(U^k) = \sum_{n=1}^N U_{nn}^k \|e_n\|_E^2$. Inserting the expression (79) for U_{nn}^k yields (84). \square

From Corollary 3.7, the norm of the discrete solution U^k can be estimated in terms of the norm of the right-hand side ℓ . We shall do this under the additional assumption of a uniform temporal mesh. For convenience of notation, we rescale the basis $\{e_n\}_n$ to $\|e_n\|_E = 1$, so that in view of (77), the numbers $(\alpha, \beta, \Pi, \theta) := (\alpha_n, \beta_n, \Pi_n, \theta_n)$ do not depend on n . Furthermore, $\theta = \Pi^2$.

Theorem 3.8. *In addition to the conditions posed in Corollary 3.7, assume that the temporal mesh is uniform. Then the discrete solution U^k to (73) satisfies the stability bound*

$$(85) \quad \|U^k\|_\pi \leq C_k \|\ell\|_{-\epsilon} \quad \text{with} \quad C_k := \gamma_k^{-2} \beta \left(1 + (\alpha - \Pi^2) \frac{\alpha^{N-1} - 1}{\alpha - 1} \right),$$

where γ_k is the discrete inf-sup constant from (51). If $\alpha = 1$ then $C_k = \gamma_k^{-2} \beta (\Pi^2 + N(1 - \Pi^2))$.

Proof. Corollary 3.7 yields

$$(86) \quad \|U^k\|_\pi = \beta \sum_{n=1}^N G_{nn}^k + \beta(\alpha - \theta) \sum_{m=1}^{N-1} G_{mm}^k \sum_{n=0}^{N-m-1} \alpha^n,$$

where we have changed the order of summation. If $\alpha \neq 1$ then it follows from the observations in (83) that either $\theta \leq \alpha < 1$ or $\theta \leq 1 < \alpha$. In both cases, $\frac{1 - \alpha^{N-n}}{1 - \alpha} \leq \frac{1 - \alpha^{N-1}}{1 - \alpha}$. Hence, evaluating the geometric sum in (86) and using the identity $\theta = \Pi^2$ yield

$$\|U^k\|_\pi = \beta \sum_{n=1}^N (1 + (\alpha - \theta) \frac{1 - \alpha^{N-n}}{1 - \alpha}) G_{nn}^k \leq \beta (1 + (\alpha - \Pi^2) \frac{1 - \alpha^{N-1}}{1 - \alpha}) \|G^k\|_\pi \leq C_k \|\ell\|_{-\epsilon}.$$

For $\alpha = 1$, the second claim follows directly from (86). \square

As a consequence of the the stability bound in the previous theorem we obtain an inf-sup condition for $\mathcal{B}^k = B - \rho^2 \Delta^k$. It is convenient to formulate it on the subspaces $\widehat{E}_\pi^k \subset E_\pi^k$ and $\widehat{F}_\epsilon^k \subset F_\epsilon^k$ of symmetric functions.

Corollary 3.9. *Suppose the temporal mesh is uniform with $\beta > 0$. Then \mathcal{B}^k in (73) satisfies the discrete inf-sup condition (note the symmetrization)*

$$(87) \quad \inf_{w \in \mathcal{S}(\widehat{E}_\pi^k)} \sup_{v \in \mathcal{S}(\widehat{F}_\epsilon^k)} \mathcal{B}^k(w, v) \geq C_k^{-1},$$

where C_k is the discrete stability constant in (85).

Proof. Fix a symmetric $w \in \widehat{E}_\pi^k$. On \widehat{F}_ϵ^k define the functional $\ell := \mathcal{B}^k w$, extending it via Hahn-Banach with equal norm to F_ϵ^k . Decompose it as $\ell =: \ell^+ - \ell^- + \ell^a$ as in (36). Then $\ell^a = 0$ by symmetry of w . Let $w^\pm \in \widehat{E}_\pi^k$ be the solution to (73) with the right-hand side ℓ^\pm . Clearly, $w = w^+ - w^-$. Therefore,

$$\|w\|_\pi \leq \|w^+\|_\pi + \|w^-\|_\pi \stackrel{(85)}{\leq} C_k (\|\ell^+\|_{-\epsilon} + \|\ell^-\|_{-\epsilon}) \stackrel{(36)}{=} C_k \|\ell\|_{-\epsilon}.$$

