Thèse n° 9087

Dynamical low rank approximation for uncertainty quantification of time-dependent problems

Présentée le 25 mai 2022

Faculté des sciences de base
Calcul scientifique et quantification de l'incertitude - Chaire CADMOS
Programme doctoral en mathématiques

pour l'obtention du grade de Docteur ès Sciences

par

Eva VIDLICKOVÁ

Acceptée sur proposition du jury

Prof. D. Kressner, président du jury
Prof. F. Nobile, directeur de thèse
Prof. A. Nouy, rapporteur
Prof. B. Vandereycken, rapporteur
Prof. N. Boumal, rapporteur

École
polytechnique
fédérale
de Lausanne

2022

Pre maminku a ocina.

Acknowledgements

My first thanks goes to Prof. Fabio Nobile for accepting me and supervising me throughout this satisfying, stressful, frustrating, exciting, hopeless, joyful but definitely enriching PhD path. I would as well like to thank my other collaborators Yoshihito Kazashi and Kody Law for their insights, effort and help.

My PhD journey was much more pleasant thanks my fellow PhD students Juan, Davide and Sundar, who accompanied me from the beginning till the end, and the rest of the CSQI group.

Mami, oci, Adelka, Blažejko a Miško, som veľmi vďačná za vašu bezpodmienečnú podporu a lásku. Robiť vás hrdými mi bolo veľkým zdrojom motivácie. Mať takú skvelú rodinu je veľkým životným šťastím a som tiež vďačná za prírastky Mifko, Barbi, Miška, Grétka, Rubinka a Dávidko.

My life in Lausanne became much more joyful after meeting the extraordinary Fabian, Giacomo, Riccardo, Giacomo, Edoardo, Gonzalo, Elisabetta, Luca, Veronica, Fero. Thank you for all the drinks, laughs, parties, hikes and bike rides.

Finally, I would like to thank Dominik for his love and support, Majka, the Prague group Viki, Lenka, Pali, Dano, Vladko and Him.

Lausanne, April 24, 2022 Evička

Abstract

The quantification of uncertainties can be particularly challenging for problems requiring long-time integration as the structure of the random solution might considerably change over time. In this respect, dynamical low-rank approximation (DLRA) is very appealing. It can be seen as a reduced basis method, thus solvable at a relatively low computational cost, in which the solution is expanded as a linear combination of a few deterministic functions with random coefficients. The distinctive feature of the DLRA is that both the deterministic functions and random coefficients are computed on the fly and are free to evolve in time, thus adjusting at each time to the current structure of the random solution. This is achieved by suitably projecting the dynamics onto the tangent space of a manifold consisting of all random functions with a fixed rank. In this thesis, we aim at further analyzing and applying the DLR methods to time-dependent problems.

Our first work considers the DLRA of random parabolic equations and proposes a class of fully discrete numerical schemes. Similarly to the continuous DLRA, our schemes are shown to satisfy a discrete variational formulation. By exploiting this property, we establish the stability of our schemes: we show that our explicit and semi-implicit versions are conditionally stable under a "parabolic" type CFL condition which does not depend on the smallest singular value of the DLR solution; whereas our implicit scheme is unconditionally stable. Moreover, we show that, in certain cases, the semi-implicit scheme can be unconditionally stable if the randomness in the system is sufficiently small. The analysis is supported by numerical results showing the sharpness of the obtained stability conditions.

The discrete variational formulation is further applied in our second work, which derives a-priori and a-posteriori error estimates for the discrete DLRA of a random parabolic equation obtained by the three newly-proposed schemes. Under the assumption that the right-hand side of the dynamical system lies in the tangent space up to a small remainder, we show that the solution converges with standard convergence rates w.r.t. the time, spatial, and stochastic discretization parameters, with constants independent of singular values.

We follow by presenting a residual-based a-posteriori error estimation for a heat equation with a random forcing term and a random diffusion coefficient which is assumed to depend affinely on a finite number of independent random variables. The a-posteriori error estimate consists of four parts: the finite element method error, the time discretization error, the stochastic collocation error, and the rank truncation error. These estimators

are then used to drive an adaptive choice of FE mesh, collocation points, time steps, and time-varying rank.
The last part of the thesis examines the idea of applying the DLR method in data assimilation problems, in particular the filtering problem. We propose two new filtering algorithms. They both rely on complementing the DLRA with a Gaussian component. More precisely, the DLR portion captures the non-Gaussian features in an evolving low-dimensional subspace through interacting particles, whereas each particle carries a Gaussian distribution on the whole ambient space. We study the effectiveness of these algorithms on a filtering problem for the Lorenz-96 system.

Keywords: Dynamical low rank, Dynamically orthogonal approximation, Uncertainty quantification, Data assimilation, Filtering problem, Error estimates, Splitting method

Résumé

La quantification des incertitudes peut être particulièrement difficile pour les problèmes nécessitant une intégration à long terme, car la structure de la solution aléatoire peut considérablement changer avec le temps. Pour répondre à cette problématique, l'approximation dynamique à faible rang (DLRA) est très attrayante. Elle peut être considérée comme une méthode de base réduite, donc soluble à un coût de calcul relativement faible, dans laquelle la solution est étendue comme une combinaison linéaire de fonctions déterministes avec des coefficients aléatoires. La caractéristique distinctive de la DLRA est que les fonctions déterministes et les coefficients aléatoires sont calculés à la volée et peuvent évoluer dans le temps, s'adaptant ainsi à chaque instant à la structure actuelle de la solution aléatoire. Ceci est réalisé en projetant de manière appropriée la dynamique sur l'espace tangent d'une variété constitué de toutes les fonctions aléatoires avec un rang fixe. Dans cette thèse, on cherche à analyser plus en profondeur et à appliquer les méthodes DLR aux problèmes dépendant du temps.

Notre première tâche considère la DLRA des équations paraboliques aléatoires et propose une classe de schémas numériques entièrement discrets. Comme pour la DLRA continue, on montre que nos schémas satisfont une formulation variationnelle discrète. En exploitant cette propriété, on établit la stabilité de nos schémas : on montre que nos versions explicite et semi-implicite sont conditionnellement stables sous une condition CFL de type "parabolique" qui ne dépend pas de la plus petite valeur singulière de la solution DLR, tandis que notre schéma implicite est inconditionnellement stable. De plus, on montre que, dans certains cas, le schéma semi-implicite peut être inconditionnellement stable si le caractère aléatoire du système est suffisamment petit. L'analyse est soutenue par des résultats numériques montrant la netteté des conditions de stabilité obtenues.

La formulation variationnelle discrète est appliquée dans la deuxième partie du travail, qui dérive des estimations d'erreur a-priori et a-posteriori pour la DLRA discrète d'une équation parabolique aléatoire obtenue par les trois schémas nouvellement proposés. En supposant que le côté droit du système dynamique se trouve dans l'espace tangent à un petit reste près, on montre que la solution converge avec des taux de convergence standard en fonction des paramètres de discrétisation temporels, spatiaux et stochastiques, avec des constantes indépendantes des valeurs singulières.

On présente ensuite une estimation de l'erreur a posteriori basée sur les résidus pour une équation de chaleur avec un terme de forçage aléatoire et un coefficient de diffusion aléatoire qui est supposé dépendre de manière affine d'un nombre fini de variables aléatoires

indépendantes. L'estimation de l'erreur a posteriori se compose de quatre parties : l'erreur d'éléments finis, l'erreur de discrétisation temporelle, l'erreur de collocation stochastique et l'erreur de troncature de rang. Ces estimateurs sont ensuite utilisés pour piloter un choix adaptatif de maillage FE, de points de collocation, de pas de temps, et de rang variable dans le temps.
La dernière partie de cette thèse examine l'idée d'appliquer la méthode DLR à des problèmes d'assimilation de données, en particulier le problème du filtrage. On propose deux nouveaux algorithmes de filtrage. Ils reposent tous deux sur le fait de compléter la DLR par une composante gaussienne. Plus précisément, la partie DLR capture les caractéristiques non gaussiennes dans un sous-espace évolutif de faible dimension par le biais de particules en interaction, tandis que chaque particule porte une distribution gaussienne sur l'ensemble de l'espace ambiant. On étudie l'efficacité de ces algorithmes sur un problème de filtrage pour le système Lorenz-96.

Mots clés : Approximation dynamique à faible rang, Approximation orthogonale dynamique, Quantification d'incertitude, Assimilation de données, Problème de filtrage, Estimations d'erreurs, Méthode de fractionnement

Contents

Acknowledgements **1**

Abstract (English/Français) **3**

Introduction **1**

0.1 Dynamical low rank approximation for forward UQ 1

0.1.1 Main contributions 4

0.2 Dynamical low rank approximation for data assimilation 7

0.2.1 Main contributions 9

I Discretization schemes for dynamical low rank approximation **11**

1 Introduction to dynamical low rank approximation **13**

1.1 Problem statement 13

1.2 Dynamical low rank approximation: equations 15

1.3 Geometrical interpretation and variational formulation 17

2 Projector-splitting schemes and their variational formulation **23**

2.1 Projector-splitting integrator for DLRA 25

2.2 A staggered time-marching scheme for DLR equations 31

2.2.1 Staggered time-marching scheme 31

2.2.2 Discrete variational formulation for the full-rank case 36

2.2.3 Discrete variational formulation for the rank-deficient case 40

2.2.4 Reinterpretation as a projector-splitting scheme 45

2.3 Linking the projector-splitting integrator and the staggered scheme . . . 46

3 Stability properties **51**

3.1 Problem specification 51

3.2 Stability of the continuous problem 53

3.3 Stability of the continuous DLR solution 54

3.4 Stability of the discrete DLR solution 55

3.4.1 Implicit Euler scheme 55

3.4.2 Explicit Euler scheme 57

3.4.3 Semi-implicit scheme 60
3.5 Example: random heat equation 63
3.5.1 Explicit Euler scheme 64
3.5.2 Semi-implicit scheme 64
3.6 Numerical results 66
3.6.1 Explicit scheme 67
3.6.2 Semi-implicit scheme 68
3.6.3 Comparison with the DDO projector-splitting scheme 72

4 A-priori error estimation 75
4.1 Problem specification 76
4.2 Error estimates without stochastic error contribution 79
4.3 Error estimate with stochastic error contribution 90

5 A-posteriori error estimation 93
5.1 A posteriori error estimation for a random heat equation 93
5.1.1 Problem statement 94
5.1.2 Discretization aspects 96
5.1.3 Fully discrete problem 99
5.1.4 Residual based a posteriori error estimation 100
5.1.5 Adaptive algorithm 109
5.1.6 Numerical results 111
5.2 A posteriori error estimation for a DLRA of a random parabolic equation 120
5.2.1 Discretization aspects 120
5.2.2 Fully discrete problem 122
5.2.3 Residual based a-posteriori error estimation for a general random parabolic equation 123
5.3 A posteriori error estimation for a DLRA of a random heat equation . . 126
5.3.1 Adaptive algorithm 135

II Dynamical low rank approximation for data assimilation 143

6 Discrete filtering problem: overview 145
6.1 Problem statement 145
6.2 Standard algorithms 147
6.2.1 The Kalman filter 148
6.2.2 Approximate Gaussian filters 149
6.2.3 The particle filter 150
6.2.4 Optimal proposal particle filter 152
6.2.5 The Gaussian mixture filter 153
6.3 Numerical example: Lorenz-96 155

7 Dynamical low rank approximation for filtering problems 161
7.1 Simple dynamical low rank approximation in the forecast step 161
7.1.1 Forecast step 162
7.1.2 Model error and analysis step: ensemble Kalman filter and particle filter 164
7.1.3 Numerical examples: ensemble Kalman filter and particle filter . 169
7.1.4 Model error and analysis step: Gaussian mixture filter 169
7.1.5 Numerical examples: Gaussian mixture filter 172
7.2 Complemented dynamical low rank approximation 173
7.2.1 Forecast step: DLRA with an independent Gaussian component 174
7.2.2 Forecast step: DLRA with a linear complement 181
7.2.3 Including the model error 186
7.2.4 Analysis step 187
7.2.5 Numerical results 189

8 Conclusions and perspectives 193

Bibliography 197

Curriculum Vitae 211

Introduction

The development in technology and science within the last decades demonstrates a growing importance of mathematical modeling and numerical simulations. The mathematical model only gives an approximate description of reality. Physical processes are often oversimplified, and the input data suffer from measurement errors or reduced availability. Therefore, a reliable mathematical model needs to account for a modeling error and uncertainty in the input data. Consider as an example the weather prediction. In order to obtain a weather forecast, one needs to compute the solution of a large system of time-dependent partial differential equations (PDEs), which only approximates the reality. The model's input data include physical parameters such as atmospheric emissions, humidity, soil moisture (and many more), and initial and boundary conditions. The exact quantification of most of these parameters is typically compromised by measurement errors, reduced amount of data, or intrinsic variability of the parameter itself. The uncertainty of the model and input parameters is naturally reflected in the uncertainty of the solution. In the context of uncertainty quantification (UQ), we can distinguish two main directions: *the forward UQ*, which quantifies the impact of uncertain inputs on the model outputs, and *the inverse UQ*, which aims to reduce the uncertainty of the model inputs using available experimental measurements of some model outputs. This thesis focuses on the forward UQ in the first part and data assimilation (which combines techniques of both forward and inverse UQ) in the second part.

0.1 Dynamical low rank approximation for forward UQ

In the present work, we consider a random evolutionary equation

$$\dot{u} = \mathcal{F}(u) \tag{1}$$

with random initial condition and random operator $\mathcal{F}$.

One of the most popular techniques for quantifying the uncertainty of such a problem is the *Monte Carlo (MC) method* (see, e.g., [Fis96; Caf98]). The solution statistics are approximated by drawing a sample of $\hat{N}$ independent realizations of the random data and solving corresponding $\hat{N}$ deterministic evolution equations. This leads to a set of $\hat{N}$

solutions $\{u_{(j)}\}_{j=1}^{\hat{N}}$ approximating the distribution of u at any time instant. The MC methods suffer from a slow convergence rate with respect to the number of particles. There have been many improvements built upon the classic MC method, as the Quasi Monte Carlo ([WS07; YY18; Gra+11; Nie92]) and the Multilevel Monte Carlo ([BL12; Cli+11]), among others.

An alternative approach to sampling methods of MC type is provided by spectral methods, which try to reconstruct the functional dependence of the solution on the random data. Suppose we can parametrize the randomness in terms of a finite-dimensional random vector ω with known distribution. Motivated by the observation that the parameter-to-solution map is smooth for many types of random equations, the method expands the random solution over a suitable stochastic basis $\{Y_i\}_{i=1}^{R}$

$$u^{\mathrm{gPC}}(t,\omega) = \sum_{i=1}^{R} U_i(t) Y_i(\omega), \tag{2}$$

where $\{U_i\}_{i=1}^{R}$ are deterministic coefficients and $\{Y_i\}_{i=1}^{R}$ are e.g. multivariate polynomial functions orthogonal w.r.t. the density function of the random vector ω. This approach is known as *generalized Polynomial Chaos (gPC) expansion*, see [XK02; XH05b; Wie38; SG04; CM47]. The coefficients $\{U_i\}_{i=1}^{R}$ can be obtained via Galerkin projection ([Bec+12; BTZ04; GS91; MK05]) or by e.g. stochastic collocation on tensor or sparse grids ([XH05b; MK10; BNT10]). Convergence rates are significantly higher compared to the standard MC methods, provided that the parameter-to-solution map has high regularity and the (effective) dimensionality of the stochastic space is not too large. This approach remains challenging, however, if the two conditions mentioned above are not met and, despite many improvements introduced via the sparse grid method or other sparsification techniques, spectral methods still suffer from the so-called curse of dimensionality, i.e., an accuracy versus cost performance that is negatively affected by the high dimension of the underlying stochastic space for many practical problems. An additional difficulty is posed by long-time integration of the problem (1). As the dependence of the solution on the random parameters may significantly vary in time, a set of fixed R stochastic basis functions might be insufficient to provide a good accuracy for long times. Naturally this can be overcome by increasing R – the number of terms in the expansion (2), however with a consequent increase in computational cost (see [Ger+10; WK06] for more details and examples).

A different strategy arises from the field of model order reduction (MOR). The underlying idea relies on the observation that for many types of problems, the collection of all realizations of the solution u of problem (1) for all possible input parameters at all times can be well approximated by a linear subspace $\mathcal{U}_R$ with a small dimension R. Assuming that we are able to parametrize the subspace $\mathcal{U}_R$ by a set of R deterministic functions

$\{U_i\}_{i=1}^R$, called reduced bases, the solution at each time can be expressed as

$$u^{\mathrm{RB}}(t) = \sum_{i=1}^{R} U_i Y_i(t) \tag{3}$$

where $\{Y_i\}_{i=1}^R$ are stochastic coefficients. The stochastic coefficients are then obtained by projecting the equations (1) onto $\mathcal{U}_R$, resulting in an R-dimensional system of equations. This technique is known as the *reduced basis method* (see e.g. [Boy+10; CQR14; CQR15; CS15; EL13] for more details). The procedure consists of two stages: an offline stage which computes the deterministic basis functions $\{U_i\}_{i=1}^R$ and an online stage in which the UQ problem (1) is reduced to a low-dimensional and consequently a low-cost problem, solving for $\{Y_i(t)\}_{i=1}^R$. The difficulty of the method lies in a good choice of the deterministic basis $\{U_i\}_{i=1}^R$ characterizing the solution subspace $\mathcal{U}_R$. One of the most popular techniques is the Proper Orthogonal Decomposition (POD) (see, e.g., [BGW07; CF11; MK10; WP02]), which collects snapshots of solutions precomputed for certain input parameters at certain time instants in a matrix and applies a singular value decomposition (SVD) to extract the R most dominant singular vectors. A considerable computational cost is required for this procedure in case of long time integration problems. Recently, greedy approaches have been proposed and applied in [Bau+15; Buf+12; CQR13; EKP11; GP05; Haa13; NRP09], trying to reduce the cost of the offline stage. However, applying any of these methods becomes challenging when the collection of solutions for different parameters considerably changes over time. For long time integration, this essentially leads to increasing R dramatically, with, again, a consequent increase in the offline and online computational cost.

We conclude that even if the solution of (1) allows a good low-rank approximation at each time instant, fixing either the deterministic basis (3) or the stochastic basis (2) for all times negatively influences the approximability properties of the expansion, leading to excessively large R for long time integration problems. The most straightforward approach to alleviate this issue is to allow both the deterministic functions $\{U_i\}_{i=1}^R$ and the stochastic functions $\{Y_i(y)\}_{i=1}^R$ to evolve in time

$$u^{\mathrm{DLR}}(t) = \sum_{i=1}^{R} U_i(t) Y_i(t), \tag{4}$$

so that they can best approximate the solution at a given time instant. This is achieved by fixing the rank R in time and imposing the functions (also called modes) $\{U_i\}_{i=1}^R$ and $\{Y_i\}_{i=1}^R$ to be linearly independent. The collection of all such random fields forms a manifold, denoted by $\mathcal{M}_R$, and the evolution of the modes is derived from projecting the governing equation (1) onto the tangent space of such manifold at the current solution u^{DLR}. This approach named *dynamically orthogonal (DO) field expansion* was first proposed in the UQ context in [SL09] and applied to problems in ocean dynamics with random data in [SL12; ULS13]. Similar ideas, using slightly different representation form for the solution, have been proposed in [CHZ13a; CHZ13b; CSK14] under the names dynamically double orthogonal and bi-orthogonal expansions, respectively. In [MNZ15] it was shown that these formulations are, in fact, equivalent. In [KKS06; MMC90; Zan+03; Bec+99], a similar approach, known as multi-configuration time-dependent Hartree (MCTDH) method, was used to approximate time-dependent Schrödinger equations. The name *Dynamical low rank approximation* comes from [HLW04; KL07a], where analogous ideas were applied in the context of matrix evolution equations (we will use the acronym DLR for dynamical low rank and DLRA for dynamical low rank approximation). This was further extended to deal with tensors in Tucker format in [KL10; KL07b; CL10], in Hierarchical Tucker or Tensor train in [LOV15a; Lub+13] and tree tensor networks in [CLW21]. In [FL18], the authors provide a thorough analysis of the geometry corresponding to this method in the matrix setting. Applications of the DLRA include stiff matrix differential equations ([Men+18; OPW19]), multi-scale linear transport equation ([EHW21]), different types of Vlasov equations [EL18; EOP20; EJ21], Navier-Stokes equations [EHY21; MN18] and wave equations [MNV17]. A dynamical low-rank approximation with a different choice of time-dependent deterministic basis was considered in [FN17].

The analysis and development of the dynamical low rank approximation method is the central topic of this thesis. In the field of forward UQ, our main focus was the analysis of efficient discretization schemes for the DLRA of random PDEs. We will follow now with descriptions of our main contributions

0.1.1 Main contributions

New fully discrete schemes for DLRA

We propose a new class of fully discrete schemes used to approximate time-dependent partial differential equations with random parameters, stemming from the evolution equations for the modes $\{U_i\}_{i=1}^R, \{Y_i\}_{i=1}^R$. This results in explicit, implicit and semi-implicit schemes. Although not evident at first sight, we show that the explicit version of our scheme can be reinterpreted as a projector-splitting scheme (whenever the discrete solution is full-rank), which is a popular DLR integrator (see e.g. [LO14; LOV15b]) with highly advantageous properties in the presence of small singular values in the solution ([KLW16]). In the rank-deficient case, our schemes may result in different solutions. In Chapter 2, we show that the discrete DLR solution obtained by the newly-proposed schemes satisfies a discrete variational formulation, analogous to the variational formulation of the continuous DLR problem (see [MNZ15, Prop. 3.4]). Such formulation is then key for analyzing the stability properties and deriving a-priori and a-posteriori error estimates, which is the goal of Chapters 3, 4 and 5.

Stability analysis

Chapter 3 is dedicated to analysing the stability properties of the proposed numerical schemes applied to a parabolic PDE with random coefficients. We first show that in this parabolic setting, the continuous DLR solution satisfies analogous stability properties as the true solution of the considered problem, in the parabolic case. We then analyze the stability of the fully discrete schemes from Chapter 2. Quite surprisingly, the stability properties of both the discrete and the continuous DLR solutions do not depend on the size of their singular values, even without any ε-approximability condition on $\mathcal{F}$. The implicit scheme is proven to be unconditionally stable. This improves the stability result which could be drawn from the error estimates derived in [KLW16]. The explicit scheme remains stable under a standard parabolic stability condition between time and space discretization parameters for an explicit propagation of parabolic equations. The semi-implicit scheme is generally only conditionally stable under again a parabolic stability condition, and becomes unconditionally stable under some restrictions on the size of the randomness of the operator. As an application of the general theory developed in this work, we consider the case of a heat equation with a random diffusion coefficient. We dedicate a section to particularize the numerical schemes and the corresponding stability results to this problem. The semi-implicit scheme turns out to be always unconditionally stable if the diffusion coefficient depends affinely on the random variables. To the best of our knowledge, this is the first work providing a stability analysis for a fully discrete DLR solution obtained by a projector-splitting scheme and the results are published in [KNV21].

A-priori error estimation

Chapter 4 deals with an a-priori error estimation for the fully discrete DLR solution of a random parabolic equation obtained by the newly-proposed schemes of Chapter 2. The considered discrete DLR solution is obtained by i) applying the finite element method with continuous piece-wise polynomials of degree $\leq r$ and elements of size h to treat the spatial discretization; ii) the Monte Carlo method with $\hat{N}$ samples to treat the stochastic discretization; iii) the time-marching scheme proposed in Chapter 2 with time step $\triangle t$. Applying the discrete variational formulation derived in Chapter 2, we are able to adapt the standard techniques derived for deterministic parabolic problems to our setting and prove an $O(\varepsilon + \triangle t + h^r + 1/\sqrt{\hat{N}})$ error bound, under the assumptions that the projection error of $\mathcal{F}(u)$ to the tangent space of $\mathcal{M}_R$ at u is of size ε, that the initial condition is of rank R, that the true solution is sufficiently smooth and an additional 'stability condition' on the deterministic basis. An important property of our error analysis is that all involved constants are independent of the smallest singular values of the solution. In the context of matrix evolution equations, the work [KLW16] proves a similar error bound $O(\varepsilon + \triangle t)$ with constants independent of the singular values. As pointed out by their authors, a limitation of their theoretical result is that it is applicable to stiff differential equations such as discretized PDEs only under a severe CFL condition $\triangle t L \ll 1$, where L is the

Lipschitz constant for $\mathcal{F}$. Such a restriction is not present in our analysis. Furthermore, in our setting, the operator $\mathcal{F}$ does not need to be uniformly bounded. On the other hand, as opposed to the work in [KLW16], our results only apply to parabolic problems.

A-posteriori error estimation

Chapter 5 is concerned with an a-posteriori error estimation for a discrete DLR solution of a random parabolic equation obtained by the newly-proposed schemes. The spatial discretization is obtained by applying the finite element method with continuous piecewise polynomials, while the stochastic discretization is performed by the stochastic collocation (SC) method [XH05a; BNT10; NTW08a]. Before tackling the problem for a DLR solution, we direct our attention to an a-posteriori error estimation for a random heat equation, with no DLRA involved. In this case, the time discretization is performed via the θ-scheme and the stochastic discretization using sparse grids. There is a vast literature on a posteriori error estimation for deterministic parabolic problems (see e.g. [EJ91; EJ95; Pic98; Ver03; LPP09; AMN06]). There is much less literature available for the a posteriori error estimation for random PDEs. When uncertainties are treated by the stochastic Galerkin method [GDS03; LMK10], a posteriori error estimations together with an adaptive algorithm have been proposed in [Kha+20; KPB18] for a linear elasticity equation and in [Eig+13; Eig+14; Eig+15; BPS14; CPB19; BX20] for an elliptic PDE. Concerning parabolic equations, the only work we are aware of considers uncertainty only in the Robin boundary condition, solved by the perturbation approach in [Gui18]. The work [GN18] derives a residual based a posteriori error estimation for an elliptic problem discretized by a stochastic collocation finite element method. There, the authors propose an algorithm that adaptively builds the sparse grid based on the a posteriori estimation of the SC error. Recently, a proof of convergence for such adaptive algorithm has been derived in [FS21; Eig+21].

Our work extends the results obtained in [GN18] to a heat equation with random right hand side and random diffusion coefficient that depends affinely on a finite number of random variables. We adopt the setting from [Ver03] to treat the spatial and time discretization errors. Our estimator allows spatial meshes and stochastic sparse grids to change in time. The estimator naturally splits into a spatial discretization estimator, time discretization estimator and stochastic discretization estimator, which are then used to drive the adaptivity with respect to all three types of discretizations. We then propose an adaptive algorithm to build a suitable time discretization and a FE mesh and sparse grid common to all time steps, so as to achieve a prescribed tolerance on a global norm of the error. We then apply this to a problem with a deterministic right hand side and a diffusion coefficient depending affinely on few random variables. These results have been published in [NV19].