Since $w \in \widehat{E}_\pi^k$ was arbitrary and $\|\ell\|_{-\epsilon} = \sup_{v \in \mathcal{S}(\widehat{F}_\epsilon^k)} \mathcal{B}^k(w, v)$, the conclusion (87) follows. \square

Now we introduce some approximations Δ^k of the trace product Δ . This is of interest primarily for the iE^* discretization. The schemes we consider are

CN_2^* : the CN^* discretization discussed in §3.2.1 with the exact trace product $\Delta^k := \Delta$.

iE_2^* : The iE^* discretization introduced in §3.2.2 with the exact trace product $\Delta^k := \Delta$.

iE_2^*/Q : iE^* with preprocessing: $\Delta^k := \Delta \circ (q_k \otimes q_k)$ with q_k from (64).

iE_2^*/\square : iE^* with the “box rule”

$$(88) \quad \Delta^k(w, v) := \sum_{n=1}^N k_n^{-1} \int_{J_n \times J_n} w(s, t) v(s, t) ds dt, \quad (w, v) \in E_\pi^k \times F_\epsilon^k.$$

This definition is motivated by observing that $\Delta(w, v)$ is the double integral of $\delta(s - t)w(s, t)v(s, t)$ over all “boxes” $J_n \times J_n$ and approximating $\delta(s - t)$ by k_n^{-1} on $J_n \times J_n$.

All these candidates for the approximate trace product Δ^k satisfy the assumptions (i)–(iii) made above. In particular, they are bilinear and continuous. Continuity is quantified in the following lemma.

Lemma 3.10. *Each of the above Δ^k is bounded on $E_\pi \times F_\epsilon$ with*

$$\Delta^k(w, v) \leq \frac{1}{2\lambda} \|w\|_\pi \|v\|_\epsilon \quad \forall (w, v) \in E_\pi \times F_\epsilon.$$

Proof. Boundedness of the exact trace product is the subject of Lemma 2.8. For the approximation with preprocessing $\Delta^k := \Delta \circ (q_k \otimes q_k)$ we have the same bound, because $\|q_k : E \rightarrow E^k\| = 1$ and therefore $\|(q_k \otimes q_k) : E_\pi \rightarrow E_\pi^k\| = 1$.

Now consider the “box rule” Δ^k as in (88). Let $(w, v) \in E_\pi \times F_\epsilon$. By [17, Theorem 2.4] we may assume that $w = w^1 \otimes w^2$. Employing $|v(s, t)| \leq \frac{1}{2} \|v\|_\epsilon$ from (32) in (88) results in the estimate $\Delta^k(w, v) \leq \frac{1}{2} \|v\|_\epsilon \sum_n \|w_1\|_{L_2(J_n)} \|w_2\|_{L_2(J_n)} \leq \frac{1}{2\lambda} \|v\|_\epsilon \|w_1\|_E \|w_2\|_E$. \square

The values of Δ^n , α , β and Π for each scheme are given in Table 1 below in terms of the time-step size $k > 0$ (assumed uniform) and the dimensionless numbers $z := \lambda k$ and $q := \rho^2 / (2\lambda)$. Recall that the basis $\{e_n\}_n \subset E^k$ is normalized to $\|e_n\|_E = 1$ to define those values. The denominator of $\beta_n = \lambda k_n b_{n-1, n}^2 / D_n$ is $D_n = \lambda k_n (b_{n-1, n}^2 - \rho^2 \Delta_{11}^n)$. Thus $D_n > 0$ necessary and sufficient for $\beta_n > 0$ in Lemma 3.6. On a uniform mesh we write $D := D_n$. We remark that $D > 0$ holds for all our schemes if the temporal mesh width k is sufficiently small, namely when $k\rho^2 \lesssim 1$.

| Scheme | $\lambda \Delta^n$ | D | $\alpha - 1$ | β | Π |
|------------------|--------------------------------------------------------------|-----------------------------------------|--------------------------------------|-----------------------|-----------------------|
| CN_2^* | $\frac{1}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ | $(1 + z/2)^2 - \frac{2}{3} qz$ | $\frac{(2/3)(2+\Pi)qz-2z}{D}$ | $\frac{(1+z/2)^2}{D}$ | $\frac{1-z/2}{1+z/2}$ |
| iE_2^* | $\frac{1}{60} \begin{pmatrix} 38 & 7 \\ 7 & 8 \end{pmatrix}$ | $\frac{1}{4}(1+z)^2 - \frac{19}{15} qz$ | $\frac{(4/15)(23+7\Pi)qz-z(2+z)}{D}$ | | |
| iE_2^*/Q | $\frac{1}{24} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ | $\frac{1}{4}(1+z)^2 - \frac{1}{6} qz$ | $\frac{(2/3)(2+\Pi)qz-z(2+z)}{4D}$ | $\frac{(1+z)^2}{4D}$ | $\frac{1}{1+z}$ |
| iE_2^*/\square | $\frac{1}{4} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ | $\frac{1}{4}(1+z)^2 - \frac{1}{2} qz$ | $\frac{1-4D}{4D}$ | | |

TABLE 1. Discretization parameters for the schemes from Section 3.6.

With Theorem 3.8 we find that $\lim_{k \rightarrow 0} C_k = C$ for the schemes CN_2^* , iE_2^*/Q , and iE_2^*/\square (but not iE_2^*), where C is the stability constant in (47) of the continuous problem (41).

3.6.4. Error analysis and convergence. In this subsection we estimate the difference between the exact solution U to (41) and the discrete solution U^k to (73). We first remark that by Lemma 3.10, the norm of $\mathcal{B}^k = B - \rho^2 \Delta^k$ is bounded by

$$\|\mathcal{B}^k\| \leq 1 + \frac{\rho^2}{2\lambda}$$

for each $\Delta^k \in \{\Delta, \Delta \circ (q_k \otimes q_k), (88)\}$. Moreover, \mathcal{B}^k satisfies the inf-sup condition (87) on $\widehat{E}_\pi^k \times \widehat{F}_\epsilon^k \subset E_\pi \times F_\epsilon$, and the dimensions of these subspaces coincide. Hence, Proposition 3.3 on quasi-optimality of the discrete solution applies. This quasi-optimality is formulated in terms of the symmetric subspace \widehat{E}_π^k , but we can improve this to E_π^k for symmetric solutions U . Indeed, if $U \in E_\pi$ is symmetric then $\|U - \frac{1}{2}(w + w^*)\|_\pi \leq \frac{1}{2}(\|U - w\|_\pi + \|(U - w)^*\|_\pi) = \|U - w\|_\pi$ for any $w \in E_\pi$, where $(\cdot)^*(s, t) := (\cdot)(t, s)$. Furthermore, the appearing residual $(\mathcal{B} - \mathcal{B}^k)U = (\Delta - \Delta^k)U$ is a

symmetric functional, whether U is symmetric or not, and therefore vanishes on anti-symmetric elements of F_ϵ . This leads to the estimate

$$\|U - U^k\|_\pi \leq (1 + C_k \|\mathcal{B}^k\|) \inf_{w \in E_\pi^k} \|U - w\|_\pi + C_k \|(\Delta - \Delta^k)U\|_{(F_\epsilon^k)'}.$$

for symmetric ℓ . Replacing C_k by $(\gamma_k^{-2} + C_k)$, the assumption of symmetry may be dropped.

This result shows convergence for the CN_2^* scheme, where $\Delta^k = \Delta$. Unfortunately, it is not useful for the iE_2^* scheme and its variants, because the best approximation from the discrete space E_π^k does not converge to U as we refine the temporal mesh, see the discussion at the end of §3.2.2. This motivates looking at the postprocessed solution

$$(89) \quad \bar{U}^k := Q_k U^k \quad \text{with} \quad Q_k := (q_k \otimes q_k)$$

for those schemes, where q_k is the projection from (64). Recall that q_k is injective on E^k . By Q_k^{-1} we will mean the inverse of $Q_k: E_\pi^k \rightarrow Q_k E_\pi^k$. In the case of the iE_2^* discretization, (65) implies

$$(90) \quad \|Q_k w\|_\pi = \frac{1}{4} \|w\|_\pi \quad \forall w \in E_\pi^k.$$