We follow by applying analogous techniques to derive an a-posteriori error bound for a DLR solution of a random parabolic equation. In this case, the stochastic discretization applies the tensor grid collocation method, spatial discretization the finite element method and the time discretization follows the projector-splitting scheme proposed in Chapter 2. Using the discrete variational formulation, we manage to split the estimator into a spatial discretization estimator, time discretization estimator, stochastic discretization estimator and a rank truncation estimator. For the case of a heat equation with diffusion coefficient affine w.r.t. a few random variables, we derive fully computable error estimators for all four error contributors. Similarly as before, we propose an adaptive algorithm to build a suitable time discretization, an FE mesh and tensor grid common to all time steps, and a rank for the DLRA which is allowed to change for different time steps, with the goal to achieve a prescribed tolerance on a global norm of the error. The performance of the algorithm is yet to be tested and is a part of our future research directions. We point the reader to [CKL22] for a different algorithm that can be used to adapt the choice of the DLR rank. To the best of our knowledge, both a-priori and a-posteriori error estimators available in this work are the first results of this kind, derived for a fully discrete DLRA of a solution to a PDE with random parameters.

0.2 Dynamical low rank approximation for data assimilation

Alongside the development of mathematical models describing many real-world phenomena, the last decades have witnessed a growing availability of data sets in almost all engineering, science, and technology areas. When the underlying mathematical model is a (possibly stochastic) dynamical system, and the data are ordered in time, the term *data assimilation* refers to the effort of combining the data with the mathematical model. The main application areas of data assimilation include atmospheric and oceanographic sciences and other areas of geoscience. With the current environmental crisis, the area of weather prediction and climate modeling offers high demands for efficient data assimilation techniques.

There are two distinct areas in this field: the *filtering problem*, which tries to update the knowledge of the state at time t (and possibly reduce its uncertainty) by taking into account the data up to time t (i.e., data from the past and present); and the *smoothing problem*, which allows updating the knowledge of the state at time t by data from the past, present, and future. In this work, we will focus on the filtering problem.

More precisely, let us denote the state (signal) at time t by $u(t)$. Suppose that the signal is governed by the (possibly stochastic) dynamics

$$\dot{u} = \mathcal{F}(u), \tag{5}$$

that the initial state $u(0)$ has probability distribution μ^0 on V, and that observations become available at discrete times $t^1, t^2, \dots$, in the form

$$z^n = H(u(t^n)) + \eta^n, \; n \in \mathbb{N},$$

where $H : V \to \mathbb{R}^l$ is an observation operator and $\eta^n \in \mathbb{R}^p, \; n = 1, 2, \dots$ are independent noise terms following a probability density function $p_n : \mathbb{R}^p \to \mathbb{R}_+$. The goal is then to compute the probability distribution of $u^n = u(t^n)$ conditional on the observations $Z^n = \{z^1, \dots, z^n\}$ collected up to time t^n, denoted by $\mathbb{P}(u^n|Z^n)$. Such calculation can be split in two steps:

- *prediction*: given $\mathbb{P}(u^n|Z^n)$ one computes first $\mathbb{P}(u^{n+1}|Z^n)$, called the forecasted/predicted distribution, by evolving $\mathbb{P}(u^n|Z^n)$ through the system (5) up to time t^{n+1};
- *analysis*: one then computes the new conditional measure $\mathbb{P}(u^{n+1}|Z^{n+1})$, called the filtering distribution, by incorporating the newly observed data z^{n+1} using the Bayes' formula.

The literature on data assimilation is vast, and mostly driven by applications in oceanography, atmospheric sciences, weather prediction or oil recovery [Kal02; Ben02; Eve09; Aba13; MH12; RC15]. We point to the book [LSZ15] for an introduction to the topic and its mathematical foundations. Various techniques have been developed over the years to deal with the problem of filtering. For linear systems with additive Gaussian noise, the Kalman filter (see [Kal60]) provides an exact algorithm to determine the filtering distribution. Several extensions of Kalman filter have been proposed to enable applications to nonlinear problems; we mention the extended Kalman filter [Jaz70] based on linearized dynamics and the ensemble Kalman filter (EnKF) [Eve09; LE96b; LGMT09; LX08], which samples the dynamics yet still relying on a Gaussian approximation in the analysis step. Kalman-based filters are robust with respect to the noise in the system and observations, however they do not reproduce the correct posterior filtering distribution for general nonlinear problems, as they all rely on invoking some Gaussian ansatz in the algorithm. This issue is alleviated in another class of filters that has been widely employed, the particle filters (PF). These are purely sampling based filters, which propagate an empirical distribution through the system and update it (still as an empirical distribution) in the analysis step. As such, they provide a consistent approximation of the true filtering distribution in the limit of an infinite number of particles (see e.g. [LSZ15; CD02; RH15; Lee+19]). The main limitation of particle filters developed so far is that their efficiency degrades with the dimensionality of the problem. This is a topic of active research in the field.

The real-world applications of data assimilation often involve a very high-dimensional problem in the forecast step. When running a full-order model is extremely expensive, one needs to rely on reduced-order modeling techniques. These typically consist in looking for a solution in a low-dimensional subspace. Works that apply some reduced-basis approaches include [SSN15] for combining POD and DEIM with 4DVAR, and [Cas+20] for combining reduced order models and deep neural networks for more efficient predictions.

Especially in the context of data assimilation, the ‘optimal’ subspace that approximates well the whole solution (or a large ensemble of particles) can significantly vary in time. In this respect, employing the dynamical low rank approximation (DLRA) in the forecast step seems very advantageous. The dominant subspace evolves in time, adjusting to the underlying dynamical system at every time instant as well as the incoming observations. The idea of applying DLR to filtering problems was recently explored in a few works. In [SL13a], the authors use a DLR approximation in the prediction step. The forecasted measure is then approximated by a Gaussian mixture model (GMM) and updated by Kalman formulas in the analysis step. This strategy was applied in [SL13b] to deal with double well diffusion problem and sudden expansion flows. A different strategy was proposed in [MQS14; SM13; MS13] to treat turbulent dynamical systems. The prediction step involves propagating an accurate approximation of the DLR solution in the low-dimensional subspace coupled with a second order Gaussian closure solution in the full space. At the beginning of the analysis step, the two solutions are then blended into a conditional Gaussian particle distribution and updated via Bayes’ formula.

0.2.1 Main contributions

In our work, we start by exploring the idea of applying simple DLRA in the forecast step, combined with standard algorithms in the analysis step (namely EnKF and PF). Applying this strategy to the 40-dimensional Lorenz-96 system of equations (a simplified model for atmospheric processes), we will see that completely dismissing the ommited modes in the DLRA leads to loss of accuracy which has significant consequences in the performance of the filters, including loss of information provided by the data. To improve upon the simple DLRA, we complement the DLR signal by a Gaussian component. This approach resembles the work introduced in [MQS14; QJM15; SM13]. However, we provide a consistent way of complementing the DLRA of the signal by enlarging the set of test functions in the DLR variational formulation (more details on the comparison of these two works is available at the end of Section 7.2.4). We propose two algorithms that complement the DLR signal by a term linear w.r.t. normally distributed random variables, which then constitutes the Gaussian component. The first algorithm imposes the DLR and linear term to be independent, whilst the second algorithm does not. At the beginning of the analysis step, we build a Gaussian mixture distribution and update it via Bayes’ formula. The numerical examples offer multiple scenarios which compare the behaviour of these two algorithms with the simple DLR in the forecast step. We will see that in the case of long observation times and large observation error, imposing independence of the DLR and linear term causes unsatisfactory performance.

The outline of the work is the following. In Chapter 6, we detail the mathematical formulation of the filtering problem, provide an overview of the standard techniques and present numerical results that apply these techniques to a 40-dimensional Lorenz-96 system of equations. Afterwards, in Chapter 7 we start with exploring the behaviour of simple DLR in the forecast step, follow by describing the two newly-proposed algorithms that complement the DLR signal by a Gaussian component and provide numerical results comparing the proposed methods.

Part I

Discretization schemes for dynamical low rank approximation

1 Introduction to dynamical low rank approximation

The considered DLR approximation of the solution is of the form

$$u(t) = \bar{u}(t) + \sum_{j=1}^{R} U_j(t) Y_j(t), \qquad t \in (0, T], \tag{1.1}$$

where R is the *rank* of the approximation, $\bar{u}(t) = \mathbb{E}[u(t)]$ is the mean value of the DLR solution, $\{U_j(t)\}_{j=1}^R$ is a time dependent set of deterministic basis functions, $\{Y_j(t)\}_{j=1}^R$ is a time dependent set of zero mean stochastic basis functions. By suitably projecting the residual of the differential equation, one can derive evolution equations for the mean value $\bar{u}$ and the deterministic and stochastic modes $\{U_j\}_{j=1}^R$, $\{Y_j\}_{j=1}^R$ (see [SL09; KL07a]). In the derivation of the method, the rank R is fixed in time. This condition is however alleviated in Chapter 5, where an adaptive algorithm for a time-dependent choice of rank is proposed. In this chapter, we start with stating the underlying problem in Section 1.1. We follow by defining the dynamical low-rank approximation in Section 1.2 and recall the equations for the modes. Further, we show that the DLR solution can be equivalently defined as a solution of a variational formulation, for which we set up a geometrical framework in Section 1.3. We point out that the details provided in this chapter are not new and serve only as a summary of well-known results.

1.1 Problem statement

We start by introducing some notation. Let $(\Omega, \mathcal{A}, \rho)$ be a probability space, where Ω is a set of outcomes, $\mathcal{A}$ a σ-algebra and $\rho : \mathcal{A} \to [0, 1]$ a probability measure. Consider the Hilbert space $L^2_\rho = L^2_\rho(\Omega)$ of real valued random variables on Ω with bounded second moments

$$L^2_\rho(\Omega) = \{v : \Omega \to \mathbb{R} \,\text{s.t.} \int_\Omega |v|^2 \,\mathrm{d}\rho < \infty\},$$

with associated scalar product $\langle v, w\rangle_{L^2_\rho} = \int_\Omega vw \,\mathrm{d}\rho$ and norm $\|v\|_{L^2_\rho} = \sqrt{\langle v, v\rangle_{L^2_\rho}}$.

Consider as well two separable Hilbert spaces H and V with scalar products $\langle\cdot,\cdot\rangle_H$, $\langle\cdot,\cdot\rangle_V$, respectively. Suppose that H and V form a Gelfand triple (V, H, V'), i.e. V is a dense subspace of H and the embedding $V \hookrightarrow H$ is continuous with a continuity constant $C_{\mathrm{P}} > 0$. Let $L^2_\rho(\Omega; V)$, $L^2_\rho(\Omega; H)$ be the Bochner spaces of square integrable V (resp. H) valued functions on Ω with scalar products

$$\begin{aligned}\langle v, w\rangle_{H,L^2_\rho} &= \int_\Omega \langle v, w\rangle_H \,\mathrm{d}\rho, \qquad v, w \in L^2_\rho(\Omega; H)\\ \langle v, w\rangle_{V,L^2_\rho} &= \int_\Omega \langle v, w\rangle_V \,\mathrm{d}\rho, \qquad v, w \in L^2_\rho(\Omega; V),\end{aligned}$$

respectively. Then, $(L^2_\rho(\Omega; V),\ L^2_\rho(\Omega; H),\ L^2_\rho(\Omega; V'))$ is a Gelfand triple as well (see e.g. [Leo17, Th. 8.17]), and we have

$$\|v\|_{H,L^2_\rho} \le C_{\mathrm{P}} \|v\|_{V,L^2_\rho} \qquad \forall v \in L^2_\rho(\Omega; V). \tag{1.2}$$

We define the mean value of an integrable random variable v as $\mathbb{E}[v] = \int_\Omega v(\omega)\,\mathrm{d}\rho(\omega)$, where the integral here denotes the Bochner integral in a suitable sense, depending on the co-domain of the random variable considered. In what follows, we will use the notation $\bar{v}$ to denote the mean value of v and $v^* := v - \bar{v}$ to denote the derivation of v from its mean value. Moreover, we let $(\cdot,\cdot)_{V'V,L^2_\rho}$ denote the duality pairing between $L^2_\rho(\Omega; V')$ and $L^2_\rho(\Omega; V)$:

$$\text{for } \mathcal{K} \in L^2_\rho(\Omega; V'),\ v \in L^2_\rho(\Omega; V), \qquad (\mathcal{K}, v)_{V'V,L^2_\rho} := \int_\Omega \Big(\mathcal{K}(\omega), v(\omega)\Big)_{V'V} \,\mathrm{d}\rho(\omega).$$

The problem considered in this work is the following random evolution equation. Given a final time $T > 0$ and a random initial condition $u_0 \in L^2_\rho(\Omega; V)$, the problem states: Find a solution $u_{\mathrm{true}} \in L^2(0, T; L^2_\rho(\Omega; V))$ with $\dot{u}_{\mathrm{true}} \in L^2(0, T; L^2_\rho(\Omega; V'))$ satisfying

$$\begin{aligned}\Big(\dot{u}_{\mathrm{true}}, v\Big)_{V'V,L^2_\rho} &= \Big(\mathcal{F}(u_{\mathrm{true}}), v\Big)_{V'V,L^2_\rho}, \quad \forall v \in L^2_\rho(\Omega; V),\ \text{a.e. } t \in (0, T]\\ u_{\mathrm{true}}(0) &= u^0,\end{aligned} \tag{1.3}$$

where $\mathcal{F}$ is a random linear or nonlinear differential operator. We assume that equation (1.3) admits a unique solution $u_{\mathrm{true}} \in L^2(0, T; L^2_\rho(\Omega; V))$ for a.e. $t \in [0, T)$. The operator $\mathcal{F}$ will be further specified in Chapter 3 to describe parabolic problems.

The DLR approximation is closely related to the Karhunen-Loève expansion, which we detail in the following. Let $u \in L^2_\rho(\Omega; H)$ be a random field. We define the covariance operator $\mathcal{C}_u : H \to H$ as

$$\langle \mathcal{C}_u v, w\rangle_H = \mathbb{E}[\langle u - \bar{u}, v\rangle_H \langle u - \bar{u}, w\rangle_H] \qquad \forall v, w \in H,$$

which is self-adjoint and compact. Consider the sequence of non-negative decreasing eigenvalues of $\mathcal{C}_u$, $\{\mu_i\}_{i=1}^{\infty}$, and the corresponding sequence of orthonormal eigenfunctions $\{Z_i\}_{i=1}^{\infty} \subset H$, satisfying

$$\mathcal{C}_u Z_i = \mu_i Z_i, \quad \langle Z_i, Z_j \rangle_H = \delta_{ij} \quad \forall i, j \in \mathbb{N}^+.$$

The random variables

$$\gamma_i(\omega) = \frac{1}{\sqrt{\mu_i}} \langle u^*, Z_i \rangle_H$$

are centered, mutually uncorrelated with unit variance, i.e.

$$\mathbb{E}[\gamma_i] = 0, \quad \mathbb{E}[\gamma_i \gamma_j] = \delta_{ij}, \qquad i, j \in N^+.$$

The Karhunen-Loève expansion of the random function $u \in L^2_\rho(\Omega; H)$ is given by

$$u(\omega) = \bar{u} + \sum_{j=1}^{\infty} \sqrt{\mu_i} \gamma_i(\omega) Z_i$$

(see e.g. [GS91; LPS14] for more details). A truncated Karhunen-Loève expansion with rank R of the random function u is a truncation of the preceding expansion, i.e.

$$u^R(\omega) = \bar{u} + \sum_{j=1}^{R} \sqrt{\mu_i} \gamma_i(\omega) Z_i \tag{1.4}$$

and results in the best R-rank approximation of u w.r.t. the $\| \cdot \|_{H, L^2_\rho}$-norm. The decay properties of eigenvalues have been investigated e.g. in the works [GS91; ST06].

1.2 Dynamical low rank approximation: equations

Dynamical low rank (DLR) approximation, or dynamically orthogonal (DO) approximation (see e.g. [KL07a; SL09; KL07b]) seeks an approximation of the solution u_{true} of problem (1.3) in the form

$$u(t) = \bar{u}(t) + \sum_{j=1}^{R} U_j(t) Y_j(t) = \bar{u}(t) + U(t) Y(t)^\intercal, \qquad t \in [0, T] \tag{1.5}$$

where $\bar{u}(t) \in V$, $U(t) = (U_1(t), \dots, U_R(t)) \subset V$ is a time dependent set of linearly independent deterministic basis functions, which we will call deterministic modes, $Y = (Y_1(t), \dots, Y_R(t)) \subset L^2_\rho$ is a time dependent set of linearly independent stochastic basis functions, called stochastic modes. We call R the rank of a function u. Notice that $U(t)$ and $Y(t)$ are interpreted here as row vectors.

There are three main formulations determining a unique representation for (1.5). The DO formulation, proposed and applied in [SL09; SL12; ULS13], keeps the deterministic modes $\{U_r\}_{r=1}^R$ orthonormal in H at all times and the uniqueness of the representation is guaranteed by the following dynamically orthogonal (DO) conditions:

$$\langle U_i(t), U_j(t)\rangle_H = \delta_{ij}, \quad \langle \dot{U}_i(t), U_j(t)\rangle_H = 0, \quad \mathbb{E}[Y_j(t)] = 0, \quad \forall 1 \leq i, j \leq R, \quad \forall t \in [0, T].$$

The dual DO formulation ([MN18]) keeps the stochastic modes orthonormal, the deterministic modes linearly independent and is characterized by the following DO conditions:

$$\langle Y_i(t), Y_j(t)\rangle_{L^2_\rho} = \delta_{ij}, \quad \langle \dot{Y}_i(t), Y_j(t)\rangle_{L^2_\rho} = 0 \quad \mathbb{E}[Y_j(t)] = 0, \quad \forall 1 \leq i, j \leq R, \quad \forall t \in [0, T]. \tag{1.6}$$

The last form is the so called double dynamically orthogonal (DDO) or bi-orthogonal formulation, where the solution is sought in the form

$$u(t) = \bar{u}(t) + \sum_{i,j=1}^R S_{ij}(t) U_i(t) Y_j(t) = \bar{u}(t) + USY^\intercal, \qquad t \in [0, T] \tag{1.7}$$

with both deterministic and stochastic modes orthonormal in their respective Hilbert spaces and the matrix $S \in \mathbb{R}^{R\times R}$ of full rank (see e.g. [CHZ13a; CHZ13b; KL07a]). The corresponding DO conditions are

$$\begin{gathered}\langle U_i(t), U_j(t)\rangle_H = \delta_{ij}, \quad \langle Y_i(t), Y_j(t)\rangle_{L^2_\rho} = \delta_{ij}, \quad \mathbb{E}[Y_j(t)] = 0,\\ \langle \dot{Y}_i(t), Y_j(t)\rangle_{L^2_\rho} = 0, \quad \langle \dot{U}_i(t), U_j(t)\rangle_H = 0, \quad \forall 1 \leq i, j \leq R, \quad \forall t \in [0, T].\end{gathered}$$

In [CSK14; MNZ15], it was shown that these formulations are equivalent. In our work, we consider the dual DO formulation (1.6).

Plugging the DLR expansion (1.5) into the equation (1.3) and following analogous steps as proposed in [SL09] leads to the DLR system of equations presented next.

Definition 1.2.1 (DLR solution)**.** We define the DLR solution of problem (1.3) as

$$u(t) = \bar{u}(t) + \sum_{i=1}^R U_i(t) Y_i(t) \quad \in L^2_\rho(\Omega; V)$$

where $\bar{u}$, $\{U_i\}_{i=1}^R$, $\{Y_i\}_{i=1}^R$ are solutions of the following system of equations:

$$(\dot{\bar{u}}, v)_{V'V} = (\mathbb{E}[\mathcal{F}(u)], v)_{V'V} \qquad \forall v \in V \tag{1.8}$$

$$(\dot{U}_j, v)_{V'V} = (\mathbb{E}[\mathcal{F}(u) Y_j], v)_{V'V} \qquad \forall v \in V,\ j = 1, \ldots, R \tag{1.9}$$

$$\dot{Y}_j - \sum_{i=1}^R (M^{-1})_{j,i} \mathcal{P}^\perp_{\mathcal{Y}} \Big[(\mathcal{F}^*(u), U_i)_{V'V} \Big] = 0 \qquad \text{in } L^2_\rho,\ j = 1, \ldots, R \tag{1.10}$$

with the initial conditions $\bar{u}(0)$, $\{Y_j(0)\}_{j=1}^R$, $\{U_j(0)\}_{j=1}^R$ such that $\bar{u}(0) \in V$, $\{Y_j(0)\}_{j=1}^R$ satisfies the conditions (1.6), $\{U_j(0)\}_{j=1}^R$ are linearly independent in V, and $\bar{u}(0)$ +

$\sum_{j=1}^{R} Y_j(0)U_j(0)$ is a good approximation of u_0. In (1.10), the matrix $M \in \mathbb{R}^{R\times R}$ is defined as $M_{ij} := \langle U_i, U_j\rangle_H$, $1 \leq i,j \leq R$ and $\mathcal{P}_{\mathcal{Y}}^{\perp}$ denotes the orthogonal projection operator in the space $L^2_\rho(\Omega)$ on the orthogonal complement of the R-dimensional subspace $\mathcal{Y} = \operatorname{span}\{Y_1, \dots, Y_R\}$, i.e.

$$\mathcal{P}_{\mathcal{Y}}^{\perp}[v] = v - \mathcal{P}_{\mathcal{Y}}[v] = v - \sum_{j=1}^{R} \langle v, Y_j\rangle_{L^2_\rho} Y_j, \qquad \text{for } v \in L^2_\rho. \tag{1.11}$$

For the initial condition one can use for instance the truncated Karhunen-Loève expansion of u^0 described in (1.4).

1.3 Geometrical interpretation and variational formulation

This subsection gives a geometrical interpretation of the DLR method and follows to a large extent derivations from [MNZ15]. Such geometrical interpretation provides a valuable insight into the method, which brings along various approaches to obtain effective discretization schemes, described in Chapter 2. We first introduce the notion of a manifold of R-rank functions, characterize its tangent space in a point as well as the orthogonal projection onto the tangent space.

The vector space consisting of all square integrable random variables with zero mean value will be denoted by $L^2_{\rho,0} = L^2_{\rho,0}(\Omega) \subset L^2_\rho(\Omega)$. The set of all random functions $v \in L^2_\rho(\Omega; V)$ with a fixed rank R forms a manifold which we will denote by $\mathcal{M}_R$ and can be parametrized in the following way.

Definition 1.3.1 (Manifold of R-rank functions)**.** By $\mathcal{M}_R \subset L^2_{\rho,0}(\Omega; V)$ we denote the manifold consisting of all rank R random functions with zero mean

$$\begin{aligned}\mathcal{M}_R = \Big\{ v^* \in L^2_{\rho,0}(\Omega; V) \mid v^* = \sum_{i=1}^{R} U_i Y_i = UY^{\intercal}, \\ \langle Y_i, Y_j\rangle_{L^2_\rho} = \delta_{ij},\ \forall 1 \leq i,j \leq R,\ \{U_i\}_{i=1}^{R} \text{ linearly independent} \Big\}.\end{aligned} \tag{1.12}$$

It is well known that $\mathcal{M}_R$ admits an infinite dimensional Riemannian manifold structure ([FHN19]).

Proposition 1.3.1 (Tangent space at $UY^{\intercal}$)**.** The tangent space $\mathcal{T}_{UY^{\intercal}}\mathcal{M}_R$ at a point

$UY^\intercal \in \mathcal{M}_R$ can be characterized as

$$\mathcal{T}_{UY^\intercal}\mathcal{M}_R = \Big\{\delta v \in L^2_{\rho,0}(\Omega; V) \,|\, \delta v = \sum_{i=1}^{R} U_i \delta Y_i + \delta U_i Y_i, \\ \delta U_i \in V,\ \delta Y_i \in L^2_{\rho,0},\ \langle \delta Y_i, Y_j\rangle_{L^2_\rho} = 0, \forall 1 \le i, j \le R\Big\}. \tag{1.13}$$

Further, we define an orthogonal projection onto the tangent space.

Proposition 1.3.2 (Orthogonal projection on $\mathcal{T}_{UY^\intercal}\mathcal{M}_R$)**.** The $L^2_{\rho,0}(\Omega; H)$-orthogonal projection $\Pi_{UY^\intercal}[v]$ of a function $v \in L^2_\rho(\Omega, H)$ onto the tangent space $\mathcal{T}_{UY^\intercal}\mathcal{M}_R$ is given by

$$\begin{aligned} \Pi_{UY^\intercal}[v] &= \sum_{i=1}^{R} \langle v, Y_i\rangle_{L^2_\rho} Y_i + \mathcal{P}^\perp_{\mathcal{Y}}[\sum_{i=1}^{R} \langle v, U_i\rangle_H (M^{-1}U^\intercal)_i] \\ &= \mathcal{P}_{\mathcal{Y}}[v] + \mathcal{P}^\perp_{\mathcal{Y}}\Big[\mathcal{P}_{\mathcal{U}}[v]\Big] = \mathcal{P}_{\mathcal{Y}}[v] + \mathcal{P}_{\mathcal{U}}[v] - \mathcal{P}_{\mathcal{Y}}\Big[\mathcal{P}_{\mathcal{U}}[v]\Big], \end{aligned} \tag{1.14}$$

where $\mathcal{U} = \operatorname{span}\{U_1, \dots, U_R\}$ and $\mathcal{P}_{\mathcal{U}}[\cdot]$ is the H-orthogonal projection onto the subspace $\mathcal{U}$.

For more details, see e.g. [MNZ15]. Note that $\Pi_{UY^\intercal}[\cdot]$ can be equivalently written as $\Pi_{UY^\intercal}[\cdot] = \mathcal{P}_{\mathcal{U}}[\cdot] + \mathcal{P}^\perp_{\mathcal{U}}\Big[\mathcal{P}_{\mathcal{Y}}[\cdot]\Big]$. In the following we will extend the domain of the projection operator $\Pi_{UY^\intercal}$. Further, we will state two lemmas used to establish Theorem 1.3.5, which presents the variational formulation of the DLR approximation.

The operator $\Pi_{UY^\intercal}$ can be extended to an operator from $L^2_\rho(\Omega; V')$ to $L^2_\rho(\Omega; V')$ as

$$\Pi_{UY^\intercal}[\mathcal{K}] := \langle \mathcal{K}, Y\rangle_{L^2_\rho} Y^\intercal + \mathcal{P}^\perp_{\mathcal{Y}}\Big[(\mathcal{K}, U)_{V'V} M^{-1}U^\intercal\Big] \qquad \forall \mathcal{K} \in L^2_\rho(\Omega; V').$$

The extended operator satisfies the following properties.