The convergence of the postprocessed solution will again be obtained via Proposition 3.3. To this end, we define $\bar{\mathcal{B}}^k := \mathcal{B}^k \circ Q_k^{-1} Q_k: E_\pi \rightarrow F'_\epsilon$ with the motivation that the postprocessed solution solves the modified discrete problem

$$(91) \quad \text{Find } \bar{U}^k \in Q_k E_\pi^k \quad \text{s.t.} \quad \bar{\mathcal{B}}^k(\bar{U}^k, v) = \ell(v) \quad \forall v \in F_\epsilon^k.$$

The operator $\bar{\mathcal{B}}^k$ is bounded with $\|\bar{\mathcal{B}}^k\| \leq 4\|\mathcal{B}^k\|$. Moreover, it follows from (90) that if \mathcal{B}^k satisfies the discrete inf-sup condition (87) on $\widehat{E}_\pi^k \times \widehat{F}_\epsilon^k$ with the constant C_k^{-1} then so does $\bar{\mathcal{B}}^k$ on $Q_k \widehat{E}_\pi^k \times \widehat{F}_\epsilon^k$ with the constant $4C_k^{-1}$. The following is our main result.

Proposition 3.11. *Let $\ell \in F'_\epsilon$ be symmetric. Assume the discrete inf-sup condition (87). Then the exact solution $U \in E_\pi$ to (41) and the postprocessed discrete solution $\bar{U}^k \in Q_k E_\pi^k$ to (73)/(91) differ by*

$$\|U - \bar{U}^k\|_\pi \leq (1 + C_k \|\mathcal{B}^k\|) \inf_{w \in Q_k E_\pi^k} \|U - w\|_\pi,$$

for the CN_2^* scheme, and by

$$(92) \quad \|U - \bar{U}^k\|_\pi \leq (1 + C_k \|\mathcal{B}^k\|) \inf_{w \in Q_k E_\pi^k} \|U - w\|_\pi + \frac{1}{4} C_k \|(\mathcal{B} - \bar{\mathcal{B}}^k)U\|_{(F_\epsilon^k)'}$$

for any of the iE_2^* schemes.

To complete the analysis we need to estimate the residual term in (92). Hence, from now on we focus entirely on the iE_2^* schemes. Recalling that $\mathcal{B} = B - \rho^2 \Delta$ and $\bar{\mathcal{B}}^k = (B - \rho^2 \Delta^k) Q_k^{-1} Q_k$ we split the residual according to

$$(93) \quad \mathcal{B} - \bar{\mathcal{B}}^k = \mathcal{B}(\text{Id} - Q_k) - B(\text{Id} - Q_k) Q_k^{-1} Q_k - \rho^2 (\Delta Q_k - \Delta^k) Q_k^{-1} Q_k$$

and address it term by term.

- The first term $T_1 := \|\mathcal{B}(\text{Id} - Q_k)U\|_{(F_\epsilon^k)'}$ in (92)/(93) goes to zero upon mesh refinement by density of the subspaces $Q_k E_\pi^k \subset E_\pi$.
- To bound the second term $T_2 := \|B(\text{Id} - Q_k) Q_k^{-1} Q_k U\|_{(F_\epsilon^k)'}$ in (92)/(93) we proceed in two steps. First, we observe that $b((\text{Id} - q_k)w, v) = ((\text{Id} - q_k)w, v)_E = ((\text{Id} - q_k)w, (\text{Id} - q_k)v)_E \leq \|w\|_E \|(\text{Id} - q_k)v\|_E$ for any $(w, v) \in E \times F$. The Poincaré–Wirtinger inequality on each temporal element yields $\|(\text{Id} - q_k)v\|_E \leq \frac{1}{\sqrt{12}} \lambda \max_n k_n \|v\|_F$ for all $v \in F^k$. Second, we write

$$(94) \quad \text{Id} - Q_k = \frac{1}{2} [(\text{Id} - q_k) \otimes (\text{Id} + q_k) + (\text{Id} + q_k) \otimes (\text{Id} - q_k)],$$

and use this identity in $B(\text{Id} - Q_k)$. Recalling $\|Q_k^{-1} Q_k U\|_\pi = 4\|Q_k U\|_\pi \leq 4\|U\|_\pi$ from (90), this gives $T_2 \leq \frac{4}{\sqrt{3}} \lambda \max_n k_n \|U\|_\pi$.