Lemma 1.3.3. *Let $UY^\intercal \in \mathcal{M}_R$. Then it holds*

$$(\mathcal{K}, \Pi_{UY^\intercal}[v])_{V'V, L^2_\rho} = (\Pi_{UY^\intercal}[\mathcal{K}], v)_{V'V, L^2_\rho}, \quad \forall v \in L^2_\rho(\Omega; V),\ \mathcal{K} \in L^2_\rho(\Omega; V'). \tag{1.15}$$

Proof. First, we show that

$$(\mathcal{K},\, \mathcal{P}_{\mathcal{Y}}[v])_{V'V, L^2_\rho} = (\mathcal{P}_{\mathcal{Y}}[\mathcal{K}],\, v)_{V'V, L^2_\rho} \quad \forall v \in L^2_\rho(\Omega; V),\ \mathcal{K} \in L^2_\rho(\Omega; V').$$

Indeed,

$$
\begin{aligned}
(\mathcal{K},\, \mathcal{P}_{\mathcal{Y}}[v])_{V'V,L^2_\rho} &= \int_\Omega \Big(\mathcal{K}, \sum_{i=1}^R \langle v, Y_i\rangle_{L^2_\rho} Y_i\Big)_{V'V} \mathrm{d}\rho = \sum_{i=1}^R \int_\Omega \Big(\mathcal{K}, \langle v, Y_i\rangle_{L^2_\rho} Y_i\Big)_{V'V} \mathrm{d}\rho \\
&= \sum_{i=1}^R \int_\Omega \Big(\mathcal{K}\, Y_i, \langle v, Y_i\rangle_{L^2_\rho}\Big)_{V'V} \mathrm{d}\rho = \sum_{i=1}^R \Big(\langle \mathcal{K}, Y_i\rangle_{L^2_\rho}, \langle v, Y_i\rangle_{L^2_\rho}\Big)_{V'V} \\
&= \sum_{i=1}^R \int_\Omega \Big(\langle \mathcal{K}, Y_i\rangle_{L^2_\rho} Y_i, v\Big)_{V'V} \mathrm{d}\rho = (\mathcal{P}_{\mathcal{Y}}[\mathcal{K}],\, v)_{V'V,L^2_\rho},
\end{aligned}
$$

where in the forth step we applied Theorem 8.13 from [Leo17].

Now we proceed with proving (1.15)

$$
\begin{aligned}
(\mathcal{K},\, \Pi_{UY^\intercal}[v])_{V'V,L^2_\rho} &= (\mathcal{K},\, \mathcal{P}_{\mathcal{Y}}[v] + \mathcal{P}^\perp_{\mathcal{Y}}[\mathcal{P}_{\mathcal{U}}[v]])_{V'V,L^2_\rho} \\
&= (\mathcal{P}_{\mathcal{Y}}[\mathcal{K}],\, v)_{V'V,L^2_\rho} + (\mathcal{P}^\perp_{\mathcal{Y}}[\mathcal{K}],\, \mathcal{P}_{\mathcal{U}}[v])_{V'V,L^2_\rho} \\
&= \Big(\mathcal{P}_{\mathcal{Y}}[\mathcal{K}],\, v\Big)_{V'V,L^2_\rho} + \Big(\mathcal{P}^\perp_{\mathcal{Y}}[\mathcal{K}],\, (v, U)_H M^{-1} U^\intercal\Big)_{V'V,L^2_\rho} \\
&= \Big(\mathcal{P}_{\mathcal{Y}}[\mathcal{K}],\, v\Big)_{V'V,L^2_\rho} + \int_\Omega \Big(\mathcal{P}^\perp_{\mathcal{Y}}[\mathcal{K}], U M^{-1}\Big)_{V'V} \Big(U^\intercal, v\Big)_H \mathrm{d}\rho \\
&= (\mathcal{P}_{\mathcal{Y}}[\mathcal{K}],\, v)_{V'V,L^2_\rho} + \Big((\mathcal{P}^\perp_{\mathcal{Y}}[\mathcal{K}], U)_{V'V} M^{-1} U^\intercal,\, v\Big)_{V'V,L^2_\rho} \\
&= (\Pi_{UY^\intercal}[\mathcal{K}], v)_{V'V,L^2_\rho}.
\end{aligned}
$$

□

We are now in the position to state the first variational formulation of the DLR equations.

Lemma 1.3.4. *Let U, Y be the solution of the system* (1.9)–(1.10)*. Then the zero-mean part of the DLR solution $u^* = UY^\intercal$ satisfies*

$$
(\dot{u}^* - \Pi_{u^*}[\mathcal{F}^*(u)],\, v)_{V'V,L^2_\rho} = 0, \qquad \forall v \in L^2_\rho(\Omega; V). \tag{1.16}
$$

Proof. First, we multiply equation (1.9) by Y_j and take its weak formulation in L^2_ρ. Summing over j results in

$$
\Big(\dot{U} Y^\intercal - \mathbb{E}\Big[\mathcal{F}(u) Y\Big] Y^\intercal, v\, w\Big)_{V'V,L^2_\rho} = 0 \quad \forall v \in V,\, w \in L^2_\rho.
$$

Notice that $\mathbb{E}\Big[\mathcal{F}^*(u)Y\Big] = \mathbb{E}\Big[\mathcal{F}(u)Y\Big]$ since $Y \subset L^2_{\rho,0}$. Analogously, we multiply (1.10) by

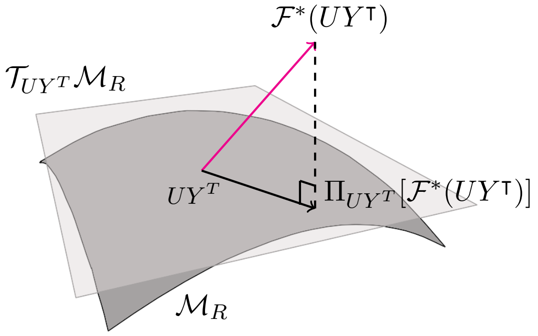

$\mathcal{F}^*(UY^\intercal)$
$\mathcal{T}_{UY^T}\mathcal{M}_R$
UY^T
$\Pi_{UY^T}[\mathcal{F}^*(UY^\intercal)]$
$\mathcal{M}_R$

Figure 1.1 – Illustration of the geometrical interpretation of dynamical low rank approximation.

U_j and take its weak formulation in V'

$$\left(U_j\dot{Y}_j - \sum_{i=1}^{R} U_j(M^{-1})_{j,i}\mathcal{P}_{\mathcal{Y}}^{\perp}\left[(\mathcal{F}^*(u), U_i)_{V'V}\right], v\,w\right)_{V'V,L^2_\rho} = 0 \qquad \forall v \in V,\ w \in L^2_\rho.$$

Summing over j, this leads to

$$\left(U\dot{Y}^\intercal - \mathcal{P}_{\mathcal{Y}}^{\perp}\left[(\mathcal{F}^*(u), U)_{V'V}M^{-1}U^\intercal\right], v\,w\right)_{V'V,L^2_\rho} = 0 \quad \forall v \in V,\ w \in L^2_\rho.$$

Summing the derived equations we obtain

$$\left(\frac{\mathrm{d}}{\mathrm{d}t}(UY^\intercal) - \Pi_{u^*}[\mathcal{F}^*(u)],\ z\right)_{V'V,L^2_\rho} = 0 \quad \forall z \in \operatorname{span}\{v\,w : v \in V,\ w \in L^2_\rho\}.$$

In particular, this holds for any z being a Bochner integrable simple function, the collection of which is dense in $L^2_\rho(\Omega; V)$ (see [Leo17, Th. 8.15]). ☐

A dynamical system (1.3) can be seen as a time-dependent vector field $\mathcal{F}$ that assigns the velocity $\mathcal{F}(v(t))$ at time t to each point v of the ambient space $L^2_\rho(\Omega; H)$. Any rank R approximation of the true solution forms a curve on the manifold $\mathcal{M}_R$ and consequently its velocity vector field must be everywhere tangent to the manifold. Lemma 1.3.4 describes the geometrical idea behind the DLR method. Having a rank R solution $u^* \in \mathcal{M}_R$, the DLRA projects the right-hand side $\mathcal{F}^*(u)$ onto the tangent space $\mathcal{T}_{u^*}\mathcal{M}_R$ at u^*, which assures that the solution of the resulting system remains in the manifold and is thus of rank R. See Figure 1.1 for an illustration.

We can finally state the variational formulation corresponding to the DLR equations (1.8)–(1.10).

Theorem 1.3.5 (DLR variational formulation)**.** *Let $\bar{u}, U, Y$ be the solution of the system* (1.8)–(1.10)*. Then the DLR solution $u = \bar{u} + UY^\intercal$ satisfies*

$$\Big(\dot{u},\, v\Big)_{V'V,L^2_\rho} = \Big(\mathcal{F}(u), v\Big)_{V'V,L^2_\rho}, \qquad \forall v = \bar{v} + v^*,\, \bar{v} \in V,\, v^* \in \mathcal{T}_{u^*}\mathcal{M}_R. \tag{1.17}$$

Proof. Based on Lemma 1.3.4 and Lemma 1.3.3 we can write

$$\begin{aligned}\Big(\dot{u}^*,\, v\Big)_{V'V,L^2_\rho} &- \Big(\Pi_{u^*}[\mathcal{F}^*(u)],\, v\Big)_{V'V,L^2_\rho} \\ &= \Big(\dot{u}^*,\, v\Big)_{V'V,L^2_\rho} - \Big(\mathcal{F}^*(u),\, \Pi_{u^*}[v]\Big)_{V'V,L^2_\rho} = 0, \qquad \forall v \in L^2_\rho(\Omega; V).\end{aligned}$$

Since $\Pi_{u^*}[v] = v$, $\forall v \in \mathcal{T}_{u^*}\mathcal{M}_R$, this results in

$$\Big(\dot{u}^* - \mathcal{F}^*(u),\, v\Big)_{V'V,L^2_\rho} = 0, \qquad \forall v \in \mathcal{T}_{u^*}\mathcal{M}_R,$$

which can be equivalently written as

$$(\dot{u}^* - \mathcal{F}^*(u),\, w + v\Big)_{V'V,L^2_\rho} = 0, \qquad \forall w \in V,\, \forall v \in \mathcal{T}_{u^*}\mathcal{M}_R, \tag{1.18}$$

exploiting the fact that $\Big(\dot{u}^* - \mathcal{F}^*(u),\, w\Big)_{V'V,L^2_\rho} = 0$, $\forall w \in V$. Likewise, equation (1.8) can be equivalently written as

$$\Big(\dot{\bar{u}} - \mathbb{E}[\mathcal{F}(u)],\, w + v\Big)_{V'V,L^2_\rho} = 0, \qquad \forall w \in V,\, \forall v \in \mathcal{T}_{u^*}\mathcal{M}_R, \tag{1.19}$$

exploiting the fact that $\Big(\dot{\bar{u}} - \mathbb{E}[\mathcal{F}(u)],\, v\Big)_{V'V,L^2_\rho} = 0$ as $\mathbb{E}[v] = 0\ \forall v \in \mathcal{T}_{u^*}\mathcal{M}_R$. Summing (1.18) and (1.19) leads to the sought equation (1.17). □

Recently, the existence and uniqueness of the dynamical low rank approximation for a class of random semi-linear evolutionary equations was established in [KN21] and for linear parabolic equations in two space dimensions with a symmetric operator in [BKU21].

2 Projector-splitting schemes and their variational formulation

In the previous chapter we saw that the DLR method can be motivated from two fairly distinct approaches: an algebraic approach, which considers a solution with a low-rank format (1.5) and derives equations for the deterministic and stochastic modes; and a geometric approach, which introduces the idea of projecting the operator on the tangent space expressed via a variational formulation (see (1.16)). In Theorem 1.3.5, we showed that these two approaches are in fact equivalent.

Similarly, to derive efficient discretization schemes, one can tackle the problem from two distinct viewpoints. An algebraic one, which involves applying Runge-Kutta methods of different orders (or other time-marching schemes) directly to the system of evolution equations for the deterministic and stochastic basis functions (1.8) – (1.10) (see e.g. [SL09; KL07a]), and a geometric one, proposed in [LO14; LOV15b], where starting from the variational formulation (1.16), the authors applied a splitting method to the projected right-hand side, resulting in a so-called projector-splitting integrator. A similar idea with a different splitting was applied in [BFFN21]. A different geometric approach was considered in [KV18], where the authors explored projected Runge-Kutta methods, where following a Runge-Kutta integration, the solution first leaves the manifold of R-rank functions by increasing its rank, and then is retracted back to the manifold.

In the presence of small singular values in the solution, the system of evolution equations becomes stiff as an inversion of a singular or nearly-singular matrix is required to solve it and applying standard explicit or implicit Runge-Kutta methods leads to instabilities (see [KLW16]). In this respect, the projector-splitting integrators (proposed in [LO14; LOV15b] and applied in e.g. [EL18; Ein19]) are very appealing. Extensions of the projector-splitting integrator to deal with symmetric matrices and rank adaptation are available in [CL19; CKL22; CL21]. In [KLW16], the authors showed that when applying the projector-splitting method for matrix differential equations one can bound the error independently of the size of the singular values, under the assumption that the projection error of $\mathcal{F}(u)$ to the tangent space of $\mathcal{M}_R$ at u is of size ε. A limitation of their theoretical result, as the authors point out, is that it requires a Lipschitz condition on $\mathcal{F}$ and is applicable to discretized PDEs only under a severe condition $\triangle tL \ll 1$ where $\triangle t$ is the step size and L is the Lipschitz constant, even for implicit schemes. Such condition is, however, not observed in numerical experiments. Analogous error bounds as in [KLW16] are obtained for the projected Runge-Kutta methods in [KV18], also for higher order schemes, under the same ε-approximability condition on $\mathcal{F}$ and under a restrictive parabolic condition on the time step.

In this work, we first recall the projector-splitting integrators and summarize some of their relevant properties. Then we propose a class of numerical schemes to approximate the evolution equations for the mean, the deterministic basis and the stochastic basis, which can be of explicit, semi-implicit or implicit type. Although not evident at first sight, we show that the explicit version of our scheme can be reinterpreted as a projector-splitting scheme, whenever the discrete solution is full-rank, and is thus equivalent to the scheme from [LO14; LOV15b]. Our derivation allows for an easy construction of implicit or semi-implicit versions. We show that the proposed discretization schemes can be written in a discrete variational formulation analogous to the continuous one (1.17), which allows for an easy geometric proof of the exactness property and becomes essential for stability and error analysis available in Chapters 3, 4 and 5.

We start by describing the discretization of the stochastic and physical variables. Afterwards, we follow by recalling the projector-splitting integrator from [LO14; LOV15b] and stating some of its beneficial properties in Section 2.1. In Section 2.2, we propose a new time-marching scheme that discretizes the DLR equations and prove that the discrete solution satisfies a discrete variational formulation. Finally, Section 2.3 is dedicated to showing a link between the new-proposed scheme and the projector-splitting integrator. The content provided in Sections 2.2–2.3 is original and taken essentially from [KNV21]. However, note that as opposed to [KNV21], where the authors considered only an elliptic linear operator $\mathcal{F}$, here the results are available for a general operator $\mathcal{F}$.

Stochastic discretization

We consider a discrete measure given by $\{\omega_k, \lambda_k\}_{k=1}^{\hat{N}}$, i.e. a set of sample points $\{\omega_k\}_{k=1}^{\hat{N}} \subset \Omega$ with $R < \hat{N} < \infty$ and a set of positive weights $\{\lambda_k\}_{k=1}^{\hat{N}}$, $\lambda_k > 0$, $\sum_{k=1}^{\hat{N}} \lambda_k = 1$, which approximates the probability measure ρ

$$\hat{\rho} \coloneqq \sum_{k=1}^{\hat{N}} \lambda_k \delta_{\omega_k} \approx \rho. \tag{2.1}$$

The discrete probability space $(\hat{\Omega} = \{\omega_k\}_{k=1}^{\hat{N}}, 2^{\hat{\Omega}}, \hat{\rho})$ will replace the original one $(\Omega, \mathcal{F}, \rho)$ in the discretization of the DLR equations. Notice, in particular, that a random variable $Z : \hat{\Omega} \mapsto \mathbb{R}$ measurable on $(\hat{\Omega}, 2^{\hat{\Omega}}, \hat{\rho})$ can be represented as a vector $z \in \mathbb{R}^{\hat{N}}$ with $z_k = Z(\omega_k)$, $k = 1, \dots, \hat{N}$. The sample points $\{\omega_k\}_{k=1}^{\hat{N}}$ can be taken as iid samples from

ρ (e.g. Monte Carlo samples) or chosen deterministically (e.g. deterministic quadrature points with positive quadrature weights). The mean value of a random variable Z with respect to the measure $\hat{\rho}$ is computed as

$$\mathbb{E}_{\hat{\rho}}[Z] = \sum_{k=1}^{\hat{N}} Z(\omega_k)\lambda_k.$$

We introduce also the semi-discrete scalar products $\langle \cdot, \cdot \rangle_{\star, L^2_{\hat{\rho}}}$ with $\star = V, H$ and their corresponding induced norms $\| \cdot \|_{\star, L^2_{\hat{\rho}}}$.

Space discretization

We consider a general finite-dimensional subspace $V_h \subset V$ whose dimension N_h is larger than R and is determined by the discretization parameter h. Eventually, we will perform a Galerkin projection of the DLR equations onto the subspace V_h. We further assume that an inverse inequality of the type

$$\|v\|_{V, L^2_{\hat{\rho}}} \leq \frac{C_{\mathrm{I}}}{h^p} \|v\|_{H, L^2_{\hat{\rho}}}, \qquad \forall v \in V_h \otimes L^2_{\hat{\rho}} \tag{2.2}$$

holds for some $p \in \mathbb{N}$ and $C_{\mathrm{I}} > 0$.

Time discretization

Concerning the time discretization, we divide the time interval into N equally spaced subintervals $0 = t_0 < t_1 < \cdots < t_N = T$ and denote the time step by $\triangle t := t_{n+1} - t_n$. We will consider various time discretization schemes specified in the rest of this chapter.

2.1 Projector-splitting integrator for DLRA

This section recalls some of the results presented in [LO14] while reformulating them to adapt to our setting and notation. The method was originally proposed to deal with time-dependent matrix evolution equations. To adhere to the finite-dimensional setting of [LO14], we consider the problem (1.3) discretized in space and random variables

$$\begin{aligned} &\left(\dot{u}_{\mathrm{true},h,\hat{\rho}}, v\right)_{V'V, L^2_{\hat{\rho}}} = \left(\mathcal{F}(u_{\mathrm{true},h,\hat{\rho}}), v\right)_{V'V, L^2_{\hat{\rho}}}, \quad \forall v \in V_h \otimes L^2_{\hat{\rho}}, \text{ a.e. } t \in (0,T] \\ &\qquad u_{\mathrm{true},h,\hat{\rho}}(0) = u^0_{h,\hat{\rho}}, \end{aligned} \tag{2.3}$$

with $u^0_{h,\hat{\rho}} \in V_h \otimes L^2_{\hat{\rho}}$ an approximation of $u^0 \in L^2_\rho(\Omega; V)$.

The DLR format of choice is the so called DDO (double dynamically orthogonal) format, i.e. both deterministic and stochastic modes are kept orthonormal. We will adapt the algorithm from [LO14] to approximate the DLR solution in the DDO format with an isolated mean, i.e.

$$u(t) = \bar{u}(t) + U(t)S(t)V(t)^{\intercal} \quad \in V_h \otimes L^2_{\hat{\rho}}. \tag{2.4}$$

with U orthonormal w.r.t. $\langle \cdot, \cdot \rangle_H$, V orthonormal in $L^2_{\hat{\rho}}$ and $S \in \mathbb{R}^{R \times R}$ of full rank.

We start by describing a continuous-in-time splitting algorithm, discretized in physical and stochastic space, and follow by proposing a fully discretized scheme.

In the following we focus on the evolution of the stochastic part of the DLR solution u^*. Let $u^{n,*}_{h,\hat{\rho}} \in V_h \otimes L^2_{\hat{\rho}}$ denote the stochastic part of the DLR solution at time t^n. Stemming from the variational formulation (1.16) for $u^* = USV^{\intercal}$:

$$\begin{aligned}(\dot{u}^*, v)_{V'V,L^2_{\rho}} &= (\Pi_{u^*}[\mathcal{F}^*(u)], v)_{V'V,L^2_{\rho}} \\ &= (\mathcal{P}_V[\mathcal{F}^*(u)] - \mathcal{P}_V[\mathcal{P}_U[\mathcal{F}^*(u)]] + \mathcal{P}_U[\mathcal{F}^*(u)], v)_{V'V,L^2_{\rho}}, \qquad \forall v \in L^2_{\rho}(\Omega; V),\end{aligned}$$

which includes three terms in the right-hand side, the first-order projector-splitting algorithm splits the evolution into three steps:

1. Solve the differential equation

$$\begin{aligned}(\dot{u}^*_I, v_h)_{V'V,L^2_{\hat{\rho}}} &= (\mathcal{P}_{V_I}[\mathcal{F}^*(u_I)], v_h)_{V'V,L^2_{\hat{\rho}}}, \qquad \forall v_h \in V_h \otimes L^2_{\hat{\rho}}, \quad t \in [t^n, t^{n+1}] \\ u_I(t^n) &= u^{n,*}_{h,\hat{\rho}},\end{aligned}$$

   where $u^*_I = U_I S_I V_I$.

2. Solve the differential equation

$$\begin{aligned}(\dot{u}^*_{II}, v_h)_{V'V,L^2_{\rho}} &= -(\mathcal{P}_{V_{II}}[\mathcal{P}_{U_{II}}[\mathcal{F}^*(u_{II})]], v_h)_{V'V,L^2_{\rho}}, \qquad \forall v_h \in V_h \otimes L^2_{\hat{\rho}}, \quad t \in [t^n, t^{n+1}] \\ u_{II}(t^n) &= u_I(t^{n+1}),\end{aligned}$$

   where $u^*_{II} = U_{II} S_{II} V_{II}$.

3. Solve the differential equation

$$\begin{aligned}(\dot{u}^*_{III}, v_h)_{V'V,L^2_{\rho}} &= (\mathcal{P}_{U_{III}}[\mathcal{F}^*(u_{III})], v_h)_{V'V,L^2_{\rho}}, \qquad \forall v_h \in V_h \otimes L^2_{\hat{\rho}}, \quad t \in [t^n, t^{n+1}] \\ u_{III}(t^n) &= u_{II}(t^{n+1}),\end{aligned}$$

   where $u^*_{III} = U_{III} S_{III} V_{III}$.

As a final step, we take $u^*_{III}(t^{n+1})$ as an approximation of $u^*_{\text{true},h,\hat{\rho}}(t^{n+1})$, the stochastic part of the solution of (2.3) at time t^{n+1}. By standard theory (see e.g. [HWL06]), this method is of first order accuracy w.r.t. $\triangle t$.

It turns out that such splitting combines well with the factorization (2.4). In the first substep, $K := US$ is updated, in the second substep S is updated and in the third substep $VS^\intercal$ is updated. The equations 1. – 3. result in the following algorithm.

Algorithm 2.1.1 (Continuous-in-time projector-splitting scheme)**.** Given the approximate solution $u_{h,\hat{\rho}}^{n,*} = U_0 S_0 V_0^\intercal$ at time t^n of the form (2.4) with

$$U_{0,j} \in V_h,\ V_{0,j} \in L^2_{\hat{\rho},0}, \quad \langle V_{0,i}, V_{0,j}\rangle_{L^2_{\hat{\rho}}} = \delta_{ij},\ \langle U_{0,i}, U_{0,j}\rangle_H = \delta_{ij}, \quad i,j = 1,\dots,R:$$

1. Solve R deterministic PDEs

   $$\begin{aligned}(\dot{K}(t),\, v_h)_{V'V} &= (\mathbb{E}_{\hat{\rho}}[\mathcal{F}^*(K(t)V_0^\intercal)V_0],\, v_h)_{V'V}, \qquad \forall v_h \in V_h, \quad t \in [t^n, t^{n+1}]\\ K(t^n) &= U_0 S_0.\end{aligned}$$

   Compute $U_1 \in V_h,\ \hat{S}_1 \in \mathbb{R}^{R\times R}$ such that $U_1\hat{S}_1 = K(t^{n+1})$ and U_1 is orthonormal in $\langle\cdot,\cdot\rangle_H$.

2. Solve the matrix differential equation (of size $R \times R$)

   $$\begin{aligned}\dot{S}(t) &= -\Big(\mathbb{E}_{\hat{\rho}}[\mathcal{F}^*(U_1 S(t) V_0^\intercal)V_0], U_1\Big)_{V'V}, \quad t \in [t^n, t^{n+1}]\\ S(t^n) &= \hat{S}_1.\end{aligned}$$

   Set $\tilde{S}_0 = S(t^{n+1})$.

3. Solve R stochastic differential equations set in $L^2_{\hat{\rho},0}$

   $$\begin{aligned}\dot{L}(t) &= (\mathcal{F}^*(U_1 L(t)^\intercal), U_1)_{V'V}\\ L(t^n) &= V_0 \tilde{S}_0^\intercal.\end{aligned}$$

   Compute $V_1 \in L^2_{\hat{\rho},0},\ S_1 \in \mathbb{R}^{R\times R}$ such that $V_1 S_1^\intercal = L(t^{n+1})$ in $L^2_{\hat{\rho},0}$ and V_1 is orthonormal in $\langle\cdot,\cdot\rangle_{L^2_{\hat{\rho}}}$.

The stochastic part of the new solution $\hat{u}_{h,\hat{\rho}}^{n+1,*}$ is then defined as

$$\hat{u}_{h,\hat{\rho}}^{n+1,*} = U_1 S_1 V_1^\intercal.$$

The algorithm was first proposed to deal with a DLR approximation of time-dependent matrices $A(t)$ which are known a-priori, avoiding the need to compute an SVD at every time step. In our setting, $A(t)$ stands for $u_{\mathrm{true},h,\hat{\rho}}(t)$ represented as a matrix of size $N_h \times \hat{N}$, where the first index corresponds to the spatial degrees of freedom (dofs) and the second index to the stochastic dofs. As an example, the algorithm can be applied in a scenario in which the operator $\mathcal{F}$ does not depend on $u_{\mathrm{true},h,\hat{\rho}}$, i.e. $\mathcal{F}(u_{\mathrm{true},h,\hat{\rho}}) = \mathcal{F}$. In this case the true solution $u_{\mathrm{true},h,\hat\rho}$ is known a-priori and satisfies

$$u_{\mathrm{true},h,\hat\rho}(t) = u_{\mathrm{true},h,\hat\rho}(0) + \int_0^t \mathcal{F}(s)\,\mathrm{d}s.$$

By $\triangle A$ let us denote the increment $\triangle A = u_{\mathrm{true},h,\hat\rho}(t^{n+1}) - u_{\mathrm{true},h,\hat\rho}(t^n) = \int_{t^n}^{t^{n+1}} \mathcal{F}(s)\,\mathrm{d}s$. Interestingly, each of the equations 1.–3. can be solved exactly in a trivial way and the resulting fully discrete scheme is summarized in the following 6–step algorithm, including the computation of the mean value.