- Consider now the third term $T_3 := \rho^2 \|(\Delta Q_k - \Delta^k) Q_k^{-1} Q_k U\|_{(\widehat{F}_\epsilon^k)}$ in (92)/(93). For the iE_2^* scheme where $\Delta^k = \Delta$, this term does not converge to zero upon mesh refinement, see §3.6.5. For the iE_2^*/Q scheme where $\Delta^k = \Delta Q_k$, this term vanishes identically.

It remains to discuss the “box rule” where $\Delta^k = (88)$. To this end, we first note that for $v \in F_\epsilon^k$

$$(95) \quad (\Delta Q_k - \Delta^k)(Q_k^{-1} Q_k U, v) = \Delta(Q_k U, v) - \Delta(Q_k U, I v) = \Delta(Q_k U, (\text{Id} - \widetilde{Q}_k)v),$$

with $\widetilde{Q}_k := I \otimes I$ and the interpolation operator I on the space of piecewise constants from (60). To estimate the expression on the right-hand side, we recall from [16, §3.2] that $C^0(\bar{J} \times \bar{J}) = C^0(\bar{J}) \otimes_\epsilon C^0(\bar{J})$. We decompose the operator $\text{Id} - \widetilde{Q}_k$ in the same way as we did for $\text{Id} - Q_k$ in (94). The estimates $\|\psi - I\psi\|_{C^0(\bar{J})} \leq (\lambda \max_n k_n)^{1/2} \|\psi\|_F$ and $\|\psi + I\psi\|_{C^0(\bar{J})} \leq \sqrt{2} \|\psi\|_F$ for $\psi \in F^k$, then imply convergence for $U \in E_\pi$ of order $\mathcal{O}(\sqrt{\max_n k_n/\lambda})$, since

$$(96) \quad |\Delta(Q_k U, (\text{Id} - \widetilde{Q}_k)v)| \leq \delta(|Q_k U|) \|v - \widetilde{Q}_k v\|_{C^0(\bar{J} \times \bar{J})} \leq \frac{\sqrt{2 \max_n k_n}}{\sqrt{\lambda}} \|U\|_\pi \|v\|_\epsilon.$$

However, we observe numerically that $T_3 \lesssim \max_n k_n$. This is because the solutions to our model problems (5c) are continuous on $\bar{J} \times \bar{J}$. Under the assumption that $U \in C^0(\bar{J} \times \bar{J})$ we split term in (95) as follows

$$(97) \quad |\Delta(Q_k U, (\text{Id} - \widetilde{Q}_k)v)| \leq |\Delta(Q_k U, (\text{Id} - I) \otimes (\text{Id} - I)v)| + 2|\Delta(Q_k U, I \otimes (\text{Id} - I)v)|.$$

The properties of the operator I mentioned above show that the first term can be bounded by $|\Delta(Q_k U, (\text{Id} - I) \otimes (\text{Id} - I)v)| \leq \max_n k_n \|U\|_\pi \|v\|_\epsilon$. In order to investigate the second term in (97), we assume first that $v = \psi \otimes \psi$ for $\psi \in F^k$. With the notation $\psi_{n-1} := \psi(t_{n-1})$ and $\bar{U}_n := Q_k U|_{J_n \times J_n}$ we obtain

$$\Delta(Q_k U, I \otimes (\text{Id} - I)v) = \sum_n \frac{\psi_n - \psi_{n-1}}{2} \psi_{n-1} k_n \bar{U}_n = -\frac{1}{4} \psi_0^2 k_1 \bar{U}_1 - \frac{1}{4} \sum_n (\psi_n - \psi_{n-1})^2 k_n \bar{U}_n.$$

For a symmetric function $v \in \widehat{F}_\epsilon^k$ we can find a representation $v = \sum_m v^m \otimes v^m$ with $v^m \in F^k$ and

$$\begin{aligned} |\Delta(Q_k U, I \otimes (\text{Id} - I)v)| &\leq \frac{1}{4} k_1 \bar{U}_1 |(\delta_0 \otimes \delta_0)(v)| + \frac{1}{4} \sum_n k_n \bar{U}_n \sum_m (v_n^m - v_{n-1}^m)^2 \\ &\leq \frac{1}{8} \max_n k_n \|U\|_{C^0(\bar{J} \times \bar{J})} \|v\|_\epsilon + \frac{1}{4} \delta(|Q_k U|) \max_n \|\delta_{t_n} - \delta_{t_{n-1}}\|_{(F^k)}^2 \|v\|_\epsilon. \end{aligned}$$

Since $\|\delta_{t_n} - \delta_{t_{n-1}}\|_{(F^k)}^2 \leq \lambda k_n$ for all n , we obtain the bound $\frac{1}{8} \max_n k_n (\|U\|_{C^0(\bar{J} \times \bar{J})} + 2\|U\|_\pi) \|v\|_\epsilon$ for the second term in (97) and $T_3 \leq \frac{\rho^2}{4} \max_n k_n (\|U\|_{C^0(\bar{J} \times \bar{J})} + 6\|U\|_\pi)$ for any symmetric $U \in E_\pi \cap C^0(\bar{J} \times \bar{J})$.

3.6.5. Non-convergence of iE_2^* with postprocessing. We introduced the approximate trace product (72) because even with postprocessing, the iE_2^* scheme with the exact trace product does not converge upon temporal mesh refinement. In fact, it is consistent with the value 2ρ for the volatility instead of ρ , as we will indicate here. First, as in (65), we have $\Delta(w, Q_k v) = \Delta(4Q_k w, Q_k v)$ for all $(w, v) \in E_\pi^k \times F_\epsilon^k$. Therefore, invoking $\delta(|w - 4Q_k w|) \leq \frac{1}{\lambda} \|w - 4Q_k w\|_\pi$ and Equation (90),

$$(98) \quad |\Delta(w, v) - 4\Delta(Q_k w, v)| = |\Delta(w - 4Q_k w, v - Q_k v)| \leq \frac{2}{\lambda} \|w\|_\pi \|v - Q_k v\|_{C^0(\bar{J} \times \bar{J})}.$$

To estimate the last term, we proceed as in (96) and use the estimates

$$\|\psi - q_k \psi\|_{C^0(\bar{J})} \leq \frac{1}{2} (\lambda \max_n k_n)^{1/2} \|\psi\|_F, \quad \|\psi + q_k \psi\|_{C^0(\bar{J})} \leq \sqrt{2} \|\psi\|_F$$

for $\psi \in F^k$. We obtain $\|v - Q_k v\|_{C^0(\bar{J} \times \bar{J})} \lesssim (\lambda \max_n k_n)^{1/2} \|v\|_\epsilon$. By the preceding subsection, the iE_2^*/Q scheme with ΔQ_k does provide a consistent approximation, so (98) shows that iE_2^* does not.

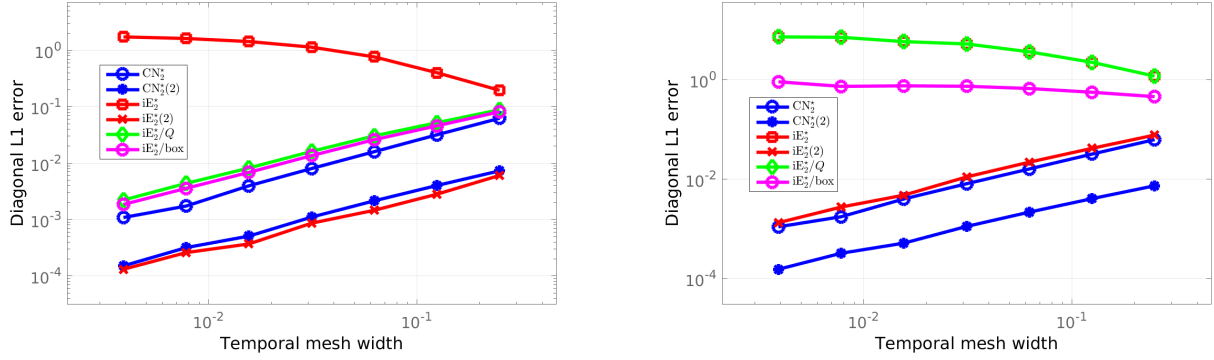


FIGURE 2. The error $\delta(|U - w|)$ as a function of the temporal mesh width for the example from Section 3.7. Left: with postprocessing, $w = \bar{U}^k$. Right: without postprocessing, $w = U^k$.