Algorithm 2.1.2 (Discrete-in-time projector-splitting scheme)**.** Given the approximated solution $u^n_{h,\hat\rho} = \bar u^n + U_0 S_0 V_0^\intercal$ at time t^n of the form (2.4) with

$$\bar u^n, U_{0,j} \in V_h,\ V_{0,j} \in L^2_{\hat\rho,0}, \quad \langle V_{0,i}, V_{0,j}\rangle_{L^2_{\hat\rho}} = \delta_{ij},\ \langle U_{0,i}, U_{0,j}\rangle_H = \delta_{ij}, \quad i,j = 1,\dots,R:$$

1. Compute the mean value $\hat{\bar u}^{n+1}$ such that
$$\langle \hat{\bar u}^{n+1}, v_h\rangle_H = \langle \bar u^n, v_h\rangle_H + \Big(\mathbb{E}_{\hat\rho}[\triangle A], v_h\Big)_{V'V} \qquad \forall v_h \in V_h.$$
2. Solve for K_1 such that
$$\langle K_1, v_h\rangle_H = \langle U_0 S_0, v_h\rangle_H + \Big(\mathbb{E}_{\hat\rho}[\triangle A V_0],\, v_h\Big)_{V'V} \qquad \forall v_h \in V_h.$$
3. Compute $U_1 \in V_h,\ \hat S_1 \in \mathbb{R}^{R\times R}$ such that
$$U_1 \hat S_1 = K_1 \text{ and } U_1 \text{ is orthonormal in } \langle\cdot,\cdot\rangle_H.$$
4. Set
$$\tilde S_0 = \hat S_1 - \Big(\mathbb{E}_{\hat\rho}[\triangle A\, V_0],\, U_1\Big)_{V'V}.$$
5. Compute $L_1 \in L^2_{\hat\rho}$ such that
$$L_1 = V_0 \tilde S_0^\intercal + \Big(\triangle A,\, U_1\Big)_{V'V}.$$
6. Compute $V_1 \in L^2_{\hat\rho,0},\ S_1 \in \mathbb{R}^{R\times R}$ such that
$$V_1 S_1^\intercal = L_1 \text{ in } L^2_{\hat\rho,0} \text{ and } V_1 \text{ is orthonormal in } \langle\cdot,\cdot\rangle_{L^2_{\hat\rho}}.$$

The new solution $\hat u^{n+1}_{h,\hat\rho}$ is then defined as

$$\hat u^{n+1}_{h,\hat\rho} = \hat{\bar u}^{n+1} + U_1 S_1 V_1^\intercal.$$

Note that, since $\triangle A = u_{\mathrm{true},h,\hat\rho}(t^{n+1}) - u_{\mathrm{true},h,\hat\rho}(t^n) \in H$, the duality pairings $(\cdot,\cdot)_{V'V}$ on the right hand side of the equations in Algorithm 2.1.2 are equal to $\langle\cdot,\cdot\rangle_H$.

If the true solution $u_{\mathrm{true},h,\hat{\rho}}$ in not known a-priori and is defined as a solution of (1.3), the work [LO14] proposes to define the increment $\triangle A$ as

$$\triangle A = \triangle t \mathcal{F}(u^n_{h,\hat{\rho}}),$$

which is a fully discretized scheme and resembles the explicit Euler method. All the scalar products $\langle \cdot, \cdot \rangle_H$ involving $\triangle A$ are consequently replaced by the dual pairing $(\cdot,\cdot)_{V'V}$.

The projector-splitting algorithm has many favourable properties. First to mention, it reproduces R-rank solutions exactly.

Theorem 2.1.3 (Exactness property)**.** *Let $u_{\mathrm{true},h,\hat{\rho}}(t) \in V_h \otimes L^2_{\hat{\rho}}$, the solution of* (2.3)*, be of rank R for $t^n \leq t \leq t^{n+1}$, so that $u_{\mathrm{true},h,\hat{\rho}}(t)$ has a factorization* (2.4)*, i.e.*

$$u_{\mathrm{true},h,\hat{\rho}}(t) = \bar{u}(t) + U(t)S(t)V(t)^\intercal.$$

Moreover, assume that the $R \times R$ matrix $\mathbb{E}[V(t^{n+1})^\intercal V(t^n)]$ is invertible. With $u^n_{h,\hat{\rho}} = u_{\mathrm{true},h,\hat{\rho}}(t^n)$, the Algorithm 2.1.2 is exact: $\hat{u}^{n+1}_{h,\hat{\rho}} = u_{\mathrm{true},h,\hat{\rho}}(t^{n+1})$.

Proof. There are multiple available proofs, see e.g. [LO14; Wal18]. In Section 2.3 we provide a new simple geometric proof. ☐

The second important property is the robustness of the algorithm to the presence of small singular values of the solution or its approximation. The DLR equations (1.8) – (1.10) involve an inversion of the matrix M, which for small singular values of the approximation becomes nearly-singular. Applying standard explicit or implicit Runge-Kutta methods leads to instabilities (see [KLW16]). Moreover, the local Lipschitz constant of the tangent space projection Π_{u^*} in (1.16) is proportional to the inverse of the smallest non-zero singular value of u (see [KL07a, Lemma 4.2]). Having small singular values cannot be easily avoided in practical applications, since the smallest singular value retained in the approximation is not expected to be much larger than the largest discarded singular value of the solution, which needs to be small to obtain good accuracy. The following theorem ensures us that under certain conditions on the operator $\mathcal{F}$, the continuous-in-time projector-splitting integrator provides an approximation whose error can be bounded independently on the singular values. Again, to adhere to the finite-dimensional setting of [KLW16], we consider the discrete problem (2.3). In addition, we assume that $\mathcal{F}(u) \in L^2_{\hat{\rho}}(\hat{\Omega}; H)$, $\forall u \in L^2_{\hat{\rho}}(\hat{\Omega}; H)$.

Theorem 2.1.4. *Let $u_{\mathrm{true},h,\hat{\rho}}$ be the solution of the problem* (2.3)*. Assume that the following conditions hold*

1. *$\mathcal{F}$ is Lipschitz-continuous and bounded*

$$\|\mathcal{F}(u)-\mathcal{F}(v)\|_{H,L^2_{\hat\rho}} \le L\|u-v\|_{H,L^2_{\hat\rho}}, \tag{2.5}$$

$$\|\mathcal{F}(u)\|_{H,L^2_{\hat\rho}} \le B, \quad \forall u,v \in L^2_{\hat\rho}(\Omega;H), \quad t\in[0,T]. \tag{2.6}$$

2. *The non-tangential part of $\mathcal{F}(u)$ is ε-small*

$$\|\Pi_u^\perp[\mathcal{F}(u)]\|_{H,L^2_{\hat\rho}} \le \varepsilon, \quad \forall u\in\mathcal{M}_R^{h,\hat\rho} \text{ in a neighbourhood of } u^*_{\mathrm{true},h,\hat\rho},\ t\in[0,T]. \tag{2.7}$$

3. *The error in the initial value is δ-small*

$$\|\hat u^0_{h,\hat\rho} - u_{\mathrm{true},h,\hat\rho}(0)\|_{H,L^2_{\hat\rho}} \le \delta$$

Let $\hat u^n_{h,\hat\rho}$ denote the rank-R approximation to $u_{\mathrm{true},h,\hat\rho}(t^n)$ obtained after n steps of the continuous projector-splitting Algorithm 2.1.1. Then, the error satisfies for all n with $(\triangle t)n \le T$

$$\|\hat u^n_{h,\hat\rho} - u_{\mathrm{true},h,\hat\rho}(t^n)\|_{H,L^2_{\hat\rho}} \le c_0\delta + c_1\varepsilon + c_2\triangle t,$$

where the constants c_1, c_2, c_3 only depend on L, B and T. In particular, the constants are independent of singular values of the exact or approximate solution.

It is further shown in [KLW16, Section 2.6.3] that an inexact solution of the matrix differential equations in the projector-splitting integrator leads to an additional error that is bounded in terms of the local errors in the inexact substeps, again with constants that do not depend on small singular values. In Chapter 4, we derive an a-priori error estimate for a DLR solution obtained by a scheme proposed in the following section. The governing equation (1.3) is set in an infinite-dimensional setting, which means that the a-priori estimate includes an error contribution w.r.t. the spatial and stochastic discretization as well. Concerning the time discretization and rank truncation, the result is analogous to Theorem 2.1.4, however, we manage to ease the conditions (2.5). In particular, we allow for an operator $\mathcal{F}$, which is not uniformly bounded and which satisfies only a one-sided Lipschitz condition

$$\langle \mathcal{F}(u)-\mathcal{F}(v), u-v\rangle_{H,L^2_{\hat\rho}} \le l\|u-v\|^2_{H,L^2_{\hat\rho}}.$$

As pointed out by authors in [KLW16, Sec. 2.6.2], the dependence on L could not have been avoided in their result so our analysis presents an improvement.

2.2 A staggered time-marching scheme for DLR equations

In this section, we define a new fully discrete DLR solution, which, on the contrary to the previous section, is derived by discretizing the DLR equations (1.8)–(1.10). We propose a staggered time-marching scheme that decouples the update of the deterministic and stochastic modes. Afterwards, we state and prove a variational formulation of the discretized problem. Finally, we will show that the proposed scheme can be formulated as a projector-splitting scheme for the Dual DO formulation and comment on its connection to the projector-splitting scheme from the previous section.

The DLR solution $u = \bar{u} + UY^{\intercal}$ appears in the right hand side of the system of equations (1.8)–(1.10), both in the operator $\mathcal{F}$ and in the projector operator onto the tangent space to the manifold. We will treat these two terms differently. Concerning the projection operator, we adopt a staggered strategy, where, given the approximate solution $u^n = \bar{u}^n + U^n {Y^n}^{\intercal}$, we first update the mean $\bar{u}^{n+1}$, then we update the deterministic basis U^{n+1} projecting on the subspace $\mathcal{Y}^n = \operatorname{span}\{Y^n\}$; finally, we update the stochastic basis Y^{n+1} projecting on the orthogonal complement of $\mathcal{Y}^n$ and on the updated subspace $\mathcal{U}^{n+1} = \operatorname{span}\{U^{n+1}\}$. Concerning the operator $\mathcal{F}$, we will discuss hereafter different discretization choices leading to explicit, semi-implicit or fully implicit algorithms.

2.2.1 Staggered time-marching scheme

We give in the next algorithm the general form of the discretization schemes that we consider in this work.

Algorithm 2.2.1. Given the approximate solution $u^n_{h,\hat{\rho}} = \bar{u}^n + \sum_{i=1}^R U^n_j Y^n_j$ at time t_n with

$$\begin{gathered} \bar{u}^n, U^n_j \in V_h, \quad Y^n_j \in L^2_{\hat{\rho}}, \quad j = 1, \ldots, R, \\ \langle Y^n_i, Y^n_j \rangle_{L^2_{\hat{\rho}}} = \delta_{ij}, \quad \mathbb{E}_{\hat{\rho}}[Y^n_j] = 0, \quad \forall\, 1 \le i, j \le R: \end{gathered}$$

1. Compute the mean value $\bar{u}^{n+1}$ such that

$$\left\langle \frac{\bar{u}^{n+1} - \bar{u}^n}{\triangle t}, v_h \right\rangle_H = \Big(\mathbb{E}_{\hat{\rho}}[\mathcal{F}(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}})], v_h \Big)_{V'V} \forall v_h \in V_h. \tag{2.8}$$

2. Compute the deterministic basis $\tilde{U}^{n+1}_j$ for $j = 1, \ldots, R$

$$\left\langle \frac{\tilde{U}^{n+1}_j - U^n_j}{\triangle t}, v_h \right\rangle_H = \Big(\mathbb{E}_{\hat{\rho}}[\mathcal{F}(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}}) Y^n_j], v_h \Big)_{V'V} \forall v_h \in V_h. \tag{2.9}$$

3. Compute the stochastic basis $\{\tilde{Y}_j^{n+1}\}_{j=1}^R$ such that

$$\frac{\tilde{Y}^{n+1} - Y^n}{\triangle t}\tilde{M}^{n+1} = \mathcal{P}^{\perp}_{\hat{\rho},\mathcal{Y}^n}\left[\left(\mathcal{F}^*(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}}), \tilde{U}^{n+1}\right)_{V'V}\right]. \tag{2.10}$$

   where $\tilde{M}^{n+1} = \langle \tilde{U}^{n+1\intercal}, \tilde{U}^{n+1}\rangle_H$, $\mathcal{P}^{\perp}_{\hat{\rho},\mathcal{Y}^n}[\cdot]$ is the analog of the projector defined in (1.11) but in the discrete space $L^2_{\hat{\rho}}$.

4. Reorthonormalize the stochastic basis: find (U^{n+1}, Y^{n+1}) s.t.

$$\sum_{j=1}^{R} Y_j^{n+1} U_j^{n+1} = \sum_{j=1}^{R} \tilde{Y}_j^{n+1} \tilde{U}_j^{n+1}, \qquad \langle Y^{n+1\intercal}, Y^{n+1}\rangle_{L^2_{\hat{\rho}}} = \mathrm{Id}. \tag{2.11}$$

5. Form the approximate solution at time step t_{n+1} as

$$u^{n+1}_{h,\hat{\rho}} = \bar{u}^{n+1} + \sum_{i=1}^{R} U_j^{n+1} Y_j^{n+1}. \tag{2.12}$$

The expression $\mathcal{F}(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}})$ stands for an unspecified time integration of the operator $\mathcal{F}(u(t))$, $t \in [t_n, t_{n+1}]$ (three choices will be specified here after) and v^* denotes the 0-mean part of a random variable $v \in L^2_{\hat{\rho}}$ with respect to the discrete measure $\hat{\rho}$, i.e. $v^* = v - \mathbb{E}_{\hat{\rho}}[v]$.

The newly computed solution $u^{n+1}_{h,\hat{\rho}}$ belongs to the tensor product space $V_h \otimes L^2_{\hat{\rho}}$, since we have $\bar{u}^{n+1}, U_j^{n+1} \in V_h$ and $Y_j \in L^2_{\hat{\rho}}$, $1 \leq j \leq R$. Note that equation (2.10) is set in $L^2_{\hat{\rho}}$. Since $L^2_{\hat{\rho}}$ is a finite dimensional vector space isomorphic to $\mathbb{R}^{\hat{N}}$, equation (2.10) can be rewritten as a deterministic linear system of $R \times \hat{N}$ equations with $R \times \hat{N}$ unknowns. This system can be decoupled into a linear system of size $R \times R$ for each collocation point. If the deterministic modes $\tilde{U}^{n+1}$ are linearly independent, the system matrix is invertible. Otherwise we interpret (2.10) in a minimal-norm least squares sense, choosing a solution $\tilde{Y}^{n+1}$, if it exists, that minimizes the norm $\|\tilde{Y}^{n+1} - Y^n\|_{L^2_{\hat{\rho}}}$. This is discussed in more details in Section 2.2.3.

The following lemma summarizes some properties satisfied by the proposed scheme (2.8)–(2.10).

Lemma 2.2.2 (Discretization properties). *Assuming that a solution* $(\tilde{Y}^{n+1}, \tilde{U}^{n+1}, \bar{u}^{n+1})$ *of* (2.8)–(2.10) *exists, the following properties hold:*

1. *Discrete DO condition:*

$$\left\langle \frac{\tilde{Y}_i^{n+1} - Y_i^n}{\triangle t}, Y_j^n \right\rangle_{L^2_{\hat{\rho}}} = 0, \quad \forall 1 \leq i, j \leq R \tag{2.13}$$

2. $\mathbb{E}_{\hat\rho}[\tilde Y^{n+1}] = 0$

3. $\langle \tilde Y^{n+1^\intercal}, Y^n\rangle_{L^2_{\hat\rho}} = \mathrm{Id}$

Proof.

1. In the following proof we assume that the matrix $\tilde M^{n+1} = \langle \tilde U^{n+1^\intercal}, \tilde U^{n+1}\rangle_H$ is full rank. For the rank-deficient case we refer the reader to the proof of Lemma 2.2.9. Let us multiply equation (2.10) by $Y^{n^\intercal}$ from the left and take the $L^2_{\hat\rho}$-scalar product. Since the second term involves $\mathcal{P}^\perp_{\hat\rho,\mathcal{Y}^n}$, the scalar product of Y^n with the second term vanishes which, under the assumption that $\tilde M^{n+1}$ is full rank, gives us the discrete DO condition
$$\left\langle Y^{n^\intercal}, \frac{\tilde Y^{n+1} - Y^n}{\triangle t}\right\rangle_{L^2_{\hat\rho}} = 0.$$
2. This is a consequence of the fact that we have $\mathbb{E}_{\hat\rho}[Y^n] = 0$ and $\mathbb{E}_{\hat\rho}\left[\left(\mathcal{F}^*(u^n, u^{n+1}), \tilde U^{n+1}\right)_{V'V}\right] = 0$.
3. This is immediate from the discrete DO property and $\langle Y^{n^\intercal}, Y^n\rangle_{L^2_{\hat\rho}} = \mathrm{Id}$.

□

To complete the discretization scheme (2.8)–(2.10) we need to specify the term $\mathcal{F}(u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho})$. The DLR system stated in (1.8)–(1.10) is coupled. Therefore, an important feature we would like to attain is to decouple the equations for the mean value, the deterministic and the stochastic modes as much as possible. We describe hereafter 3 strategies for the discretization of the operator evaluation term $\mathcal{F}(u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho})$.

Explicit Euler scheme

The explicit Euler scheme performs the discretization

$$\mathcal{F}(u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho}) = \mathcal{F}(u^n_{h,\hat\rho}).$$

It decouples the system (2.8)–(2.10) since, for the computation of the new modes, we require only the knowledge of the already-computed modes. The equations for the stochastic modes $\{\tilde Y^{n+1}_j\}_{j=1}^R$ are coupled together through the matrix $\tilde M^{n+1} = \langle \tilde U^{n+1^\intercal}, \tilde U^{n+1}\rangle_H \in \mathbb{R}^{R\times R}$ but are otherwise decoupled between collocation points (i.e. $\hat N$ linear systems of size R have to be solved).

Implicit Euler scheme

The implicit Euler scheme performs the discretization

$$\mathcal{F}(u_{h,\hat{\rho}}^{n}, u_{h,\hat{\rho}}^{n+1}) = \mathcal{F}(u_{h,\hat{\rho}}^{n+1}).$$

This method couples the system (2.8)–(2.10) in a non-trivial way, which is why we do not focus on this method in our numerical results. We mention it in the stability estimates section (Section 3.3) for its interesting stability properties.

Semi-implicit scheme

The following technique is proposed for a more specified operator $\mathcal{F}$ that generalizes a random parabolic equation. Assume that our operator $\mathcal{F}$ can be decomposed as

$$\mathcal{F}(u) = f - (\mathcal{L}_{\text{det}}(u) + \mathcal{L}_{\text{stoch}}(u)),$$

with both $\mathcal{L}_{\text{det}}$ and $\mathcal{L}_{\text{stoch}}$ linear w.r.t u. The operator $\mathcal{L}_{\text{det}} : V \to V'$ is a deterministic operator such that it induces a bounded and coercive bilinear form $\langle \cdot, \cdot \rangle_{\mathcal{L}_{\text{det}}}$ on V

$$\langle u, v \rangle_{\mathcal{L}_{\text{det}}} := (\mathcal{L}_{\text{det}}(u), v)_{V'V}, \qquad u, v \in V \tag{2.14}$$

and that its action on a function $v = v_1 v_2$ with $v_1 \in V$, $v_2 \in L^2_\rho$ is defined as

$$\mathcal{L}_{\text{det}}(v) = \mathcal{L}_{\text{det}}(v_1) v_2.$$

Then, $\mathcal{L}_{\text{det}}$ is also a linear operator $\mathcal{L}_{\text{det}} : L^2_\rho(\Omega; V) \to L^2_\rho(\Omega; V')$ (as well as $\mathcal{L}_{\text{det}} : L^2_{\hat{\rho}}(\hat{\Omega}; V) \mapsto L^2_{\hat{\rho}}(\hat{\Omega}; V')$) and induces a bounded coercive bilinear form on $L^2_\rho(\Omega; V)$

$$\langle u, v \rangle_{\mathcal{L}_{\text{det}}, \rho} = \int_\Omega (\mathcal{L}_{\text{det}}(u), v)_{V'V} \, \mathrm{d}\rho.$$

We propose a semi-implicit time integration of the operator evaluation term

$$\mathcal{F}(u_{h,\hat{\rho}}^{n}, u_{h,\hat{\rho}}^{n+1}) = f^{n,n+1} - (\mathcal{L}_{\text{det}}(u_{h,\hat{\rho}}^{n+1}) + \mathcal{L}_{\text{stoch}}(u_{h,\hat{\rho}}^{n})) \tag{2.15}$$

where for $f^{n,n+1}$ we can either take $f^{n,n+1} = f(t_{n+1})$ or $f^{n,n+1} = f(t_n)$ or any convex combination of both. The resulting scheme is detailed in the next lemma.

Lemma 2.2.3. *The semi-implicit integration scheme* (2.15) *combined with the general steps* (2.8)–(2.10) *is equivalent to the following set of equations*

$$\begin{aligned} &\langle \bar{u}^{n+1}, v_h \rangle_H + \triangle t \langle \bar{u}^{n+1}, v_h \rangle_{\mathcal{L}_{\text{det}}} \\ &\quad = \langle \bar{u}^{n}, v_h \rangle_H - \triangle t (\mathbb{E}_{\hat{\rho}}[\mathcal{L}_{\text{stoch}}(u_{h,\hat{\rho}}^{n}) - f^{n,n+1}], v_h)_{V'V} \qquad \forall v_h \in V_h \end{aligned} \tag{2.16}$$

$$\begin{aligned} &\langle \tilde{U}_j^{n+1}, v_h \rangle_H + \triangle t \langle \tilde{U}_j^{n+1}, v_h \rangle_{\mathcal{L}_{\text{det}}} \\ &\quad = \langle \tilde{U}_j^{n}, v_h \rangle_H - \triangle t (\mathbb{E}_{\hat{\rho}}[(\mathcal{L}_{\text{stoch}}(u_{h,\hat{\rho}}^{n}) - f^{n,n+1}) Y_j^n], v_h)_{V'V} \quad \forall v_h \in V_h \end{aligned} \tag{2.17}$$

$$\begin{aligned}&\left(\tilde{Y}^{n+1} - Y^n\right)\left(\tilde{M}^{n+1} + \triangle t\langle \tilde{U}^{n+1^\intercal}, \tilde{U}^{n+1}\rangle_{\mathcal{L}_{\mathrm{det}}}\right)\\ &\qquad = -\triangle t\mathcal{P}^{\perp}_{\hat{\rho},\mathcal{Y}^n}[(\mathcal{L}^*_{\mathrm{stoch}}(u^n_{h,\hat{\rho}}) - f^{n,n+1^*}, \tilde{U}^{n+1})_{V'V}] \qquad \text{in } L^2_{\hat{\rho}}. \qquad (2.18)\end{aligned}$$

Proof. The equation for the mean (2.8) using the semi-implicit scheme (2.15) can be written as

$$\begin{aligned}\left\langle \frac{\bar{u}^{n+1} - \bar{u}^n}{\triangle t}, v_h\right\rangle_H &+ \underbrace{\left(\mathbb{E}_{\hat{\rho}}[\mathcal{L}_{\mathrm{det}}(\bar{u}^{n+1})], v_h\right)_{V'V}}_{T_1} + \underbrace{\left(\mathbb{E}_{\hat{\rho}}[\mathcal{L}_{\mathrm{det}}(\tilde{U}^{n+1}Y^{n+1^\intercal})], v_h\right)_{V'V}}_{T_2}\\ &= -\left(\mathbb{E}_{\hat{\rho}}[\mathcal{L}_{\mathrm{stoch}}(u^n_{h,\hat{\rho}}) - f^{n,n+1}], v_h\right)_{V'V}.\end{aligned}$$

Noticing that

$$\begin{aligned}T_1 &= \left(\mathcal{L}_{\mathrm{det}}(\bar{u}^{n+1}), v_h\right)_{V'V} = \langle \bar{u}^{n+1}, v_h\rangle_{\mathcal{L}_{\mathrm{det}}}\\ T_2 &= \left(\mathcal{L}_{\mathrm{det}}(\tilde{U}^{n+1})\mathbb{E}_{\hat{\rho}}[Y^{n+1^\intercal}], v_h\right)_{V'V} = 0\end{aligned}$$

gives us equation (2.16). Concerning the equation for the deterministic modes we derive

$$\begin{aligned}&\left\langle \frac{\tilde{U}_j^{n+1} - U_j^n}{\triangle t}, v_h\right\rangle_H + \underbrace{\left(\mathbb{E}_{\hat{\rho}}[\mathcal{L}_{\mathrm{det}}(\bar{u}^{n+1})Y_j^n], v_h\right)_{V'V}}_{T_3}\\ &+ \underbrace{\left(\mathbb{E}_{\hat{\rho}}[\mathcal{L}_{\mathrm{det}}(\tilde{U}^{n+1}\tilde{Y}^{n+1^\intercal})Y_j^n], v_h\right)_{V'V}}_{T_4} = -\left(\mathbb{E}_{\hat{\rho}}[(\mathcal{L}_{\mathrm{stoch}}(u^n_{h,\hat{\rho}}) - f^{n,n+1})Y_j^n], v_h\right)_{V'V}.\end{aligned}$$

The term T_3 vanishes since $\mathbb{E}_{\hat{\rho}}[Y^n] = 0$ and the term T_4 can be further expressed as

$$\begin{aligned}T_4 &= \left(\mathcal{L}_{\mathrm{det}}(\tilde{U}^{n+1})\mathbb{E}_{\hat{\rho}}[\tilde{Y}^{n+1^\intercal}Y_j^n], v_h\right)_{V'V} = \left(\mathcal{L}_{\mathrm{det}}(\tilde{U}_j^{n+1}), v_h\right)_{V'V}\\ &= \langle \tilde{U}_j^{n+1}, v_h\rangle_{\mathcal{L}_{\mathrm{det}}},\end{aligned}$$

where we used the discrete DO condition (2.13). Finally, the stochastic equation (2.10) can be written as

$$\begin{aligned}&\left(\frac{\tilde{Y}^{n+1} - Y^n}{\triangle t}\right)(\tilde{M}^{n+1}) + \underbrace{\mathcal{P}^{\perp}_{\hat{\rho},\mathcal{Y}^n}\left[\left(\mathcal{L}^*_{\mathrm{det}}(\bar{u}^{n+1}), \tilde{U}^{n+1}\right)_{V'V}\right]}_{T_5}\\ &\underbrace{+\mathcal{P}^{\perp}_{\hat{\rho},\mathcal{Y}^n}\left[\left(\mathcal{L}^*_{\mathrm{det}}(\tilde{U}^{n+1}\tilde{Y}^{n+1^\intercal}), \tilde{U}^{n+1}\right)_{V'V}\right]}_{T_6} = -\mathcal{P}^{\perp}_{\hat{\rho},\mathcal{Y}^n}\left[\left(\mathcal{L}^*_{\mathrm{stoch}}(u^n_{h,\hat{\rho}}) - f^{n,n+1^*}, \tilde{U}^{n+1}\right)_{V'V}\right].\end{aligned}$$

The term T_5 vanishes since $\mathcal{L}^*_{\det}(\bar{u}^{n+1}) = 0$. As for T_6, we derive

$$\begin{aligned} T_6 &= \Big(\mathcal{L}_{\det}(\tilde{U}^{n+1})\tilde{Y}^{n+1^\intercal}, \tilde{U}^{n+1}\Big)_{V'V} - \Big(\mathcal{L}_{\det}(\tilde{U}^{n+1})\underbrace{\mathbb{E}_{\hat{\rho}}[\tilde{Y}^{n+1^\intercal}Y^n]}_{\mathrm{Id}}Y^{n^\intercal}, \tilde{U}^{n+1}\Big)_{V'V} \\ &\qquad - \mathcal{P}^{\perp}_{\hat{\rho},\mathcal{Y}^n}\Big[\Big(\mathcal{L}_{\det}(\tilde{U}^{n+1})\underbrace{\mathbb{E}_{\hat{\rho}}[\tilde{Y}^{n+1^\intercal}]}_{=0}, \tilde{U}^{n+1}\Big)_{V'V}\Big] \\ &= \langle \tilde{U}^{n+1}, \tilde{U}^{n+1}\rangle_{\mathcal{L}_{\det}}(\tilde{Y}^{n+1^\intercal} - Y^{n^\intercal}) \end{aligned}$$

which leads us to the sought equation (2.18).