3.7. Numerical example. In the following numerical experiment we implement the schemes CN_2^* , iE_2^* , iE_2^*/Q , and iE_2^*/\square proposed in Section 3.6 to solve the discrete variational problem (73). In addition, we apply the discretizations of polynomial degree $p = 2$ from §3.2.3 with the exact trace product Δ , denoted by $\text{CN}_2^*(2)$ and $\text{iE}_2^*(2)$. We choose $T = 2$, $\lambda = 3$, $\rho^2 = \lambda/2$, and for the right-hand side $\ell(v) := v(0)$, motivated by (38). The error against the exact solution from (5c) is measured as the L_1 error on the diagonal, $E(w) := \delta(|U - w|)$ for $w = U^k$ (without postprocessing) and $w = \bar{U}^k$ (with postprocessing). The results are shown in Figure 2. Convergence of the schemes is summarized in the following table. The convergence, where present, is of first order in the temporal mesh width.

| | CN_2^* | $\text{CN}_2^*(2)$ | iE_2^* | $\text{iE}_2^*(2)$ | iE_2^*/Q | iE_2^*/\square |
|------------------|-----------------|--------------------|-----------------|--------------------|-------------------|-------------------------|
| (99) \bar{U}^k | ✓ | ✓ | × | ✓ | ✓ | ✓ |
| U^k | ✓ | ✓ | × | ✓ | × | × |

These results are in line with the convergence results established in Section 3.6. The schemes of polynomial degree $p = 2$ exhibit only first order converge, presumably due to the limited smoothness of the second moment across the diagonal. However, they do not require pre- or postprocessing for convergence. The stability of the $\text{iE}_2^*(2)$ scheme, in particular, does not depend on the temporal mesh width as long as it is equidistant, see (63), but for now, this statement is limited to the range (71).

4. CONCLUSIONS

We have considered model stochastic ODEs with additive and multiplicative Brownian noise (1)/(2), and have derived the deterministic equations in variational form satisfied by the first (16) and second moment (22)/(38) of the solution. The equations for the second moment are posed on tensor products of function spaces, which can be taken as Hilbert tensor products (17) in the additive case, whereas projective–injective tensor product spaces (26) as trial–test spaces are required in the multiplicative case. The well-posedness of those equations is evident in the additive case (22) by the isometry property of the operator (19), but the multiplicative case, analyzed in Theorem 2.11, requires more work due to the presence of the trace product (25) in the operator.

We have discussed Petrov–Galerkin discretizations of two basic kinds for the first moment: CN^* in §3.2.1 and iE^* in §3.2.2. The main difference is in the stability behavior documented in Figure 1, wherein CN^* requires the CFL number to be small, as opposed to iE^* which can be made stable (3.1) under mild restrictions on the temporal mesh. Higher order generalizations followed

in §3.2.3. From these, tensor product Petrov–Galerkin discretizations are constructed in Section 3.4. We have discussed the additive case briefly in Section 3.5 in order to focus the multiplicative case in Section 3.6.

Trying to harness the favorable stability properties of the iE^* discretization, two problems arise in the multiplicative case: lack of density of the trial spaces (see §3.2.2) and inconsistent interaction of the basis functions with the trace product (see §3.6.5). The first issue is addressed by postprocessing (89) and the second by a modification of the trace product (we have suggested the two variants iE_2^*/Q and iE_2^*/\square). Unfortunately, postprocessing, as analyzed in the framework of variational crimes in (92), again entails a CFL restriction. Postprocessing is not required for the higher order discretizations (see Figure 2 and Table (99)), but their stability beyond the trivial range (71) remains to be verified.

These insights should prove useful for developing numerical methods for stochastic partial differential evolution equations.

5. ACKNOWLEDGEMENT

The authors thank S. Larsson for valuable input. RA was supported by French ANR-12-MONU-0013 and Swiss NSF #164616. RA is grateful to the guest researcher program of the Department of Mathematical Sciences, Chalmers University of Technology, for enabling a productive four-week visit.