□

We see from (2.16)–(2.18) that, similarly to the explicit Euler scheme, the equations for the mean, deterministic modes and stochastic modes are decoupled. If the spatial discretization of the PDEs (2.16) and (2.17) is performed by the Galerkin approximation, the final linear system involves the inversion of the matrix

$$A_{ij} = \langle \varphi_j, \varphi_i\rangle_H + \triangle t\langle \varphi_j, \varphi_i\rangle_{\mathcal{L}_{\det}},$$

where $\{\varphi_i\}$ is the basis of V_h in which the solution is represented. Both the mass matrix $\langle \varphi_j, \varphi_i\rangle_H$ and the stiffness matrix $\langle \varphi_j, \varphi_i\rangle_{\mathcal{L}_{\det}}$ are positive definite and do not evolve with time, so that an LU factorization can be computed once and for all at the beginning of the simulation. Concerning the stochastic equation (2.18), we need to solve a linear system with the matrix $\tilde{M}^{n+1} + \triangle t\langle \tilde{U}^{n+1^\intercal}, \tilde{U}^{n+1}\rangle_{\mathcal{L}_{\det}}$ for each collocation point ω_k, where $\tilde{M}^{n+1} = \langle \tilde{U}^{n+1^\intercal}, \tilde{U}^{n+1}\rangle_H$. This is in contrast to the explicit Euler method, where the system involves only the matrix $\tilde{M}^{n+1}$. The matrix $\tilde{M}^{n+1} + \triangle t\langle \tilde{U}^{n+1^\intercal}, \tilde{U}^{n+1}\rangle_{\mathcal{L}_{\det}}$ is positive definite with the smallest singular value bigger than that of $\tilde{M}^{n+1}$. Notice, however, that if $\tilde{M}^{n+1}$ is rank deficient, also the matrix $\tilde{M}^{n+1} + \triangle t\langle \tilde{U}^{n+1^\intercal}, \tilde{U}^{n+1}\rangle_{\mathcal{L}_{\det}}$ will be so. The computational complexity of the semi-implicit scheme w.r.t. the explicit scheme depends on the operator $\mathcal{L}_{\det}$.

Note that there exists a unique discrete DLR solution for the explicit and semi-implicit version of Algorithm 2.2.1 also in the rank-deficient case (see Lemma 2.2.10 below). The existence of solutions for the implicit version remains still an open question.

2.2.2 Discrete variational formulation for the full-rank case

This subsection will closely follow the structure of the Subsection 1.3. We will introduce analogous geometrical concepts for the discrete setting, i.e. manifold of R-rank functions, tangent space and orthogonal projection, and will show in Theorem 2.2.7 that the scheme from Algorithm 2.2.1 can be written in a (discrete) variational formulation, assuming that the matrix $\tilde{M}^{n+1}$ stays full-rank.

Definition 2.2.1 (Discrete manifold of R-rank functions)**.** By $\mathcal{M}_R^{h,\hat{\rho}} \subset V_h \otimes L^2_{\hat{\rho},0}$ we denote the manifold of all rank R functions with zero mean that belong to the (possibly finite dimensional) space $V_h \otimes L^2_{\hat{\rho}}$, namely

$$\begin{aligned}\mathcal{M}_R^{h,\hat{\rho}} = \Bigg\{ & v^* \in V_h \otimes L^2_{\hat{\rho},0} \,|\, v^* = \sum_{i=1}^{R} U_i Y_i, \quad \{Y_i\}_{i=1}^R \subset L^2_{\hat{\rho},0} \\ & \langle Y_i, Y_j\rangle_{L^2_{\hat{\rho}}} = \delta_{ij},\ \forall 1 \le i,j \le R,\ \{U_i\}_{i=1}^R \subset V_h \text{ linearly independent}\Bigg\}.\end{aligned} \tag{2.19}$$

Proposition 2.2.4 (Discrete tangent space at $UY^\intercal$)**.** The tangent space $\mathcal{T}_{UY^\intercal}\mathcal{M}_R^{h,\hat{\rho}}$ at a point $UY^\intercal \in \mathcal{M}_R^{h,\hat{\rho}}$ is formed as

$$\begin{aligned}\mathcal{T}_{UY^\intercal}\mathcal{M}_R^{h,\hat{\rho}} = \Bigg\{ & \delta v \in V_h \otimes L^2_{\hat{\rho},0} \,|\, \delta v = \sum_{i=1}^{R} U_i \delta Y_i + \delta U_i Y_i, \\ & \delta U_j \in V_h,\ \delta Y_i \in L^2_{\hat{\rho},0},\ \langle \delta Y_i, Y_j\rangle_{L^2_{\hat{\rho}}} = 0,\ \forall 1 \le i,j \le R\Bigg\}.\end{aligned} \tag{2.20}$$

The projection $\Pi^{h,\hat{\rho}}_{UY^\intercal}$ is defined in the discrete space $V_h \otimes L^2_{\hat{\rho}}$ analogously to its continuous version (1.14). It holds

$$\Pi^{h,\hat{\rho}}_{UY^\intercal} : V_h \otimes L^2_{\hat{\rho}} \to \mathcal{T}_{UY^\intercal}\mathcal{M}_R^{h,\hat{\rho}} \subset V_h \otimes L^2_{\hat{\rho}}, \qquad \forall UY^\intercal \in \mathcal{M}_R^{h,\hat{\rho}}.$$

A discrete analogue of Lemma 1.3.3 holds, i.e.

$$(\mathcal{K}, \Pi^{h,\hat{\rho}}_{UY^\intercal}[v])_{V'V,L^2_{\hat{\rho}}} = (\Pi^{h,\hat{\rho}}_{UY^\intercal}[\mathcal{K}], v)_{V'V,L^2_{\hat{\rho}}}, \qquad \forall v \in V_h \otimes L^2_{\hat{\rho}},\ \mathcal{K} \in V_h' \otimes L^2_{\hat{\rho}}. \tag{2.21}$$

The solution of the proposed numerical scheme (2.8)–(2.11) satisfies a discrete variational formulation analogous to the variational formulation (1.17). To show this, we first present a technical lemma which will be important in deriving the variational formulation.

Lemma 2.2.5. *Let* $u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}}$ *be the discrete DLR solution at* t_n, t_{n+1}*, respectively, from the scheme in Algorithm 2.2.1. Then the zero-mean parts* $u^{n,*}_{h,\hat{\rho}}, u^{n+1,*}_{h,\hat{\rho}}$ *satisfy*

1. $u^{n^*}_{h,\hat{\rho}} \in \mathcal{T}_{\tilde{U}^{n+1}Y^{n\intercal}}\mathcal{M}_R^{h,\hat{\rho}}$,
2. $u^{n+1,*}_{h,\hat{\rho}} \in \mathcal{T}_{\tilde{U}^{n+1}Y^{n\intercal}}\mathcal{M}_R^{h,\hat{\rho}}$.

Proof.

1. The solution $u^{n,*}_{h,\hat{\rho}}$ can be written as

$$u^{n,*}_{h,\hat{\rho}} = \tilde{U}^{n+1}0^\intercal + U^n Y^{n\intercal}.$$

Since $\langle 0^\intercal, Y^n\rangle_{L^2_{\hat\rho}} = 0$, using the definition (2.20) we have

$$u^{n,*}_{h,\hat\rho} \in \mathcal{T}_{\tilde U^{n+1}Y^{n\intercal}}\mathcal{M}^{h,\hat\rho}_R.$$

2. The newly computed solution $u^{n+1,*}_{h,\hat\rho}$ can be expressed as

$$u^{n+1,*}_{h,\hat\rho} = \tilde U^{n+1}(\tilde Y^{n+1} - Y^n)^\intercal + \tilde U^{n+1}Y^{n\intercal}.$$

Based on Lemma 2.2.2(1.) we know that $\langle \tilde Y^{n+1\intercal} - Y^{n\intercal}, Y^n\rangle_{L^2_{\hat\rho}} = 0$, i.e. again using the definition (2.20) we have $u^{n+1,*}_{h,\hat\rho} \in \mathcal{T}_{\tilde U^{n+1}Y^{n\intercal}}\mathcal{M}^{h,\hat\rho}_R$.

□

Remark 1. Note that for any function of the form $v = \tilde U^{n+1}K^\intercal$ or $v = JY^{n\intercal}$ with $K \in (L^2_{\hat\rho})^R$, $J \in (V_h)^R$, it holds $v \in \mathcal{T}_{\tilde U^{n+1}Y^{n\intercal}}\mathcal{M}^{h,\hat\rho}_R$ since we have

$$JY^{n\intercal} = \tilde U^{n+1}0^\intercal + JY^{n\intercal}, \quad \mathbb{E}_{\hat\rho}[0^\intercal Y^n] = 0$$

$$\tilde U^{n+1}K^\intercal = \tilde U^{n+1}(\mathcal{P}^\perp_{\hat\rho,Y^n}[K])^\intercal + \tilde U^{n+1}(\mathcal{P}_{\hat\rho,Y^n}[K])^\intercal, \quad \langle(\mathcal{P}^\perp_{\hat\rho,Y^n}[K])^\intercal, Y^n\rangle_{L^2_{\hat\rho}} = 0.$$

Since $\mathcal{T}_{\tilde U^{n+1}Y^{n\intercal}}\mathcal{M}^{h,\hat\rho}_R$ is a vector space, it includes any linear combination of $u^{n,*}_{h,\hat\rho}$ and $u^{n+1,*}_{h,\hat\rho}$. The following lemma is an analogue of Lemma 1.3.4 and will become useful when we derive the discrete variational formulation.

Lemma 2.2.6. *Let $u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho}$ be the discrete DLR solutions at times t_n, t_{n+1} as defined in Algorithm 2.2.1. Then the zero-mean parts $u^{n+1^*}_{h,\hat\rho}, u^{n^*}_{h,\hat\rho}$ satisfy*

$$\left(\frac{(u^{n+1}_{h,\hat\rho} - u^n_{h,\hat\rho})^*}{\triangle t} - \Pi^{h,\hat\rho}_{\tilde U^{n+1}Y^{n\intercal}}[\mathcal{F}^*(u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho})], v_h\right)_{V'V,L^2_{\hat\rho}} = 0, \quad \forall v_h \in V_h \otimes L^2_{\hat\rho}. \quad (2.22)$$

Proof. Multiplying (2.9) by Y^n_j and summing over j, we obtain

$$\left\langle \frac{\tilde U^{n+1}Y^{n\intercal} - u^{n,*}_{h,\hat\rho}}{\triangle t}, v_h\right\rangle_H - \left(\mathbb{E}_{\hat\rho}[\mathcal{F}(u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho})Y^n]Y^{n\intercal}, v_h\right)_{V'V} = 0, \quad \forall v_h \in V_h. \quad (2.23)$$

Noticing that

$$\begin{aligned}\mathbb{E}_{\hat\rho}[\mathcal{F}(u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho})Y^n]Y^{n\intercal} &= \mathbb{E}_{\hat\rho}[\mathcal{F}^*(u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho})Y^n]Y^{n\intercal}\\ &= \mathcal{P}_{\hat\rho,\mathcal{Y}^n}[\mathcal{F}^*(u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho})],\end{aligned}$$

and taking the weak formulation of (2.23) in $L^2_{\hat\rho}$ results in

$$\langle \tilde U^{n+1} Y^{n\intercal}, v_h\rangle_{H,L^2_{\hat\rho}} = \langle u^{n,*}_{h,\hat\rho}, v_h\rangle_{H,L^2_{\hat\rho}} + \triangle t\Big(\mathcal{P}_{\hat\rho,\mathcal{Y}^n}[\mathcal{F}^*(u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho})], v_h\Big)_{V'V,L^2_{\hat\rho}} \qquad \forall v_h \in V_h \otimes L^2_{\hat\rho}. \tag{2.24}$$

Similarly, multiplying (2.10) by $\tilde U^{n+1}$, and further writing (2.10) in a weak form in $L^2_{\hat\rho}$, we obtain

$$\Big\langle \frac{u^{n+1,*}_{h,\hat\rho} - \tilde U^{n+1} Y^{n\intercal}}{\triangle t} - \mathcal{P}^{\perp}_{\hat\rho,\mathcal{Y}^n}\Big[\Big(\mathcal{F}^*(u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho}), \tilde U^{n+1}\Big)_{V'V} (\tilde M^{n+1})^{-1} \tilde U^{n+1\intercal}\Big],\, w\Big\rangle_{L^2_{\hat\rho}} = 0, \qquad \forall w \in L^2_{\hat\rho}. \tag{2.25}$$

Since

$$\mathcal{P}^{\perp}_{\hat\rho,\mathcal{Y}^n}\Big[\Big(\mathcal{F}^*(u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho}), \tilde U^{n+1}\Big)_{V'V} (\tilde M^{n+1})^{-1} \tilde U^{n+1\intercal}\Big] = \mathcal{P}^{\perp}_{\hat\rho,\mathcal{Y}^n}[\mathcal{P}_{\tilde{\mathcal{U}}^{n+1}}[\mathcal{F}^*(u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho})]],$$

taking the weak formulation of (2.25) in V_h results in

$$\langle u^{n+1,*}_{h,\hat\rho}, v_h\rangle_{H,L^2_{\hat\rho}} = \langle \tilde U^{n+1} Y^{n\intercal}, v_h\rangle_{H,L^2_{\hat\rho}} + \triangle t\Big(\mathcal{P}^{\perp}_{\hat\rho,\mathcal{Y}^n}\Big[\mathcal{P}_{\tilde{\mathcal{U}}^{n+1}}[\mathcal{F}^*(u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho})]\Big], v_h\Big)_{V'V,L^2_{\hat\rho}} \qquad \forall v_h \in V_h \otimes L^2_{\hat\rho}. \tag{2.26}$$

Finally, summing equations (2.24) and (2.26) results in (2.22). □

We now proceed with the discrete variational formulation.

Theorem 2.2.7 (Discrete variational formulation)**.** *Let $u^n_{h,\hat\rho}$ and $u^{n+1}_{h,\hat\rho}$ be the discrete DLR solution at times t_n, t_{n+1}, respectively, $n = 0, \dots, N-1$, as defined in Algorithm 2.2.1. Then it holds*

$$\Big\langle \frac{u^{n+1}_{h,\hat\rho} - u^n_{h,\hat\rho}}{\triangle t},\, v_h\Big\rangle_{H,L^2_{\hat\rho}} = \Big(\mathcal{F}(u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho}),\, v_h\Big)_{V'V,L^2_{\hat\rho}}, \qquad \forall v_h = \bar v_h + v^*_h \text{ with } \bar v_h \in V_h \text{ and } v^*_h \in \mathcal{T}_{\tilde U^{n+1} Y^{n\intercal}} \mathcal{M}^{h,\hat\rho}_R. \tag{2.27}$$

Proof. Thanks to Lemma 2.2.5 we have $(u^{n+1}_{h,\hat\rho} - u^n_{h,\hat\rho})^* \in \mathcal{T}_{\tilde U^{n+1} Y^n} \mathcal{M}^{h,\hat\rho}_R$, and we can derive

$$\begin{aligned} \Big\langle \frac{(u^{n+1}_{h,\hat\rho} - u^n_{h,\hat\rho})^*}{\triangle t}, v_h\Big\rangle_{H,L^2_{\hat\rho}} &= \Big\langle \Pi^{h,\hat\rho}_{\tilde U^{n+1} Y^{n\intercal}}\Big[\frac{(u^{n+1}_{h,\hat\rho} - u^n_{h,\hat\rho})^*}{\triangle t}\Big], v_h\Big\rangle_{H,L^2_{\hat\rho}} \\ &= \Big\langle \frac{(u^{n+1}_{h,\hat\rho} - u^n_{h,\hat\rho})^*}{\triangle t}, \Pi^{h,\hat\rho}_{\tilde U^{n+1} Y^{n\intercal}}\Big[v_h\Big]\Big\rangle_{H,L^2_{\hat\rho}} \end{aligned} \tag{2.28}$$

and formula (2.21) gives us

$$\Big(\Pi^{h,\hat\rho}_{\tilde U^{n+1}Y^{n\intercal}}[\mathcal{F}^*(u^n_{h,\hat\rho},u^{n+1}_{h,\hat\rho})],\,v_h\Big)_{V'V,L^2_{\hat\rho}} \\ = \Big(\mathcal{F}^*(u^n_{h,\hat\rho},u^{n+1}_{h,\hat\rho}),\,\Pi^{h,\hat\rho}_{\tilde U^{n+1}Y^{n\intercal}}[v_h]\Big)_{V'V,L^2_{\hat\rho}}. \quad (2.29)$$

Summing (2.28), (2.29) and applying Lemma 2.2.6 results in

$$\Big\langle\frac{(u^{n+1}_{h,\hat\rho}-u^n_{h,\hat\rho})^*}{\triangle t},\,v_h\Big\rangle_{H,L^2_{\hat\rho}} = \Big(\mathcal{F}^*(u^n_{h,\hat\rho},u^{n+1}_{h,\hat\rho}),\,v_h\Big)_{V'V,L^2_{\hat\rho}} \\ \forall v_h\in\mathcal{T}_{\tilde U^{n+1}Y^{n\intercal}}\mathcal{M}^{h,\hat\rho}_R.$$

Now summing this to equation (2.8) we obtain

$$\Big\langle\frac{\bar u^{n+1}_{h,\hat\rho}-\bar u^n_{h,\hat\rho}+(u^{n+1}_{h,\hat\rho}-u^n_{h,\hat\rho})^*}{\triangle t},\,w_h+v_h\Big\rangle_{H,L^2_{\hat\rho}} \\ = \Big(\mathbb{E}_{\hat\rho}[\mathcal{F}(u^n_{h,\hat\rho},u^{n+1}_{h,\hat\rho})]+\mathcal{F}^*(u^n_{h,\hat\rho},u^{n+1}_{h,\hat\rho}),\,w_h+v_h\Big)_{V'V,L^2_{\hat\rho}} \\ \forall w_h\in V_h,\ \forall v_h\in\mathcal{T}_{\tilde U^{n+1}Y^{n\intercal}}\mathcal{M}^{h,\hat\rho}_R \quad (2.30)$$

which is equivalent to the final result (2.27). In (2.30) we have employed

$$\Big\langle\frac{(u^{n+1}_{h,\hat\rho}-u^n_{h,\hat\rho})^*}{\triangle t},\,w_h\Big\rangle_{H,L^2_{\hat\rho}} - \Big(\mathcal{F}^*(u^n_{h,\hat\rho},u^{n+1}_{h,\hat\rho}),\,w_h\Big)_{V'V,L^2_{\hat\rho}} = 0,\quad \forall w_h\in V_h$$

$$\Big\langle\frac{\bar u^{n+1}_{h,\hat\rho}-\bar u^n_{h,\hat\rho}}{\triangle t},\,v_h\Big\rangle_{H,L^2_{\hat\rho}} - \Big(\mathbb{E}_{\hat\rho}[\mathcal{F}(u^n_{h,\hat\rho},u^{n+1}_{h,\hat\rho})],\,v_h\Big)_{V'V,L^2_{\hat\rho}} = 0,\quad \forall v_h\in\mathcal{T}_{\tilde U^{n+1}Y^{n\intercal}}\mathcal{M}^{h,\hat\rho}_R,$$

which holds as $\mathbb{E}[v_h]=0$, $\forall v_h\in\mathcal{T}_{\tilde U^{n+1}Y^{n\intercal}}\mathcal{M}^{h,\hat\rho}_R$. □

2.2.3 Discrete variational formulation for the rank-deficient case

The discrete variational formulation established in the previous section is valid only in the case of the deterministic basis $\tilde U^{n+1}$ being linearly independent, since the proof of Theorem 2.2.7 implicitly involves the inverse of $\tilde M^{n+1}=\langle\tilde U^{n+1\intercal},\tilde U^{n+1}\rangle_H$. In this subsection, we show that a discrete variational formulation can be generalized for the rank-deficient case.

When applying the discretization scheme proposed in step 3. of Algorithm 2.2.1 with a rank-deficient matrix $\tilde M^{n+1}$, we recall that the solution $\tilde Y^{n+1}$ is defined as the solution of (2.10) minimizing $\|\tilde Y^{n+1}-Y^n\|_{L^2_{\hat\rho}}$. Note that minimizing $\|\tilde Y^{n+1}-Y^n\|_{L^2_{\hat\rho}}$ is equivalent to minimizing the norm $\|\tilde{Y}^{n+1}(\omega_k) - Y^n(\omega_k)\|_{\mathbb{R}^R}$ for every sample point ω_k, $k = 1, \dots, \hat{N}$, where $\|\cdot\|^2_{\mathbb{R}^R} = \langle \cdot, \cdot \rangle_{\mathbb{R}^R}$ denotes the Euclidean scalar product in $\mathbb{R}^R$.

In what follows we will exploit the fact that the vector space $L^2_{\hat{\rho}}$ is isomorphic to $\mathbb{R}^{\hat{N}}$. In particular, it holds that $(\tilde{Y}^{n+1} - Y^n)^\intercal \in \mathbb{R}^{R\times\hat{N}}$, where each column of $(\tilde{Y}^{n+1} - Y^n)^\intercal$ is given by $(\tilde{Y}^{n+1} - Y^n)(\omega_k)$, $k = 1, \dots, \hat{N}$. With a little abuse of notation, we use $\tilde{U}^{n+1} : \mathbb{R}^R \to V_h$ to denote a linear operator which takes real coefficients and returns the corresponding linear combination of the basis functions $\tilde{U}^{n+1}$. By $\tilde{U}^{n+1\intercal} : V_h \to \mathbb{R}^R$ we denote its dual.

Lemma 2.2.8. *For any discrete solution $\tilde{Y}^{n+1}$ of equation* (2.10) *that minimizes the norm $\|\tilde{Y}^{n+1} - Y^n\|_{L^2_{\hat{\rho}}}$, it holds that every column of the increment $(\tilde{Y}^{n+1} - Y^n)^\intercal$ lies in the $\langle \cdot, \cdot \rangle_{\mathbb{R}^R}$-orthogonal complement of the kernel of $\tilde{M}^{n+1}$, i.e.*

$$(\tilde{Y}^{n+1} - Y^n)^\intercal \ \in \ \Big(\ker(\tilde{M}^{n+1})^\perp\Big)^{\hat{N}}$$

where $\ker(\tilde{M}^{n+1}) = \{v \in \mathbb{R}^R : \ \tilde{M}^{n+1} v = 0\}$.

Proof. Seeking a contradiction, let us suppose that $(\tilde{Y}^{n+1} - Y^n)^\intercal \notin \Big(\ker(\tilde{M}^{n+1})^\perp\Big)^{\hat{N}}$. Let

$$Z^\intercal := \tilde{Y}^{n+1\intercal} - \mathcal{P}_{\ker(\tilde{M}^{n+1})}[\tilde{Y}^{n+1\intercal} - Y^{n\intercal}] \ \neq \ \tilde{Y}^{n+1}, \tag{2.31}$$

where $\mathcal{P}_{\ker(\tilde{M}^{n+1})}[v] \in \mathbb{R}^{R\times\hat{N}}$ for $v \in \mathbb{R}^{R\times\hat{N}}$ denotes the column-wise application of $\langle \cdot, \cdot \rangle_{\mathbb{R}^R}$-orthogonal projection onto the kernel of $\tilde{M}^{n+1}$. Then, such constructed Z satisfies

$$\begin{aligned}\|(Z - Y^n)(\omega_k)\|_{\mathbb{R}^R} &= \|\Big(\tilde{Y}^{n+1} - Y^n - \mathcal{P}_{\ker(\tilde{M}^{n+1})}[\tilde{Y}^{n+1} - Y^n]\Big)(\omega_k)\|_{\mathbb{R}^R} \\ &< \|(\tilde{Y}^{n+1} - Y^n)(\omega_k)\|_{\mathbb{R}^R},\end{aligned}$$

and solves (2.10):

$$\begin{aligned}\tilde{M}^{n+1}(Z - Y^n)^\intercal &= \tilde{M}^{n+1}(\tilde{Y}^{n+1} - Y^n)^\intercal \\ &= \triangle t \mathcal{P}^\perp_{\hat{\rho},\mathcal{Y}^n}\left[\Big(\mathcal{F}^*(u^n_{h,\hat{\rho}}, \bar{u}^{n+1} + \tilde{U}^{n+1}\tilde{Y}^{n+1\intercal}), \tilde{U}^{n+1}\Big)_{V'V}\right] \\ &= \triangle t \mathcal{P}^\perp_{\hat{\rho},\mathcal{Y}^n}\left[\Big(\mathcal{F}^*(u^n_{h,\hat{\rho}}, \bar{u}^{n+1} + \tilde{U}^{n+1}Z^\intercal), \tilde{U}^{n+1}\Big)_{V'V}\right],\end{aligned}$$

where in the last step we used that $\ker(\tilde{M}^{n+1}) = \ker(\tilde{U}^{n+1})$. This leads to a contradiction that $\tilde{Y}^{n+1}$ was the solution minimizing $\|\tilde{Y}^{n+1} - Y^n\|_{L^2_{\hat{\rho}}}$. □

When showing the equivalence between the DLR variational formulation (1.17) and the

DLR system of equations (1.8)–(1.10) in the continuous setting, the DO condition

$$\langle \dot{Y}_i, Y_j \rangle_{L^2_{\hat\rho}} = 0, \quad \forall 1 \le i, j \le R \tag{2.32}$$

plays an important role. In an analogous way, the discrete DO condition (property 1. from Lemma 2.2.2 for the full-rank case) plays an important role when showing the equivalence between the discrete DLR system of equations and the discrete DLR variational formulation.