(R. Andreev) UNIV PARIS DIDEROT, SORBONNE PARIS CITÉ, LJLL (UMR 7598 CNRS), F-75205 PARIS, FRANCE
E-mail address: roman.andreev@upmc.fr

(K. Kirchner) DEPARTMENT OF MATHEMATICAL SCIENCES, CHALMERS UNIVERSITY OF TECHNOLOGY AND UNIVERSITY OF GOTHENBURG, SE-412 96 GOTHENBURG, SWEDEN
E-mail address: kristin.kirchner@chalmers.se

REFERENCES

- [1] R. Andreev. Quasi-optimality of approximate solutions in normed vector spaces. Technical Report hal-01338040, HAL, 2016. 16
- [2] R. Andreev and J. Schweitzer. Conditional space-time stability of collocation Runge–Kutta for parabolic evolution equations. *Electron. Trans. Numer. Anal.*, 41:62–80, 2014. 10, 12, 13, 14, 15
- [3] I. Babuška. Error-bounds for finite element method. *Numer. Math.*, 16(4):322–333, 1971. 4
- [4] I. Babuška and T. Janik. The h-p version of the finite element method for parabolic equations. I. The p version in time. *Numer. Meth. Part. D. E.*, 5:363–399, 1989. 10
- [5] I. Babuška and T. Janik. The h-p version of the finite element method for parabolic equations. II. The h-p version in time. *Numer. Meth. Part. D. E.*, 6:343–369, 1990. 10, 11, 12
- [6] A. Barth, A. Lang, and C. Schwab. Multilevel Monte Carlo method for parabolic stochastic partial differential equations. *BIT*, 53(1):3–27, 2013. 1
- [7] M. B. Giles. Multilevel Monte Carlo path simulation. *Oper. Res.*, 56(3):607–617, 2008. 1
- [8] M. B. Giles and B. J. Waterhouse. Multilevel quasi-Monte Carlo path simulation. In *Advanced financial modelling*, volume 8 of *Radon Ser. Comput. Appl. Math.*, pages 165–181. Walter de Gruyter, Berlin, 2009. 1
- [9] K. Kirchner, A. Lang, and S. Larsson. Covariance structure of parabolic stochastic partial differential equations with multiplicative Lévy noise. arXiv:1506.00624. 1, 8, 9
- [10] F. Y. Kuo, C. Schwab, and I. H. Sloan. Quasi-Monte Carlo finite element methods for a class of elliptic partial differential equations with random coefficients. *SIAM J. Numer. Anal.*, 50(6):3351–3374, 2012. 1
- [11] F. Y. Kuo, C. Schwab, and I. H. Sloan. Multi-level quasi-Monte Carlo finite element methods for a class of elliptic PDEs with random coefficients. *Found. Comput. Math.*, 15(2):411–449, 2015. 1
- [12] A. Lang, S. Larsson, and C. Schwab. Covariance structure of parabolic stochastic partial differential equations. *Stoch. Partial Differ. Equ. Anal. Comput.*, 1(2):351–364, 2013. 1, 5
- [13] B. Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Universitext. Springer-Verlag, Berlin, 6th edition, 2013. 2, 3, 5
- [14] T. Palmer. *Banach Algebras and the General Theory of *-Algebras: Volume 2, *-Algebras*. Banach Algebras and the General Theory of *-algebras. Cambridge University Press, 2001. 7
- [15] M. Reed and B. Simon. *Methods of modern mathematical physics. I. Functional analysis*. Academic Press Inc., New York, 1980. 4
- [16] R. Ryan. *Introduction to Tensor Products of Banach Spaces*. Springer Monographs in Mathematics. Springer, 2002. 6, 8, 23
- [17] R. Schatten. *A Theory of Cross-Spaces*. Annals of Mathematics Studies, no. 26. Princeton University Press, Princeton, N. J., 1950. 7, 8, 21
- [18] A. Stern. Banach space projections and Petrov–Galerkin estimates. *Numer. Math.*, 130(1):125–133, 2015. 15, 16
- [19] R. A. Todor. *Sparse perturbation algorithms for elliptic PDE’s with stochastic data*. PhD thesis, ETH Zürich, 2005. ETH Diss. Nr. 16192. 5
- [20] J. Xu and L. Zikatanov. Some observations on Babuška and Brezzi theories. *Numer. Math.*, 94(1):195–202, 2003. 12, 16