Lemma 2.2.9. *Any discrete solution $\tilde{Y}^{n+1}$ of equation* (2.10) *which minimizes the norm $\|\tilde{Y}^{n+1} - Y^n\|_{L^2_{\hat\rho}}$, satisfies the discrete DO condition*

$$\left\langle \left(\frac{\tilde{Y}^{n+1} - Y^n}{\triangle t} \right)^{\intercal}, Y^n \right\rangle_{L^2_{\hat\rho}} = 0. \tag{2.33}$$

Proof. Let $\tilde{Y}^{n+1}$ be a solution of (2.10) minimizing $\|\tilde{Y}^{n+1} - Y^n\|_{L^2_{\hat\rho}}$. Thanks to Lemma 2.2.8 we know that

$$\left(\tilde{Y}^{n+1} - Y^n\right)^{\intercal} \in \left(\ker(\tilde{M}^{n+1})^{\perp}\right)^{\hat{N}}.$$

Now, let $\tilde{M}^{n+1^+}$ denote the pseudoinverse of $\tilde{M}^{n+1}$. Since

$$\tilde{M}^{n+1^+} \tilde{M}^{n+1} v = v$$

for any $v \in \ker(\tilde{M}^{n+1})^{\perp}$, the solution $\tilde{Y}^{n+1}$ of equation (2.10) satisfies

$$\tilde{Y}^{n+1^{\intercal}} = Y^{n^{\intercal}} + \triangle t \tilde{M}^{n+1^+} \mathcal{P}^{\perp}_{\hat\rho, \mathcal{Y}^n} \left[\left(\mathcal{F}^*(u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho}), \tilde{U}^{n+1} \right)^{\intercal}_{V'V} \right]. \tag{2.34}$$

Thus, if we have

$$\mathbb{E}_{\hat\rho}\left[Y^{n^{\intercal}} \left(\mathcal{P}^{\perp}_{\hat\rho, \mathcal{Y}^n} \left[\left(\mathcal{F}^*(u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho}), \tilde{U}^{n+1} \right)_{V'V} \right] \tilde{M}^{n+1^+} \right) \right] = 0,$$

then the statement will follow. But for the column space of

$$\mathcal{P}^{\perp}_{\hat\rho, \mathcal{Y}^n} \left[\left(\mathcal{F}^*(u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho}), \tilde{U}^{n+1} \right)_{V'V} \right] \tilde{M}^{n+1^+} \in \mathbb{R}^{\hat{N} \times R}$$

it holds

$$\begin{aligned} &\operatorname{span}\left\{ \mathcal{P}^{\perp}_{\hat\rho, \mathcal{Y}^n} \left[\left(\mathcal{F}^*(u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho}), \tilde{U}^{n+1} \right)_{V'V} \right] \tilde{M}^{n+1^+} \right\} \\ &\qquad\qquad \subset \operatorname{span}\left\{ \mathcal{P}^{\perp}_{\hat\rho, \mathcal{Y}^n} \left[\left(\mathcal{F}^*(u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho}), \tilde{U}^{n+1} \right)_{V'V} \right] \right\} \subset \mathcal{Y}^{n^{\perp}}_{\hat\rho} \end{aligned}$$

with $\mathcal{Y}^{n^{\perp}}_{\hat\rho} \subset \mathbb{R}^{\hat{N}}$ being the orthogonal complement to $\mathcal{Y}^n$ in the scalar product $\langle \cdot, \cdot \rangle_{L^2_{\hat\rho}}$. Now the proof is complete. □

In the following lemma we address the question of existence of a unique solution when applying the explicit and semi-implicit scheme.

Lemma 2.2.10. *For the explicit and semi-implicit scheme, as described in Section 2.2.1, there exists a unique discrete solution* $\tilde{Y}^{n+1}$ *of equation* (2.10) *minimizing the norm* $\|\tilde{Y}^{n+1} - Y^n\|_{L^2_{\hat{\rho}}}$.

Proof. We will start with the semi-implicit scheme. By virtue of Lemma 2.2.3, under the discrete DO condition (2.33), applying the semi-implicit scheme to equation (2.10) is equivalent to solving equation (2.18). We will first focus our attention to equation (2.18) and show that there exists a unique solution minimizing $\|\tilde{Y}^{n+1} - Y^n\|_{L^2_{\hat{\rho}}}$. This solution will satisfy the discrete DO and consequently is a unique minimizing solution of (2.10). Equation (2.18) can be rewritten as

$$B\,(\tilde{Y}^{n+1} - Y^n)^\intercal = \text{RHS} \qquad \text{in } L^2_{\hat{\rho}}, \tag{2.35}$$

where

$$\begin{aligned} B &= \tilde{M}^{n+1} + \triangle t \langle \tilde{U}^{n+1\intercal}, \tilde{U}^{n+1} \rangle_{\mathcal{L}_{\text{det}}} \\ \text{RHS} &= -\triangle t \Big(\tilde{U}^{n+1\intercal}, \mathcal{P}^{\perp}_{\hat{\rho}, \mathcal{Y}^n} [\mathcal{L}^*_{\text{stoch}}(u^n_{h,\hat{\rho}}) - f^{n,n+1^*}] \Big)_{VV'}. \end{aligned}$$

Since RHS above lies in the range of $\tilde{U}^{n+1\intercal}$, which is the same as the range of B, a solution of (2.35) exists. Moreover, since the matrix B is positive definite on the space $\ker(B)^\perp$, any solution can be expressed as $(\tilde{Y}^{n+1} - Y^n + W)^\intercal$ with $W^\intercal \in \Big(\ker(B)\Big)^{\hat{N}}$ and a unique $\tilde{Y}^{n+1\intercal} \in \mathbb{R}^{R \times \hat{N}}$ such that $(\tilde{Y}^{n+1} - Y^n)^\intercal \in \Big(\ker(B)^\perp\Big)^{\hat{N}}$. The solution $\tilde{Y}^{n+1}$ minimizes each column $\|(\tilde{Y}^{n+1} - Y^n)(\omega_k)\|_{\mathbb{R}^R}$, $k = 1, \ldots, \hat{N}$ and thus it is the unique solution of (2.35) that minimizes norm $\|\tilde{Y}^{n+1} - Y^n\|_{L^2_{\hat{\rho}}}$. We observe that the established solution $\tilde{Y}^{n+1}$ of equation (2.35) satisfies the discrete DO condition (2.33). The argument is analogous to the proof of Lemma 2.2.9, but instead of $\tilde{M}^{n+1}$ here we take B. Therefore, the statement for the semi-implicit scheme follows. The explicit case can be shown by following analogous steps with

$$\begin{aligned} B &= \tilde{M}^{n+1}, \\ \text{RHS} &= \triangle t \Big(\tilde{U}^{n+1\intercal}, \mathcal{P}^{\perp}_{\hat{\rho}, \mathcal{Y}^n} [\mathcal{F}^*(u^n_{h,\hat{\rho}})] \Big)_{VV'}. \end{aligned}$$

□

Now we can proceed with showing the discrete variational formulation. It is not generally easy to deal with the notion of a tangent space at a certain point on the manifold in the rank-deficient case. In the following theorem we will, however, show that an analogous discrete variational formulation holds. Given $U \in (V_h)^R$ and $Y \in (L^2_{\hat{\rho},0})^R$, we define the vector space $\mathcal{T}_{UY^\intercal}$ as

$$\mathcal{T}_{UY^\intercal} = \Big\{\delta v \in V_h \otimes L^2_{\hat\rho,0} \mid \delta v = \sum_{i=1}^{R} U_i \delta Y_i + \delta U_i Y_i \\ \delta U_i \in V_h,\ \delta Y_i \in L^2_{\hat\rho,0},\ \langle \delta Y_i, Y_j\rangle_{L^2_{\hat\rho}} = 0 \quad \forall i,j = 1,\dots,R\Big\}.$$

It is easy to verify that, analogously to Lemma 2.2.5, the (possibly rank-deficient) discrete DLR solutions $u^n_{h,\hat\rho}$ and $u^{n+1}_{h,\hat\rho}$ at times t_n, t_{n+1}, as defined in Algorithm 2.2.1 satisfy

$$u^n_{h,\hat\rho} \in \mathcal{T}_{\tilde U^{n+1} Y^{n\intercal}}, \qquad u^{n+1}_{h,\hat\rho} \in \mathcal{T}_{\tilde U^{n+1} Y^{n\intercal}}. \tag{2.36}$$

Theorem 2.2.11. *Let $u^n_{h,\hat\rho}$ and $u^{n+1}_{h,\hat\rho}$ be the (possibly rank-deficient) discrete DLR solution at times t_n, t_{n+1}, respectively, $n = 0,\dots,N-1$, as defined in Algorithm 2.2.1. Then the following variational formulation holds*

$$\begin{aligned} \Big\langle \frac{u^{n+1}_{h,\hat\rho} - u^n_{h,\hat\rho}}{\triangle t},\, v_h \Big\rangle_{H,L^2_{\hat\rho}} = \Big(\mathcal{F}(u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho}),\, v_h\Big)_{V'V,L^2_{\hat\rho}}, \\ \forall v_h = \bar v_h + v_h^* \textit{ with } \bar v_h \in V_h \textit{ and } v_h^* \in \mathcal{T}_{\tilde U^{n+1} Y^{n\intercal}}. \end{aligned} \tag{2.37}$$

Proof. First, consider equation (2.9) with $v_h = \tilde U^{n+1}_j$. Summing over j results in

$$\Big(\mathbb{E}_{\hat\rho}[(\mathcal{F}^*(u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho}))Y^n], \tilde U^{n+1}\Big)_{V'V} = \frac{1}{\triangle t}\Big(\tilde M^{n+1} - \langle U^{n\intercal}, \tilde U^{n+1}\rangle_H\Big). \tag{2.38}$$

Let us proceed with the equation (2.10):

$$\begin{aligned} 0 &= \frac{\tilde Y^{n+1} - Y^n}{\triangle t}\tilde M^{n+1} - \mathcal{P}^{\perp}_{\hat\rho,\mathcal{Y}^n}\Big[\Big(\mathcal{F}^*(u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho}), \tilde U^{n+1}\Big)_{V'V}\Big] \\ &= \frac{\tilde Y^{n+1} - Y^n}{\triangle t}\tilde M^{n+1} - \Big(\mathcal{F}^*(u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho}), \tilde U^{n+1}\Big)_{V'V} \\ &\qquad + Y^n\Big(\mathbb{E}_{\hat\rho}[(\mathcal{F}^*(u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho}))Y^{n\intercal}], \tilde U^{n+1}\Big)_{V'V} \\ &= \frac{\tilde Y^{n+1}\langle \tilde U^{n+1\intercal}, \tilde U^{n+1}\rangle_H - Y^n\tilde M^{n+1} + Y^n\tilde M^{n+1} - Y^n\langle U^{n\intercal}, \tilde U^{n+1}\rangle_H}{\triangle t} \\ &\qquad - \Big(\mathcal{F}^*(u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho}), \tilde U^{n+1}\Big)_{V'V} \\ &= \Big\langle \frac{(u^{n+1}_{h,\hat\rho} - u^n_{h,\hat\rho})^*}{\triangle t},\, \tilde U^{n+1}\Big\rangle_H - \Big(\mathcal{F}^*(u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho}),\, \tilde U^{n+1}\Big)_{V'V} \end{aligned}$$

Taking a weak formulation in $L^2_{\hat{\rho},0}$ results in

$$\left\langle \frac{(u^{n+1}_{h,\hat{\rho}} - u^n_{h,\hat{\rho}})^*}{\triangle t},\, w_h \right\rangle_{H,L^2_{\hat{\rho}}} - \left(\mathcal{F}^*(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}}),\, w_h \right)_{V'V,L^2_{\hat{\rho}}} = 0$$
$$\forall w_h = \tilde{U}^{n+1} \delta Y^{\intercal},\ \delta Y \in (L^2_{\hat{\rho},0})^R. \quad (2.39)$$

Concerning equation (2.9), we proceed as follows: $\forall v_h \in (V_h)^R$

$$\begin{aligned} 0 &= \left\langle \frac{\tilde{U}^{n+1} - U^n}{\triangle t}, v_h \right\rangle_H - \left(\mathbb{E}_{\hat{\rho}}[(\mathcal{F}(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}}))Y^n], v_h \right)_{V'V} \\ &= \left\langle \frac{\tilde{U}^{n+1}\mathbb{E}_{\hat{\rho}}[\tilde{Y}^{n+1\intercal} Y^n] - U^n \mathbb{E}_{\hat{\rho}}[Y^{n\intercal} Y^n]}{\triangle t}, v_h \right\rangle_H - \left(\mathbb{E}_{\hat{\rho}}[(\mathcal{F}(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}}))Y^n], v_h \right)_{V'V} \\ &= \left\langle \frac{(u^{n+1}_{h,\hat{\rho}} - u^n_{h,\hat{\rho}})^*}{\triangle t},\, v_h Y^{n\intercal} \right\rangle_{H,L^2_{\hat{\rho}}} - \left(\mathcal{F}^*(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}}),\, v_h Y^{n\intercal} \right)_{V'V,L^2_{\hat{\rho}}} \\ &\qquad \forall v_h \in (V_h)^R, \end{aligned} \quad (2.40)$$

where in the second step we applied $\mathbb{E}_{\hat{\rho}}[\tilde{Y}^{n+1\intercal} Y^n] = \mathrm{Id}$ which holds thanks to the discrete DO condition from Lemma 2.2.9. Summing equation (2.39) and (2.40) we obtain

$$\left\langle \frac{(u^{n+1}_{h,\hat{\rho}} - u^n_{h,\hat{\rho}})^*}{\triangle t},\, w_h \right\rangle_{H,L^2_{\hat{\rho}}} - \left(\mathcal{F}^*(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}}),\, w_h \right)_{V'V,L^2_{\hat{\rho}}} = 0$$
$$\forall w_h \in \mathcal{T}_{\tilde{U}^{n+1} Y^{n\intercal}}.$$

The rest of the proof follows the same steps as in the proof of Theorem 2.2.7, i.e. summing the mean value equation (2.8) and noting that some terms vanish. □

2.2.4 Reinterpretation as a projector-splitting scheme

The proposed Algorithm 2.2.1 was derived from the DLR system of equations (1.8)–(1.10). This subsection is dedicated to showing that this scheme can in fact be formulated as a projector-splitting scheme for the time discretization of the Dual DO approximation of (1.3). Afterwards, we will continue by showing its connection to the projector-splitting scheme described in Section 2.1, which was proposed in [LO14; LOV15b] and further analyzed in [KLW16].

In what follows, we will focus on the evolution of $u^{n,*}_{h,\hat{\rho}}$, i.e. the 0-mean part of the discrete DLR solution $u^n_{h,\hat{\rho}}$.

Lemma 2.2.12. *The discretized system of equations* (2.9)–(2.10) *can be equivalently* *reformulated as*

$$\langle \tilde{u}_{h,\hat{\rho}}, v_h\rangle_{H,L^2_{\hat{\rho}}} = \langle u^{n,*}_{h,\hat{\rho}}, v_h\rangle_{H,L^2_{\hat{\rho}}} + \triangle t\Big(\mathcal{P}_{\hat{\rho},\mathcal{Y}^n}[\mathcal{F}^*(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}})], v_h\Big)_{V'V,L^2_{\hat{\rho}}} \tag{2.41}$$

$$\begin{aligned}\langle u^{n+1,*}_{h,\hat{\rho}}, v_h\rangle_{H,L^2_{\hat{\rho}}} &= \langle \tilde{u}_{h,\hat{\rho}}, v_h\rangle_{H,L^2_{\hat{\rho}}} \\ &+ \triangle t\Big(\mathcal{P}^{\perp}_{\hat{\rho},\mathcal{Y}^n}\Big[\mathcal{P}_{\tilde{\mathcal{U}}^{n+1}}[\mathcal{F}^*(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}})]\Big], v_h\Big)_{V'V,L^2_{\hat{\rho}}}, \qquad (2.42)\\ &\qquad\qquad \forall v_h \in V_h \otimes L^2_{\hat{\rho}},\end{aligned}$$

where $\tilde{u}_{h,\hat{\rho}} = \tilde{U}^{n+1} Y^{n\intercal}$.

Proof. These equations are essentially equations (2.24) and (2.26), which are shown to hold in the proof of Lemma 2.2.6. □

We recall that from Lemma 1.3.4, the zero-mean part of the continuous DLR approximation $u^* = UY^\intercal$ satisfies

$$\begin{aligned}(\dot{u}^* - \Pi_{u^*}[\mathcal{F}^*(u) - f^*],\, v)_{V'V,L^2_\rho} \\ = (\dot{u}^* - \mathcal{P}_{\mathcal{Y}}[\mathcal{F}^*(u)] - \mathcal{P}^{\perp}_{\mathcal{Y}}[\mathcal{P}_{\mathcal{U}}[\mathcal{F}^*(u)]],\, v)_{V'V,L^2_\rho} = 0, \\ \forall v \in L^2_\rho(\Omega; V).\end{aligned}$$

Lemma 2.2.12 therefore shows that the time integration scheme corresponds to a projection-splitting scheme in which first the projection $\mathcal{P}_{\mathcal{Y}}[\mathcal{F}^*(u)]$ and then the projection $\mathcal{P}^{\perp}_{\mathcal{Y}}[\mathcal{P}_{\mathcal{U}}[\mathcal{F}^*(u)]]$ are applied.

2.3 Linking the projector-splitting integrator from Section 2.1 with the staggered scheme of Section 2.2

The projector-splitting scheme of Section 2.1 is a time integration scheme successfully used for the integration of dynamical low rank approximation in the DDO formulation. This section provides a detailed look into the comparison of the Algorithm 2.2.1 and the Algorithm 2.1.2. We will see that, if the solution is full rank, these schemes are in fact equivalent.

The projector-splitting integrator was originally proposed to deal with R-rank approximation of time-dependent matrices. For the case of time-dependent differential equations, the authors in [LO14] propose to apply $\triangle A = \triangle t \mathcal{F}(u^n_{h,\hat{\rho}})$. In this work, we consider a more general expression

$$\triangle A = \triangle t\Big(\mathcal{F}(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}})\Big)$$

where $\mathcal{F}(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}})$ can be any of the explicit, implicit or semi-implicit discretizations detailed in Section 2.2.1.

Now, let us compare the steps of Algorithm 2.1.2 to Algorithm 2.2.1. We can easily observe that $\hat{\bar{u}}^{n+1} = \bar{u}^{n+1}$. Since $Y^n = V_0$, we can see that equation (2.9) is equivalent to step 1 with $U^n = U_0 S_0$, i.e. $K_1 = \tilde{U}^{n+1}$. Further, we have

$$\tilde{M}^{n+1} = \langle \tilde{U}^{n+1^\intercal}, \tilde{U}^{n+1} \rangle_H = \hat{S}_1^\intercal \langle U_1^\intercal, U_1 \rangle_H \hat{S}_1 = \hat{S}_1^\intercal \hat{S}_1.$$

Equation (2.10) can be reformulated as

$$\begin{aligned}\tilde{Y}^{n+1}\hat{S}_1^\intercal \hat{S}_1 = Y^n \hat{S}_1^\intercal \hat{S}_1 - \triangle t Y^n \Big(\mathbb{E}_{\hat{\rho}}\Big[Y^{n^\intercal}(\mathcal{F}^*(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}}))\Big], U_1\Big)_{V'V} \hat{S}_1 \\ + \triangle t \Big(\mathcal{F}^*(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}}), U_1\Big)_{V'V} \hat{S}_1,\end{aligned}$$

which, provided $\hat{S}_1$ is invertible, is equivalent to

$$\begin{aligned}\tilde{Y}^{n+1}\hat{S}_1^\intercal = Y^n \Big(\hat{S}_1^\intercal - \triangle t \Big(\mathbb{E}_{\hat{\rho}}\Big[Y^{n^\intercal}(\mathcal{F}^*(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}}))\Big], U_1\Big)_{V'V}\Big) \\ + \triangle t \Big(\mathcal{F}^*(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}}), U_1\Big)_{V'V}.\end{aligned}$$

Note that the expression in brackets in the first term on the right hand side is exactly the transpose of $\tilde{S}_0$ from step 3:

$$\hat{S}_1^\intercal - \triangle t \Big(\mathbb{E}_{\hat{\rho}}\Big[Y^{n^\intercal}(\mathcal{F}^*(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}}))\Big], U_1\Big)_{V'V} = \tilde{S}_0^\intercal,$$

from which we deduce

$$L_1 = \tilde{Y}^{n+1}\hat{S}_1^\intercal.$$

Finally, we have

$$\hat{u}^{n+1,*}_{h,\hat{\rho}} = U_1 S_1 V_1^\intercal = U_1 L_1^\intercal = U_1 \hat{S}_1 \tilde{Y}^{n+1^\intercal} = \tilde{U}^{n+1}\tilde{Y}^{n+1^\intercal} = u^{n+1,*}_{h,\hat{\rho}}.$$

We conclude that the scheme in Algorithm 2.2.1 and the scheme in Algorithm 2.1.2 coincide in exact arithmetic, provided the matrix S_1 is invertible. However, the numerical behavior of the two schemes differs when S_1 is singular or close to singular. For $\tilde{M}^{n+1}$ close to singular, solving equation (2.10) might lead to numerical instabilities. This problem seems to be avoided in the projector-splitting scheme from [LO14; LOV15b], as no matrix inversion is involved. Such ill conditioning is however hidden in performing step 3. of Algorithm 2.1.2, since the QR or SVD decomposition can become unstable for ill-conditioned matrices (see [GVL96, chap. 5]). In the case of a rank deficient basis $\{\tilde{U}^{n+1}\}$, Algorithm 2.2.1 updates the stochastic basis by solving equation (2.10) in a least square sense while minimizing the norm $\|\tilde{Y}^{n+1} - Y^n\|_{L^2_{\hat{\rho}}}$. The previous subsection showed that such solution satisfies the discrete variational formulation which plays a crucial role in stability estimation (see Section 3.4). On the other hand, Algorithm 2.1.2

relies on the somehow arbitrary completion of the basis $\{U_1\}$ in the step 3. In presence of rank deficiency, the two algorithms can deliver different solutions (see section 3.6.3 for a numerical comparison). In the following theorem we show, that the discrete solution obtained by the projector-splitting integrator of Algorithm 2.1.2 satisfies a similar discrete variational formulation.

Lemma 2.3.1. *Let $\hat{u}^n_{h,\hat{\rho}}, \hat{u}^{n+1}_{h,\hat{\rho}}$ be the discrete DLR solutions at times t_n, t_{n+1} as defined in Algorithm 2.1.2. Then the zero-mean parts $\hat{u}^{n+1^*}_{h,\hat{\rho}}, \hat{u}^{n^*}_{h,\hat{\rho}}$ satisfy*

$$\left(\frac{(\hat{u}^{n+1}_{h,\hat{\rho}} - \hat{u}^n_{h,\hat{\rho}})^*}{\triangle t} - \Pi^{h,\hat{\rho}}_{U_1 Y^{n\intercal}}[\mathcal{F}^*(\hat{u}^n_{h,\hat{\rho}}, \hat{u}^{n+1}_{h,\hat{\rho}})], v_h\right)_{V'V,L^2_{\hat{\rho}}} = 0, \quad \forall v_h \in V_h \otimes L^2_{\hat{\rho}}. \tag{2.43}$$

Proof. Simply following a reversed order of the steps of the Algorithm 2.1.2, we derive for $\forall v_h \in V_h \otimes L^2_{\hat{\rho}}$

$$\begin{aligned}
\langle \hat{u}^{n+1^*}_{h,\hat{\rho}}, v_h\rangle_{H,L^2_{\hat{\rho}}} &= \langle U_1 L_1^\intercal, v_h\rangle_{H,L^2_{\hat{\rho}}} = \left\langle U_1\tilde{S}_0 V_0^\intercal + \triangle t U_1 (U_1, \mathcal{F}^*(\hat{u}^n_{h,\hat{\rho}}, \hat{u}^{n+1}_{h,\hat{\rho}}))_{VV',L^2_{\hat{\rho}}}, v_h\right\rangle_{H,L^2_{\hat{\rho}}} \\
&= \Big\langle U_1\hat{S}_1 V_0^\intercal - \triangle t\, U_1(\mathbb{E}_{\hat{\rho}}[\mathcal{F}^*(\hat{u}^n_{h,\hat{\rho}}, \hat{u}^{n+1}_{h,\hat{\rho}})V_0], U_1)_{V'V,L^2_{\hat{\rho}}} V_0^\intercal \\
&\qquad + \triangle t\, U_1(U_1, \mathcal{F}^*(\hat{u}^n_{h,\hat{\rho}}, \hat{u}^{n+1}_{h,\hat{\rho}}))_{VV',L^2_{\hat{\rho}}}, v_h\Big\rangle_{H,L^2_{\hat{\rho}}} \\
&= \Big\langle K_1 V_0^\intercal - \triangle t\, U_1(\mathbb{E}_{\hat{\rho}}[\mathcal{F}^*(\hat{u}^n_{h,\hat{\rho}}, \hat{u}^{n+1}_{h,\hat{\rho}})V_0], U_1)_{V'V,L^2_{\hat{\rho}}} V_0^\intercal \\
&\qquad + \triangle t\, U_1(U_1, \mathcal{F}^*(\hat{u}^n_{h,\hat{\rho}}, \hat{u}^{n+1}_{h,\hat{\rho}}))_{VV',L^2_{\hat{\rho}}}, v_h\Big\rangle_{H,L^2_{\hat{\rho}}} \\
&= \Big\langle U_0 S_0 V_0^\intercal + \mathbb{E}_{\hat{\rho}}[\mathcal{F}^*(\hat{u}^n_{h,\hat{\rho}}, \hat{u}^{n+1}_{h,\hat{\rho}})V_0]V_0^\intercal - \triangle t\, U_1(\mathbb{E}_{\hat{\rho}}[\mathcal{F}^*(\hat{u}^n_{h,\hat{\rho}}, \hat{u}^{n+1}_{h,\hat{\rho}})V_0], U_1)_{V'V,L^2_{\hat{\rho}}} V_0^\intercal \\
&\qquad + \triangle t\, U_1(U_1, \mathcal{F}^*(\hat{u}^n_{h,\hat{\rho}}, \hat{u}^{n+1}_{h,\hat{\rho}}))_{VV',L^2_{\hat{\rho}}}, v_h\Big\rangle_{H,L^2_{\hat{\rho}}} \\
&= \Big\langle \hat{u}^{n^*}_{h,\hat{\rho}} + \triangle t\, \Pi^{h,\hat{\rho}}_{U_1 V_0^\intercal}[\mathcal{F}^*(\hat{u}^n_{h,\hat{\rho}}, \hat{u}^{n+1}_{h,\hat{\rho}})], v_h\rangle_{V'V,L^2_{\hat{\rho}}}.
\end{aligned}$$

Since $Y^n = V_0$, we arrive at the sought statement. ☐

Lemma 2.3.2. *Let $u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}}$ be the discrete DLR solution at t_n, t_{n+1}, respectively, from the scheme in Algorithm 2.1.2. Then the zero-mean parts $u^{n,*}_{h,\hat{\rho}}, u^{n+1,*}_{h,\hat{\rho}}$ satisfy*

1. $u^{n^*}_{h,\hat{\rho}} \in \mathcal{T}_{U_1 Y^{n\intercal}}\mathcal{M}^{h,\hat{\rho}}_R$,
2. $u^{n+1,*}_{h,\hat{\rho}} \in \mathcal{T}_{U_1 Y^{n\intercal}}\mathcal{M}^{h,\hat{\rho}}_R$.

Proof. The proof follows analogous steps as the proof of Lemma 2.2.5. ☐

Theorem 2.3.3. *Let $\hat{u}^n_{h,\hat{\rho}}$ and $\hat{u}^{n+1}_{h,\hat{\rho}}$ be the (possibly rank-deficient) discrete DLR solution at times t_n, t_{n+1}, respectively, $n = 0, \ldots, N-1$, as defined in Algorithm 2.1.2. Then the* *following variational formulation holds*

$$\Big\langle \frac{\hat{u}^{n+1}_{h,\hat{\rho}} - \hat{u}^{n}_{h,\hat{\rho}}}{\triangle t}, v_h \Big\rangle_{H,L^2_{\hat{\rho}}} = \Big(\mathcal{F}(\hat{u}^{n}_{h,\hat{\rho}}, \hat{u}^{n+1}_{h,\hat{\rho}}), v_h\Big)_{V'V,L^2_{\hat{\rho}}}, \qquad \forall v_h = \bar{v}_h + v_h^* \text{ with } \bar{v}_h \in V_h \text{ and } v_h^* \in \mathcal{T}_{U_1 Y^{n\intercal}} \mathcal{M}^{h,\hat{\rho}}_R. \tag{2.44}$$

Proof. Applying the variational formulation (2.43) together with Lemma 2.3.2 and (2.21) we derive

$$\Big\langle \frac{(\hat{u}^{n+1}_{h,\hat{\rho}} - \hat{u}^{n}_{h,\hat{\rho}})^*}{\triangle t}, v_h \Big\rangle_{H,L^2_{\hat{\rho}}} = \Big(\mathcal{F}^*(\hat{u}^{n}_{h,\hat{\rho}}, \hat{u}^{n+1}_{h,\hat{\rho}}), v_h\Big)_{V'V,L^2_{\hat{\rho}}}, \quad \forall v_h \in \mathcal{T}_{U_1 Y^{n\intercal}} \mathcal{M}^{h,\hat{\rho}}_R.$$

To incorporate the mean value in the variational formulation we follow analogous steps as in the proof of Theorem 2.2.7. □

The variational formulation (2.44) holds in the rank-deficient case as well.

Remark 2. Note that the ordering of the equations in Algorithm 2.2.1 is crucial. When dealing with the DO formulation, i.e. orthonormal deterministic basis and linearly independent stochastic basis, we shall first update the stochastic basis and then evolve the deterministic basis. For a reversed ordering the Theorem 2.2.7 would not hold.

Note that in the full-rank case (when $\tilde{M}^{n+1}$ is full rank), it holds

$$\mathcal{T}_{U_1 Y^{n\intercal}} \mathcal{M}^{h,\hat{\rho}}_R = \mathcal{T}_{\tilde{U}^{n+1} Y^{n\intercal}} \mathcal{M}^{h,\hat{\rho}}_R.$$

However, in the rank deficient case,

$$\mathcal{T}_{U_1 Y^{n\intercal}} \mathcal{M}^{h,\hat{\rho}}_R \neq \mathcal{T}_{\tilde{U}^{n+1} Y^{n\intercal}}.$$

Comparing the variational formulations (2.27) and (2.44), it is clear, why the two discrete DLR solutions are equal in the full-rank case but differ in the rank deficient case.

The discrete variational formulation provides a geometric insight into the projector-splitting algorithm. It becomes useful when analysing stability, estimating the error caused by discretization as well as providing a geometric proof for the exactness property, available in the following.

Theorem 2.3.4 (Exactness property). *Let* $u_{\text{true},h,\hat{\rho}}(t) \in V_h \otimes L^2_{\hat{\rho}}$ *be of rank* R *for* $t^n \leq t \leq t^{n+1}$*, so that* $u_{\text{true},h,\hat{\rho}}(t)$ *has a factorization* (2.4)*, i.e.*

$$u_{\text{true},h,\hat{\rho}}(t) = \bar{u}(t) + U(t)S(t)V(t)^\intercal.$$

Moreover, assume that the $R \times R$ *matrix* $\mathbb{E}_{\hat{\rho}}[V(t^{n+1})^\intercal V(t^n)]$ *is invertible. With* $u^n_{h,\hat{\rho}} =$

$u_{\text{true},h,\hat\rho}(t^n)$, and $\triangle A = u_{\text{true},h,\hat\rho}(t^{n+1}) - u_{\text{true},h,\hat\rho}(t^n)$, the Algorithm 2.1.2 is exact: $\hat u_{h,\hat\rho}^{n+1} = u_{\text{true},h,\hat\rho}(t^{n+1})$.

Proof. By assumption, $\mathbb{E}_{\hat\rho}[V(t^{n+1})^\intercal V(t^n)]$ is invertible, therefore the range of $u_{\text{true},h,\hat\rho}(t^{n+1})$ lies in the range of the updated deterministic modes U_1 given by the integrator. This implies that

$$\mathcal{P}_{U_1}[u_{\text{true},h,\hat\rho}(t^{n+1})] = u_{\text{true},h,\hat\rho}(t^{n+1}) \quad \text{which gives us} \quad u_{\text{true},h,\hat\rho}(t^{n+1}) \in \mathcal{T}_{U_1 Y^{n\intercal}}\mathcal{M}_R^{h,\hat\rho}.$$

Applying Lemma 2.3.2 and the variational formulation (2.43), we derive

$$\begin{aligned}\hat u_{h,\hat\rho}^{n+1} &= \hat u_{h,\hat\rho}^{n} + \bar u_{\text{true},h,\hat\rho}(t^{n+1}) - \bar u_{\text{true},h,\hat\rho}(t^n) + \Pi^{h,\hat\rho}_{U_1 Y^{n\intercal}}[u^*_{\text{true},h,\hat\rho}(t^{n+1}) - u^*_{\text{true},h,\hat\rho}(t^n)]\\ &= u_{\text{true},h,\hat\rho}(t^n) + u_{\text{true},h,\hat\rho}(t^{n+1}) - u_{\text{true},h,\hat\rho}(t^n) = u_{\text{true},h,\hat\rho}(t^{n+1}).\end{aligned}$$

□

Note that the exactness property holds for the discrete DLR solution obtained by Algorithm 2.2.1 as well.

3 Stability properties

The main goal of this chapter is to prove the stability of the newly-proposed numerical schemes from Section 2.2 applied to a parabolic problem with random coefficients. The stability of the implicit and explicit Euler schemes applied to deterministic parabolic problems (with no DLR approximation) is well analyzed (see e.g. [EG04b]). A natural question is to what extent constraining the dynamics to the low rank manifold influences the stability properties. We will start this chapter by specifying the operator $\mathcal{F}$ from (1.3) to describe parabolic problems with random coefficients in Section 3.1. In Section 3.2, we will first recall some stability properties of the true solution u_{true} of problem (1.3). Then, in Section 3.3 we will see that these properties hold for the continuous DLR solution as well. It turns out that our discretization schemes satisfy analogous stability properties, as we will see in Section 3.4, under some stability conditions for the explicit and semi-implicit scheme. The sharpness of the obtained stability conditions on the time step and spatial discretization is supported by the numerical results provided in Section 3.6. In the rest of this chapter we will assume that a solution of problem (1.3), continuous DLR solution and discrete DLR solution exist. The results presented in this chapter are original and based on the paper [KNV21].

3.1 Problem specification

In this section we will introduce a random operator $\mathcal{L}$ particularizing the operator $\mathcal{F}$ from (1.3) to comprise parabolic equations.

Let us consider a random operator L with values in the space of linear bounded operators from V to V' that is uniformly bounded and coercive, i.e. a Borel measurable function

$$\begin{array}{rlcl} L: & \Omega & \rightarrow & \mathfrak{L}(V, V') \\ & \omega & \mapsto & L(\omega) \end{array}$$

such that there exist $C_{\mathcal{L}}, C_{\mathcal{B}} > 0$ satisfying

$$\Big(L(\omega)v, v\Big)_{V'V} \geq C_{\mathcal{L}}\|v\|_V^2 \qquad \forall \omega \in \Omega,\ \forall v \in V, \tag{3.1}$$

$$\Big(L(\omega)v, w\Big)_{V'V} \leq C_{\mathcal{B}}\|v\|_V\|w\|_V \qquad \forall \omega \in \Omega,\ \forall v, w \in V. \tag{3.2}$$

Associated to the random operator L, we introduce the operator $\mathcal{L}$, defined as

$$\begin{array}{llll} \mathcal{L}: & L^2_\rho(\Omega; V) & \to & L^2_\rho(\Omega; V') \\ & u & \mapsto & \mathcal{L}(u): \quad \mathcal{L}(u)(\omega) = L(\omega)u(\omega) \in V' \quad \forall \omega \in \Omega. \end{array}$$

Notice that for any strongly measurable $u : \Omega \to V$, the map $\omega \in \Omega \mapsto L(\omega)u(\omega) \in V'$ is strongly measurable, V' being separable (see [KNV21, Proposition A]). From the uniform boundedness of L it follows immediately that, if u is square integrable, then $\mathcal{L}(u)$ is square integrable as well and $\|\mathcal{L}(u)\|_{L^2_\rho(\Omega;V')} \leq C_{\mathcal{B}}\|u\|_{L^2_\rho(\Omega;V)}$, $\forall u \in L^2_\rho(\Omega; V)$. The operator $\mathcal{L}$ induces a bilinear form on $L^2_\rho(\Omega; V)$ defined as

$$\langle v, w\rangle_{\mathcal{L},\rho} := \int_\Omega \Big(\mathcal{L}(v)(\omega), w(\omega)\Big)_{V'V} \mathrm{d}\rho(\omega), \qquad v, w \in L^2_\rho(\Omega; V),$$

which is coercive and bounded with coercivity and continuity constant $C_{\mathcal{L}}$ and $C_{\mathcal{B}}$, respectively, i.e.

$$\begin{aligned} \langle v, v\rangle_{\mathcal{L},\rho} &\geq C_{\mathcal{L}}\|v\|^2_{V,L^2_\rho}, \\ \langle u, v\rangle_{\mathcal{L},\rho} &\leq C_{\mathcal{B}}\|u\|_{V,L^2_\rho}\|v\|_{V,L^2_\rho}. \end{aligned}$$

Then, given a final time $T > 0$, a random forcing term $f \in L^2(0, T; L^2_\rho(\Omega; H))$ and a random initial condition $u_0 \in L^2_\rho(\Omega; V)$, we consider now the following parabolic problem: Find a solution $u_{\text{true}} \in L^2(0, T; L^2_\rho(\Omega; V))$ with $\dot{u}_{\text{true}} \in L^2(0, T; L^2_\rho(\Omega; V'))$ satisfying

$$\begin{aligned} \Big(\dot{u}_{\text{true}}, v\Big)_{V'V,L^2_\rho} + \Big(\mathcal{L}(u_{\text{true}}), v\Big)_{V'V,L^2_\rho} &= \langle f, v\rangle_{H,L^2_\rho}, \\ &\forall v \in L^2_\rho(\Omega; V),\ \text{a.e. } t \in (0, T] \\ u_{\text{true}}(0) &= u_0. \end{aligned} \tag{3.3}$$

The general theory of parabolic equations (see e.g. [Wlo87]) can be applied to problem (3.3), at least in the case of $L^2_\rho(\Omega; V), L^2_\rho(\Omega; H), L^2_\rho(\Omega; V')$ being separable, e.g. when Ω is a Polish space and $\mathcal{A}$ is the corresponding Borel σ-algebra. We conclude then that problem (3.3) has a unique solution u_{true} which depends continuously on f and u_0. We note that the theory of parabolic equations would allow for less regular data $f \in L^2(0, T; L^2_\rho(\Omega; V'))$ and $u_0 \in L^2_\rho(\Omega; H)$. However, in this work we restrict our attention to the case $f \in L^2(0, T; L^2_\rho(\Omega; H))$, $u_0 \in L^2_\rho(\Omega; V)$.

Concerning the discretization proposed in Chapter 2, note that the semi-discrete bilinear form $\langle\cdot,\cdot\rangle_{\mathcal{L},\hat{\rho}}$ defined as

$$\langle v, w\rangle_{\mathcal{L},\hat{\rho}} = \sum_{k=1}^{\hat{N}} L(\omega_k)v(\omega_k)w(\omega_k)\lambda_k$$

is coercive and bounded, with the same coercivity and continuity constants $C_{\mathcal{L}}$, $C_{\mathcal{B}}$, defined in (3.1), (3.2), respectively.

We will state two types of estimates: the first one holds for an operator $\mathcal{L}$ as described in this section and a second one additionally assuming the operator $\mathcal{L}$ to be symmetric. Note that in the second case the bilinear coercive form $\langle\cdot,\cdot\rangle_{\mathcal{L},\rho}$ is a scalar product on $L^2_\rho(\Omega; V)$.

3.2 Stability of the continuous problem

We state here some standard stability estimates concerning the solution u_{true} of problem (3.3).

Proposition 3.2.1. Let $u_{\text{true}} \in L^2(0,T; L^2_\rho(\Omega; V))$ be the solution of problem (3.3). Then, the following estimates hold:

1.
$$\begin{aligned}\|u_{\text{true}}(T)\|^2_{H,L^2_\rho} + C_{\mathcal{L}}\int_0^T \|u_{\text{true}}(t)\|^2_{V,L^2_\rho}\,\mathrm{d}t \\ \leq \|u_{\text{true}}(0)\|^2_{H,L^2_\rho} + \frac{C_{\mathrm{P}}^2}{C_{\mathcal{L}}}\left\|f\right\|^2_{L^2(0,T;L^2_\rho(\Omega;H))};\end{aligned} \tag{3.4}$$

2. if, in addition, $\mathcal{L}$ is symmetric and $\dot{u}_{\text{true}} \in L^2(0,T; L^2_\rho(\Omega; H))$, we have
$$\begin{aligned}\|u_{\text{true}}(T)\|^2_{\mathcal{L},\rho} + \int_0^T \|\dot{u}_{\text{true}}(t)\|^2_{H,L^2_\rho}\,\mathrm{d}t \\ \leq \|u_{\text{true}}(0)\|^2_{\mathcal{L},\rho} + \left\|f\right\|^2_{L^2(0,T;L^2_\rho(\Omega;H))},\end{aligned} \tag{3.5}$$

where $C_{\mathcal{L}} > 0$ is the coercivity constant defined in (3.1) and C_{P} is the continuous embedding constant defined in (1.2).

For $f = 0$ and $t_1, t_2 \in [0,T],\ t_1 \leq t_2$, we have:

3.
$$\|u_{\text{true}}(t_2)\|_{H,L^2_\rho} \leq \|u_{\text{true}}(t_1)\|_{H,L^2_\rho}, \tag{3.6}$$

4. moreover, if $\mathcal{L}$ is symmetric and $\dot{u}_{\text{true}} \in L^2(0,T;L^2_\rho(\Omega;H))$, we have
$$\|u_{\text{true}}(t_2)\|_{\mathcal{L},\rho} \leq \|u_{\text{true}}(t_1)\|_{\mathcal{L},\rho}. \tag{3.7}$$

Proof. As for part 1, choose u_{true} as a test function in the variational formulation (3.3). Using [Zei90, Prop. 23.23] results in

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u_{\text{true}}\|^2_{H,L^2_\rho} + \langle u_{\text{true}}, u_{\text{true}}\rangle_{\mathcal{L},\rho} = \langle f, u_{\text{true}}\rangle_{H,L^2_\rho} &\leq C_{\mathrm{P}}\|f\|_{H,L^2_\rho}\|u_{\text{true}}\|_{V,L^2_\rho} \\ &\leq \frac{C_{\mathrm{P}}^2}{2C_{\mathcal{L}}}\|f\|^2_{H,L^2_\rho} + \frac{C_{\mathcal{L}}}{2}\|u_{\text{true}}\|^2_{V,L^2_\rho} \quad \text{for a.e. } t \in (0,T].\end{aligned}$$

Multiplying by 2 and integrating over $[0,T]$ gives the sought estimate. Part 2. is proved in a similar way by considering $\dot{u}_{\text{true}}$ as a test function. We can derive

$$\begin{aligned}\|\dot{u}_{\text{true}}\|^2_{H,L^2_\rho} + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u_{\text{true}}\|^2_{\mathcal{L},\rho} = \langle f, \dot{u}_{\text{true}}\rangle_{H,L^2_\rho} &\leq \|f\|_{H,L^2_\rho}\|\dot{u}_{\text{true}}\|_{H,L^2_\rho} \\ &\leq \frac{\|f\|^2_{H,L^2_\rho}}{2} + \frac{\|\dot{u}_{\text{true}}\|^2_{H,L^2_\rho}}{2}\end{aligned}$$

and obtain the result by multiplying by 2 and integrating over $[0,T]$.

Part 3. and part 4. are consequences of part 1. and 2., where the final integration is realized over $[t_1,t_2]$ instead of $[0,T]$. ☐

3.3 Stability of the continuous DLR solution

Constraining the dynamics to the R-rank manifold does not destroy the stability properties from Proposition 3.2.1.

Theorem 3.3.1. *Let $u \in L^2(0,T;L^2_\rho(\Omega;V))$ with $\dot{u} \in L^2(0,T;L^2_\rho(\Omega;V))$ be the continuous DLR solution defined in Definition 1.2.1. Then u satisfies the same inequalities* (3.4)*,* (3.5)*,* (3.6)*,* (3.7) *as the true solution u_{true}.*

Proof. Part 1: Let $u = \bar{u} + UY^\intercal$ with $UY^\intercal \in \mathcal{M}_R$. Then, we have $u^* = u - \bar{u} \in \mathcal{T}_{u^*}\mathcal{M}_R$. Indeed, since
$$u^* = \sum_{i=1}^{R} U_i 0 + U_i Y_i \quad \in L^2_{\rho,0}(\Omega;V)$$
with $\langle 0, Y_i\rangle_{L^2_\rho} = 0$, we can take u as a test function in the variational formulation (1.17). The rest of the proof follows the same steps as in the proof of Proposition 3.2.1.

Part 2: we express

$$\dot{u}^* = \sum_{j=1}^{R} \dot{U}_j Y_j + U_j \dot{Y}_j \quad \in \mathcal{T}_{u^*}\mathcal{M}_R$$

since $\langle Y_i, \dot{Y}_j \rangle_{L^2_\rho} = \delta_{ij}$ and $\dot{u}^* \in L^2_\rho(\Omega; V)$. As $\dot{\bar{u}} \in V$ we can consider $\dot{u}$ as a test function in the variational formulation (1.17) and arrive at the sought result.

Part 3. and 4. is obtained analogously. $\square$

3.4 Stability of the discrete DLR solution

Now we proceed with showing stability properties of the fully discretized DLR system from Algorithm 2.2.1 for the three different operator evaluation terms corresponding to implicit Euler, explicit Euler and semi-implicit scheme. For each of them we will establish boundedness of norms and a decrease of norms for the case of zero forcing term f.

The following simple lemma will be repeatedly used throughout.

Lemma 3.4.1. *Let $\langle \cdot, \cdot \rangle : (V_h \otimes L^2_{\hat{\rho}}) \times (V_h \otimes L^2_{\hat{\rho}}) \to \mathbb{R}$ be a symmetric bilinear form. Then it holds*

$$\begin{aligned}
\langle v, w - v \rangle &= \frac{1}{2}\Big(\langle w, w \rangle - \langle v, v \rangle - \langle w - v, w - v \rangle \Big) \\
\langle w, w - v \rangle &= \frac{1}{2}\Big(\langle w, w \rangle - \langle v, v \rangle + \langle w - v, w - v \rangle \Big) \\
\langle v, w + v \rangle &= \frac{1}{2}\Big(\langle v, v \rangle - \langle w, w \rangle + \langle w + v, w + v \rangle \Big)
\end{aligned}$$

for any $v, w \in V_h \otimes L^2_{\hat{\rho}}$.

3.4.1 Implicit Euler scheme

Applying an implicit operator evaluation, i.e. $\mathcal{L}(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}}) = \mathcal{L}(u^{n+1}_{h,\hat{\rho}})$ results in a discretization scheme with the following stability properties.

Theorem 3.4.2. *Let $\{u^n_{h,\hat{\rho}}\}_{n=0}^N$ be the discrete DLR solution as defined in Algorithm 2.2.1 with $\mathcal{L}(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}}) = \mathcal{L}(u^{n+1}_{h,\hat{\rho}})$. Then the following estimates hold:*

1.

$$\|u^N_{h,\hat{\rho}}\|^2_{H,L^2_{\hat{\rho}}} + \triangle t C_{\mathcal{L}} \sum_{n=0}^{N-1} \|u^{n+1}_{h,\hat{\rho}}\|^2_{V,L^2_{\hat{\rho}}} \leq \|u^0_{h,\hat{\rho}}\|^2_{H,L^2_{\hat{\rho}}} + \triangle t \frac{C_{\mathrm{P}}^2}{C_{\mathcal{L}}} \sum_{n=0}^{N-1} \|f(t_{n+1})\|^2_{H,L^2_{\hat{\rho}}},$$

2. *if $\mathcal{L}$ is a symmetric operator we have*

$$\|u_{h,\hat\rho}^N\|^2_{\mathcal{L},\hat\rho} + \triangle t \sum_{n=0}^{N-1} \left\| \frac{u_{h,\hat\rho}^{n+1} - u_{h,\hat\rho}^n}{\triangle t} \right\|^2_{H,L^2_{\hat\rho}} \leq \|u_{h,\hat\rho}^0\|^2_{\mathcal{L},\hat\rho} + \triangle t \sum_{n=0}^{N-1} \|f(t_{n+1})\|^2_{H,L^2_{\hat\rho}},$$

for any time and space discretization parameters $\triangle t$, $h > 0$ with $C_{\mathcal{L}}, C_{\mathrm{P}} > 0$ the coercivity and continuous embedding constant defined in (3.1)*,* (1.2)*, respectively.*

In particular, for $f = 0$ and $n = 0, \dots, N-1$ it holds:

3. $\|u_{h,\hat\rho}^{n+1}\|_{H,L^2_{\hat\rho}} \leq \|u_{h,\hat\rho}^n\|_{H,L^2_{\hat\rho}}$,
4. *if $\mathcal{L}$ is a symmetric operator we have* $\|u_{h,\hat\rho}^{n+1}\|_{\mathcal{L},\hat\rho} \leq \|u_{h,\hat\rho}^n\|_{\mathcal{L},\hat\rho}$.

Proof. Thanks to Theorem 2.2.7, we know that the discretized DLR system of equations with implicit operator evaluation can be written in a variational formulation as

$$\begin{aligned} &\Big\langle \frac{u_{h,\hat\rho}^{n+1} - u_{h,\hat\rho}^n}{\triangle t}, v_h \Big\rangle_{H,L^2_{\hat\rho}} + \Big\langle u_{h,\hat\rho}^{n+1}, v_h \Big\rangle_{\mathcal{L},\hat\rho} = \Big\langle f(t_{n+1}), v_h \Big\rangle_{H,L^2_{\hat\rho}}, \\ &\qquad\qquad \forall v_h = \bar v_h + v_h^* \text{ with } \bar v_h \in V_h \text{ and } v_h^* \in \mathcal{T}_{\tilde U^{n+1} Y^n} \mathcal{M}_R^{h,\hat\rho}, \end{aligned} \tag{3.8}$$

$n = 0, \dots, N-1$.

1. Based on Lemma 2.2.5 we take $v_h = u_{h,\hat\rho}^{n+1}$ as a test function in the variational formulation (3.8). Using Lemma 3.4.1 results in

   $$\begin{aligned} &\|u_{h,\hat\rho}^{n+1}\|^2_{H,L^2_{\hat\rho}} - \|u_{h,\hat\rho}^n\|^2_{H,L^2_{\hat\rho}} + \|u_{h,\hat\rho}^{n+1} - u_{h,\hat\rho}^n\|^2_{H,L^2_{\hat\rho}} + 2\triangle t \langle u_{h,\hat\rho}^{n+1}, u_{h,\hat\rho}^{n+1}\rangle_{\mathcal{L},\hat\rho} \\ &\quad = 2\triangle t (f(t_{n+1}), u_{h,\hat\rho}^{n+1})_{V'V,L^2_{\hat\rho}} \leq \triangle t \frac{C_{\mathrm{P}}^2}{C_{\mathcal{L}}} \|f(t_{n+1})\|^2_{H,L^2_{\hat\rho}} + \triangle t C_{\mathcal{L}} \|u_{h,\hat\rho}^{n+1}\|^2_{V,L^2_{\hat\rho}}. \end{aligned}$$

   Using the coercivity condition (3.1) and summing over $n = 0, \dots, N-1$ gives us the sought result.

2. Now, consider $v_h = (u_{h,\hat\rho}^{n+1} - u_{h,\hat\rho}^n)/\triangle t$. Using Lemma 3.4.1, the variational formulation results in

   $$\begin{aligned} &\left\| \frac{u_{h,\hat\rho}^{n+1} - u_{h,\hat\rho}^n}{\triangle t} \right\|^2_{H,L^2_{\hat\rho}} + \frac{1}{2\triangle t}\Big(\|u_{h,\hat\rho}^{n+1}\|^2_{\mathcal{L},\hat\rho} - \|u_{h,\hat\rho}^n\|^2_{\mathcal{L},\hat\rho} + \|u_{h,\hat\rho}^{n+1} - u_{h,\hat\rho}^n\|^2_{\mathcal{L},\hat\rho} \Big) \\ &\quad = \Big\langle f(t_{n+1}), \frac{u_{h,\hat\rho}^{n+1} - u_{h,\hat\rho}^n}{\triangle t} \Big\rangle_{H,L^2_{\hat\rho}} \leq \frac{\|f(t_{n+1})\|^2_{H,L^2_{\hat\rho}}}{2} + \frac{1}{2} \left\| \frac{u_{h,\hat\rho}^{n+1} - u_{h,\hat\rho}^n}{\triangle t} \right\|^2_{H,L^2_{\hat\rho}}. \end{aligned}$$

   Multiplying by $2\triangle t$ and summing over $n = 0, \dots, N-1$ leads us to the result.

Parts 3. and 4. follow from part 1. and 2. without summing over $n = 0, \dots, N-1$. □

3.4.2 Explicit Euler scheme

Concerning the explicit Euler scheme (see subsection 2.2.1), which applies the time discretization $\mathcal{L}(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}}) = \mathcal{L}(u^n_{h,\hat{\rho}})$, the following stability result holds.

Theorem 3.4.3. *Let $\{u^n_{h,\hat{\rho}}\}^N_{n=0}$ be the discrete DLR solution as defined in Algorithm 2.2.1 with $\mathcal{L}(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}}) = \mathcal{L}(u^n_{h,\hat{\rho}})$. Then the following estimates hold:*

1.

$$\|u^N_{h,\hat{\rho}}\|^2_{H,L^2_{\hat{\rho}}} + \triangle t C_{\mathcal{L}}(1-\kappa) \sum_{n=0}^{N-1} \|u^{n+1}_{h,\hat{\rho}}\|^2_{V,L^2_{\hat{\rho}}} \leq \|u^0_{h,\hat{\rho}}\|^2_{H,L^2_{\hat{\rho}}} + \frac{\triangle t C_{\mathrm{P}}^2}{C_{\mathcal{L}}} \sum_{n=0}^{N-1} \|f(t_n)\|^2_{H,L^2_{\hat{\rho}}}$$

for $0 < \kappa$ and $\triangle t, h$ satisfying

$$\frac{\triangle t}{h^{2p}} \leq \frac{\kappa\, C_{\mathcal{L}}}{C_{\mathrm{I}}^2\, C_{\mathcal{B}}^2}. \tag{3.9}$$

2. If $\mathcal{L}$ is a symmetric operator we have

$$\|u^N_{h,\hat{\rho}}\|_{\mathcal{L},\hat{\rho}} \leq \|u^0_{h,\hat{\rho}}\|_{\mathcal{L},\hat{\rho}} + \frac{\triangle t}{\kappa} \sum_{n=0}^{N-1} \|f(t_n)\|^2_{H,L^2_{\hat{\rho}}},$$

for $\triangle t, h$ satisfying

$$\frac{\triangle t}{h^{2p}} \leq \frac{2-\kappa}{C_{\mathrm{I}}^2 C_{\mathcal{B}}} \quad \text{with } 0 < \kappa < 2. \tag{3.10}$$

Here $C_{\mathcal{L}}, C_{\mathcal{B}}, C_{\mathrm{P}} > 0$ are the coercivity, continuity and continuous embedding constants defined in (3.1), (3.2), (1.2), *respectively and C_{I} is the inverse inequality constant introduced in* (2.2).

For $f = 0$ and $n = 0, \dots, N-1$ it holds:

3. $\|u^{n+1}_{h,\hat{\rho}}\|_{H,L^2_{\hat{\rho}}} \leq \|u^n_{h,\hat{\rho}}\|_{H,L^2_{\hat{\rho}}}$,

under a weakened condition $\quad \dfrac{\triangle t}{h^{2p}} \leq \dfrac{2C_{\mathcal{L}}}{C_{\mathrm{I}}^2 C_{\mathcal{B}}^2}.$

4. *If $\mathcal{L}$ is a symmetric operator we have*

$$\|u^{n+1}_{h,\hat\rho}\|_{\mathcal{L},\hat\rho} \le \|u^{n}_{h,\hat\rho}\|_{\mathcal{L},\hat\rho},$$
$$\|u^{n+1}_{h,\hat\rho}\|_{H,L^2_{\hat\rho}} \le \|u^{n}_{h,\hat\rho}\|_{H,L^2_{\hat\rho}},$$

under a weakened condition

$$\frac{\triangle t}{h^{2p}} \le \frac{2}{C_{\mathrm{I}}^2 C_{\mathcal{B}}}.$$

Proof. Thanks to the Theorem 2.2.7 we can rewrite the system of equations in the variational formulation

$$\begin{gathered}\Big\langle \frac{u^{n+1}_{h,\hat\rho} - u^{n}_{h,\hat\rho}}{\triangle t},\, v_h \Big\rangle_{H,L^2_{\hat\rho}} + \Big\langle u^{n}_{h,\hat\rho},\, v_h \Big\rangle_{\mathcal{L},\hat\rho} = \Big\langle f(t_n),\, v_h \Big\rangle_{H,L^2_{\hat\rho}}, \\ \forall v_h = \bar v_h + v_h^* \text{ with } \bar v_h \in V_h \text{ and } v_h^* \in \mathcal{T}_{\tilde U^{n+1} Y^n} \mathcal{M}_R^{h,\hat\rho}.\end{gathered} \tag{3.11}$$

1. Based on Lemma 2.2.5 we take $v_h = u^{n+1}_{h,\hat\rho}$ as a test function in the variational formulation (3.11) and using Lemma 3.4.1 results in

$$\begin{aligned}&\|u^{n+1}_{h,\hat\rho}\|^2_{H,L^2_{\hat\rho}} - \|u^{n}_{h,\hat\rho}\|^2_{H,L^2_{\hat\rho}} + \|u^{n+1}_{h,\hat\rho} - u^{n}_{h,\hat\rho}\|^2_{H,L^2_{\hat\rho}} + 2\triangle t \langle u^{n}_{h,\hat\rho}, u^{n+1}_{h,\hat\rho}\rangle_{\mathcal{L},\hat\rho} \\ &\qquad = 2\triangle t (f(t_n), u^{n+1}_{h,\hat\rho})_{V'V,L^2_{\hat\rho}} \le \triangle t \frac{C_{\mathrm{P}}^2}{C_{\mathcal{L}}} \|f(t_n)\|^2_{H,L^2_{\hat\rho}} + \triangle t C_{\mathcal{L}} \|u^{n+1}_{h,\hat\rho}\|_{V,L^2_{\hat\rho}}.\end{aligned}$$

   We further proceed by estimating

$$\begin{aligned}&2\triangle t \langle u^{n}_{h,\hat\rho}, u^{n+1}_{h,\hat\rho}\rangle_{\mathcal{L},\hat\rho} = 2\triangle t \langle u^{n}_{h,\hat\rho} - u^{n+1}_{h,\hat\rho}, u^{n+1}_{h,\hat\rho}\rangle_{\mathcal{L},\hat\rho} + 2\triangle t \langle u^{n+1}_{h,\hat\rho}, u^{n+1}_{h,\hat\rho}\rangle_{\mathcal{L},\hat\rho} \\ &\ge -2\triangle t C_{\mathcal{B}} \|u^{n+1}_{h,\hat\rho} - u^{n}_{h,\hat\rho}\|_{V,L^2_{\hat\rho}} \|u^{n+1}_{h,\hat\rho}\|_{V,L^2_{\hat\rho}} + 2\triangle t C_{\mathcal{L}} \|u^{n+1}_{h,\hat\rho}\|^2_{V,L^2_{\hat\rho}} \\ &\ge -\kappa \triangle t C_{\mathcal{L}} \|u^{n+1}_{h,\hat\rho}\|^2_{V,L^2_{\hat\rho}} + 2\triangle t C_{\mathcal{L}} \|u^{n+1}_{h,\hat\rho}\|^2_{V,L^2_{\hat\rho}} - \triangle t \frac{C_{\mathrm{I}}^2 C_{\mathcal{B}}^2}{\kappa\, h^{2p} C_{\mathcal{L}}} \|u^{n+1}_{h,\hat\rho} - u^{n}_{h,\hat\rho}\|^2_{H,L^2_{\hat\rho}}\end{aligned} \tag{3.12}$$

   where, in the third step, we used the inequality

$$\|u^{n+1}_{h,\hat\rho} - u^{n}_{h,\hat\rho}\|_{H,L^2_{\hat\rho}} \ge \frac{h^p}{C_{\mathrm{I}}} \|u^{n+1}_{h,\hat\rho} - u^{n}_{h,\hat\rho}\|_{V,L^2_{\hat\rho}},$$

   which holds based on assumption (2.2). Combining the terms, using the condition (3.9) and summing over $n = 0, \ldots, N-1$ finishes the proof.

2. Lemma 2.2.5 enables us to take $u^{n+1}_{h,\hat\rho} - u^{n}_{h,\hat\rho}$ as a test function in (3.11). This results

in

$$\begin{aligned}\frac{1}{\triangle t}\|u^{n+1}_{h,\hat\rho}-u^{n}_{h,\hat\rho}\|^2_{H,L^2_{\hat\rho}}+\langle u^{n}_{h,\hat\rho},u^{n+1}_{h,\hat\rho}-u^{n}_{h,\hat\rho}\rangle_{\mathcal{L},\hat\rho}&=\langle f(t_n),u^{n+1}_{h,\hat\rho}-u^{n}_{h,\hat\rho}\rangle_{H,L^2_{\hat\rho}}\\&\le\frac{\triangle t\|f(t_n)\|^2_{H,L^2_{\hat\rho}}}{2\kappa}+\frac{\kappa\,\|u^{n+1}_{h,\hat\rho}-u^{n}_{h,\hat\rho}\|^2_{H,L^2_{\hat\rho}}}{2\triangle t}.\end{aligned}\tag{3.13}$$

Using Lemma 3.4.1 we obtain

$$\begin{aligned}\|u^{n+1}_{h,\hat\rho}\|^2_{\mathcal{L},\hat\rho}&\le\|u^{n}_{h,\hat\rho}\|^2_{\mathcal{L},\hat\rho}+\frac{\triangle t}{\kappa}\|f(t_n)\|^2_{H,L^2_{\hat\rho}}+\|u^{n+1}_{h,\hat\rho}-u^{n}_{h,\hat\rho}\|^2_{\mathcal{L},\hat\rho}\\&\quad-\frac{2-\kappa}{\triangle t}\|u^{n+1}_{h,\hat\rho}-u^{n}_{h,\hat\rho}\|^2_{H,L^2_{\hat\rho}}\\&\le\|u^{n}_{h,\hat\rho}\|^2_{\mathcal{L},\hat\rho}+\frac{\triangle t}{\kappa}\|f(t_n)\|^2_{H,L^2_{\hat\rho}}+\left(1-\frac{(2-\kappa)\,h^{2p}}{C_{\mathrm{I}}^2C_{\mathcal{B}}\triangle t}\right)\|u^{n+1}_{h,\hat\rho}-u^{n}_{h,\hat\rho}\|^2_{\mathcal{L},\hat\rho}\\&\le\|u^{n}_{h,\hat\rho}\|^2_{\mathcal{L},\hat\rho}+\frac{\triangle t}{\kappa}\|f(t_n)\|^2_{H,L^2_{\hat\rho}}\end{aligned}$$

where, in the second step, we used the assumption (2.2), (3.2) and the fact that $\left(1-\frac{(2-\kappa)h^{2p}}{C_{\mathrm{I}}^2C_{\mathcal{B}}\triangle t}\right)\le 0$, thanks to the stability condition (3.10).

3. The proof of part 3. follows the same steps as the proof of Part 1. We have

$$\|u^{n+1}_{h,\hat\rho}\|^2_{H,L^2_{\hat\rho}}-\|u^{n}_{h,\hat\rho}\|^2_{H,L^2_{\hat\rho}}+\|u^{n+1}_{h,\hat\rho}-u^{n}_{h,\hat\rho}\|^2_{H,L^2_{\hat\rho}}+2\triangle t\langle u^{n}_{h,\hat\rho},u^{n+1}_{h,\hat\rho}\rangle_{\mathcal{L},\hat\rho}=0.$$

In (3.12) we choose $\kappa=2$ and conclude the result.

4. The proof of the forth property follows the same steps as the proof of Part 2. Since there is no need to use the Young's inequality in (3.13), the condition on $\triangle t/h^{2p}$ is weakened:

$$\begin{aligned}\|u^{n+1}_{h,\hat\rho}\|^2_{\mathcal{L},\hat\rho}&=\|u^{n}_{h,\hat\rho}\|^2_{\mathcal{L},\hat\rho}+\|u^{n+1}_{h,\hat\rho}-u^{n}_{h,\hat\rho}\|^2_{\mathcal{L},\hat\rho}-\frac{2}{\triangle t}\|u^{n+1}_{h,\hat\rho}-u^{n}_{h,\hat\rho}\|^2_{H,L^2_{\hat\rho}}\\&\le\|u^{n}_{h,\hat\rho}\|^2_{\mathcal{L},\hat\rho}+\left(1-\frac{2h^{2p}}{C_{\mathrm{I}}^2C_{\mathcal{B}}\triangle t}\right)\|u^{n+1}_{h,\hat\rho}-u^{n}_{h,\hat\rho}\|^2_{\mathcal{L},\hat\rho}.\end{aligned}$$

As for the estimate in the $\|\cdot\|_{H,L^2_{\hat\rho}}$-norm we can derive

$$\begin{aligned}\|u^{n+1}_{h,\hat\rho}\|^2_{H,L^2_{\hat\rho}}&=\|u^{n}_{h,\hat\rho}\|^2_{H,L^2_{\hat\rho}}-\frac{\triangle t}{2}\Big(\|u^{n}_{h,\hat\rho}\|^2_{\mathcal{L},\hat\rho}-\|u^{n+1}_{h,\hat\rho}\|^2_{\mathcal{L},\hat\rho}\\&\qquad\qquad+\|u^{n+1}_{h,\hat\rho}+u^{n}_{h,\hat\rho}\|^2_{\mathcal{L},\hat\rho}\Big)\\&\le\|u^{n}_{h,\hat\rho}\|^2_{H,L^2_{\hat\rho}},\end{aligned}$$

where in the last inequality we applied $\|u^{n+1}_{h,\hat\rho}\|_{\mathcal{L},\hat\rho}\le\|u^{n}_{h,\hat\rho}\|_{\mathcal{L},\hat\rho}$ for $\frac{\triangle t}{h^{2p}}\le\frac{2}{C_{\mathrm{I}}^2C_{\mathcal{B}}}$.

□

3.4.3 Semi-implicit scheme

This subsection is dedicated to analyzing the semi-implicit scheme introduced in subsection 2.2.1 which applies the discretization $\mathcal{L}(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}}) = \mathcal{L}_{\text{det}}(u^{n+1}_{h,\hat{\rho}}) + \mathcal{L}_{\text{stoch}}(u^n_{h,\hat{\rho}})$.

Apart from the inverse inequality (2.2) we will be using two additional inequalities. Let us assume there exists a constant $C_{\text{det}} > 0$ such that

$$\langle u, u\rangle_{\mathcal{L}_{\text{det}},\hat{\rho}} \geq C_{\text{det}}\, \langle u, u\rangle_{\mathcal{L},\hat{\rho}}, \qquad \forall u \in V_h \otimes L^2_{\hat{\rho}}. \tag{3.14}$$

This constant plays an important role in the stability estimation as it quantifies the extent to which the operator is evaluated implicitly. Its significance is summarized in Theorem 3.4.4. In addition we introduce a constant C_{stoch} that bounds the stochasticity of the operator

$$|(\mathcal{L}_{\text{stoch}}(u), v)_{V'V,L^2_{\hat{\rho}}}| \leq C_{\text{stoch}} \|u\|_{V,L^2_{\hat{\rho}}} \|v\|_{V,L^2_{\hat{\rho}}} \tag{3.15}$$

Theorem 3.4.4. *Let $\{u^n_{h,\hat{\rho}}\}_{n=0}^N$ be the discrete DLR solution as defined in Algorithm 2.2.1 with $\mathcal{L}(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}}) = \mathcal{L}_{\text{det}}(u^{n+1}_{h,\hat{\rho}}) + \mathcal{L}_{\text{stoch}}(u^n_{h,\hat{\rho}})$ with $\mathcal{L}_{\text{det}}$ and $\mathcal{L}_{\text{stoch}}$ satisfying* (3.14) *and* (3.15)*, respectively. Then it holds*

1.

$$\|u^N_{h,\hat{\rho}}\|^2_{H,L^2_{\hat{\rho}}} + \triangle t C_{\mathcal{L}}(1-\kappa) \sum_{n=0}^{N-1} \|u^{n+1}_{h,\hat{\rho}}\|^2_{V,L^2_{\hat{\rho}}} \leq \|u^0_{h,\hat{\rho}}\|^2_{H,L^2_{\hat{\rho}}} + \frac{\triangle t C_{\text{P}}^2}{C_{\mathcal{L}}} \sum_{n=0}^{N-1} \|f^{n,n+1}\|^2_{H,L^2_{\hat{\rho}}}$$

for $\kappa > 0$ and $\triangle t, h$ satisfying

$$\frac{\triangle t}{h^{2p}} \leq \frac{\kappa\, C_{\mathcal{L}}}{C_{\text{I}}^2\, C_{\text{stoch}}^2}. \tag{3.16}$$

2. If $\mathcal{L}$ is a symmetric operator we have

$$\|u^N_{h,\hat{\rho}}\|_{\mathcal{L},\hat{\rho}} \leq \|u^0_{h,\hat{\rho}}\|_{\mathcal{L},\hat{\rho}} + \frac{\triangle t}{\kappa} \sum_{n=0}^{N-1} \|f^{n,n+1}\|^2_{H,L^2_{\hat{\rho}}} \tag{3.17}$$

for $\triangle t, h$ satisfying

$$\frac{\triangle t}{h^{2p}} \leq \begin{cases} +\infty & \text{if } C_{\text{det}} \geq \frac{1}{2} \\ \frac{2-\kappa}{C_{\text{I}}^2 C_{\mathcal{B}}(1-2C_{\text{det}})} & \text{if } C_{\text{det}} < \frac{1}{2} \end{cases}$$

Here $C_{\mathcal{L}}, C_{\mathcal{B}}, C_{\mathrm{P}}, C_{\mathrm{I}} > 0$ *are the coercivity, continuity, continuous embedding and inverse inequality constants defined in* (3.1), (3.2), (1.2), (2.2), *respectively. The constants* $C_{\mathrm{det}}, C_{\mathrm{stoch}}$ *were introduced in* (3.14), (3.15).

For $f = 0$ *and* $\mathcal{L}$ *symmetric we have*

3.

$$\|u_{h,\hat\rho}^{n+1}\|_{\mathcal{L},\hat\rho} \leq \|u_{h,\hat\rho}^{n}\|_{\mathcal{L},\hat\rho}, \qquad n = 0, \dots, N-1 \tag{3.18}$$

with $\triangle t, h$ *satisfying a weakened condition*

$$\frac{\triangle t}{h^{2p}} \leq \begin{cases} +\infty & \text{if } C_{\mathrm{det}} \geq \frac{1}{2} \\ \frac{2}{C_{\mathrm{I}}^2 C_{\mathcal{B}}(1-2C_{\mathrm{det}})} & \text{if } C_{\mathrm{det}} < \frac{1}{2} \end{cases} \tag{3.19}$$

Proof. The variational formulation of the discrete DLR problem from Algorithm 2.2.1 reads in this case

$$\begin{aligned} &\Big\langle \frac{u_{h,\hat\rho}^{n+1} - u_{h,\hat\rho}^{n}}{\triangle t}, v_h \Big\rangle_{H,L^2_{\hat\rho}} + \Big\langle u_{h,\hat\rho}^{n+1}, v_h \Big\rangle_{\mathcal{L}_{\mathrm{det}},\hat\rho} + \Big(\mathcal{L}_{\mathrm{stoch}}(u_{h,\hat\rho}^{n}), v_h\Big)_{V'V,L^2_{\hat\rho}} \\ &= \Big\langle f^{n,n+1}, v_h \Big\rangle_{H,L^2_{\hat\rho}} \quad \forall v_h = \bar v_h + v_h^* \text{ with } \bar v_h \in V_h \text{ and } v_h^* \in \mathcal{T}_{\tilde U^{n+1} Y^n} \mathcal{M}_R^{h,\hat\rho}. \end{aligned} \tag{3.20}$$

1. We will consider $v_h = u_{h,\hat\rho}^{n+1}$ as a test function in (3.20) and we derive

$$\begin{aligned} &\|u_{h,\hat\rho}^{n+1}\|^2_{H,L^2_{\hat\rho}} + 2\triangle t \langle u_{h,\hat\rho}^{n+1}, u_{h,\hat\rho}^{n+1}\rangle_{\mathcal{L},\hat\rho} \\ &= \|u_{h,\hat\rho}^{n}\|^2_{H,L^2_{\hat\rho}} - \|u_{h,\hat\rho}^{n+1} - u_{h,\hat\rho}^{n}\|^2_{H,L^2_{\hat\rho}} + 2\triangle t \langle f^{n,n+1}, u_{h,\hat\rho}^{n+1}\rangle_{H,L^2_{\hat\rho}} \\ &\quad + 2\triangle t (\mathcal{L}_{\mathrm{stoch}}(u_{h,\hat\rho}^{n+1} - u_{h,\hat\rho}^{n}), u_{h,\hat\rho}^{n+1})_{V'V,L^2_{\hat\rho}} \\ &\leq \|u_{h,\hat\rho}^{n}\|^2_{H,L^2_{\hat\rho}} - \|u_{h,\hat\rho}^{n+1} - u_{h,\hat\rho}^{n}\|^2_{H,L^2_{\hat\rho}} + \triangle t \frac{C_{\mathrm{P}}^2}{C_{\mathcal{L}}} \|f^{n,n+1}\|^2_{H,L^2_{\hat\rho}} + \triangle t C_{\mathcal{L}} \|u_{h,\hat\rho}^{n+1}\|^2_{V,L^2_{\hat\rho}} \\ &\quad + \kappa \triangle t C_{\mathcal{L}} \|u_{h,\hat\rho}^{n+1}\|^2_{V,L^2_{\hat\rho}} + \triangle t \frac{C_{\mathrm{I}}^2 C_{\mathrm{stoch}}^2}{\kappa h^{2p} C_{\mathcal{L}}} \|u_{h,\hat\rho}^{n+1} - u_{h,\hat\rho}^{n}\|^2_{H,L^2_{\hat\rho}}. \end{aligned}$$

   Combining the terms and summing over $n = 0, \dots, N-1$ finishes the proof.

2. We will proceed by taking $v_h = u_{h,\hat\rho}^{n+1} - u_{h,\hat\rho}^{n}$ in the variational formulation (3.20)

since $(u^{n+1}_{h,\hat{\rho}} - u^{n}_{h,\hat{\rho}})^* \in \mathcal{T}_{\tilde{U}^{n+1}Y^n}\mathcal{M}^{h,\hat{\rho}}_R$ (Lemma 2.2.5). We obtain

$$
\begin{aligned}
&\frac{1}{\triangle t}\|u^{n+1}_{h,\hat{\rho}} - u^{n}_{h,\hat{\rho}}\|^2_{H,L^2_{\hat{\rho}}} + \langle u^{n+1}_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}} - u^{n}_{h,\hat{\rho}}\rangle_{\mathcal{L}_{\det},\hat{\rho}} \qquad (3.21)\\
&\quad + \Big(\mathcal{L}_{\mathrm{stoch}}(u^{n}_{h,\hat{\rho}}), u^{n+1}_{h,\hat{\rho}} - u^{n}_{h,\hat{\rho}}\Big)_{V'V,L^2_{\hat{\rho}}} \pm \Big(\mathcal{L}_{\det}(u^{n}_{h,\hat{\rho}}), u^{n+1}_{h,\hat{\rho}} - u^{n}_{h,\hat{\rho}}\Big)_{V'V,L^2_{\hat{\rho}}}\\
&= \frac{1}{\triangle t}\|u^{n+1}_{h,\hat{\rho}} - u^{n}_{h,\hat{\rho}}\|^2_{H,L^2_{\hat{\rho}}} + \langle u^{n+1}_{h,\hat{\rho}} - u^{n}_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}} - u^{n}_{h,\hat{\rho}}\rangle_{\mathcal{L}_{\det},\hat{\rho}}\\
&\quad + \langle u^{n}_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}} - u^{n}_{h,\hat{\rho}}\rangle_{\mathcal{L},\hat{\rho}}\\
&= \langle f^{n,n+1}, u^{n+1}_{h,\hat{\rho}} - u^{n}_{h,\hat{\rho}}\rangle_{H,L^2_{\hat{\rho}}} \le \frac{\triangle t}{2\kappa}\|f^{n,n+1}\|^2_{H,L^2_{\hat{\rho}}} + \frac{\kappa}{2\triangle t}\|u^{n+1}_{h,\hat{\rho}} - u^{n}_{h,\hat{\rho}}\|^2_{H,L^2_{\hat{\rho}}}. \qquad (3.22)
\end{aligned}
$$

Using Lemma 3.4.1 we further derive

$$
\begin{aligned}
\|u^{n+1}_{h,\hat{\rho}}\|^2_{\mathcal{L},\hat{\rho}} \le{}& \|u^{n}_{h,\hat{\rho}}\|^2_{\mathcal{L},\hat{\rho}} + \frac{\triangle t}{\kappa}\|f^{n,n+1}\|^2_{H,L^2_{\hat{\rho}}} + \|u^{n+1}_{h,\hat{\rho}} - u^{n}_{h,\hat{\rho}}\|^2_{\mathcal{L},\hat{\rho}}\\
&- \frac{2-\kappa}{\triangle t}\|u^{n+1}_{h,\hat{\rho}} - u^{n}_{h,\hat{\rho}}\|^2_{H,L^2_{\hat{\rho}}}\\
&- 2\langle u^{n+1}_{h,\hat{\rho}} - u^{n}_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}} - u^{n}_{h,\hat{\rho}}\rangle_{\mathcal{L}_{\det},\hat{\rho}}\\
\le{}& \|u^{n}_{h,\hat{\rho}}\|^2_{\mathcal{L},\hat{\rho}} + \frac{\triangle t}{\kappa}\|f^{n,n+1}\|^2_{H,L^2_{\hat{\rho}}} + \|u^{n+1}_{h,\hat{\rho}} - u^{n}_{h,\hat{\rho}}\|^2_{\mathcal{L},\hat{\rho}}\\
&- \frac{(2-\kappa)h^{2p}}{C_{\mathrm{I}}^2 C_{\mathcal{B}}\triangle t}\|u^{n+1}_{h,\hat{\rho}} - u^{n}_{h,\hat{\rho}}\|^2_{\mathcal{L},\hat{\rho}} - 2C_{\det}\|u^{n+1}_{h,\hat{\rho}} - u^{n}_{h,\hat{\rho}}\|^2_{\mathcal{L},\hat{\rho}}\\
={}& \|u^{n}_{h,\hat{\rho}}\|^2_{\mathcal{L},\hat{\rho}} + \frac{\triangle t}{\kappa}\|f^{n,n+1}\|^2_{H,L^2_{\hat{\rho}}}\\
&+ \Big(1 - \frac{(2-\kappa)h^{2p}}{C_{\mathrm{I}}^2 C_{\mathcal{B}}\triangle t} - 2C_{\det}\Big)\|u^{n+1}_{h,\hat{\rho}} - u^{n}_{h,\hat{\rho}}\|^2_{\mathcal{L},\hat{\rho}},
\end{aligned}
$$

where in the second step we used the inequalities (2.2), (3.2) and (3.14). From the condition on $\triangle t, h$ after summing over $n = 0, \dots, N-1$ the equation (3.17) follows.

3. To treat the case of $f = 0$ we follow analogous steps as in Part 2. We consider $\kappa = 0$ as there is no need for the Young inequality in (3.22).

□

Theorem 3.4.4 tells us that when $\mathcal{L}$ is a symmetric operator, using the semi-implicit scheme leads to a conditionally stable solution if $C_{\det} \in (0, \frac{1}{2})$ and an unconditionally stable solution, if $C_{\det} \ge \frac{1}{2}$ (small randomness).

Remark 3. The discrete variational formulation (2.27) as well as the stability estimates presented in this section hold for the full-rank solution of the projector-splitting scheme from [LO14] with the ordering K, S, L, as presented in subsection 2.2.4. However, these results do not hold with the ordering K, L, S, which was discussed in [LO14]. This might be another reason why K, L, S performs poorly when compared to K, S, L (see [LO14, sec.5.2]).

Remark 4. All of the derived estimates for the discrete DLR solution obtained by Algorithm 2.2.1 hold also for the case of $\{u^n_{h,\hat{\rho}}\}_{n=0}^N$ being rank-deficient for some $n = 0, \dots, N$ as a consequence of Theorem 2.2.11 and the property (2.36). They hold as well for the discrete DLR solution obtained by the Algorithm 2.1.2 thanks to the variational formulation (2.44) and Lemma 2.3.2.

3.5 Example: random heat equation

In this section we will specifically address the case of a random heat equation. We will analyze what the underlying assumptions require of this problem, present the explicit and semi-implicit discretization schemes applied to a heat equation and state their stability properties.

Let $D \subset \mathbb{R}^d$, $1 \leq d \leq 3$ be a polygonal domain. Let $V = H^1_0(D) =: H^1_0$, $H = L^2(D) =: L^2$, $V' = H^{-1}(D) =: H^{-1}$ and $\mathcal{L}(x,\xi)(v) = -\nabla \cdot (a(x,\xi)\nabla v)$ with

$$0 < a_{\min} \leq a(x,\xi) \leq a_{\max} < \infty, \quad \forall x \in D, \forall \xi \in \Omega. \tag{3.23}$$

In this case, the scalar products $\langle v,w\rangle_{H,L^2_\rho}$, $\langle v,w\rangle_{V,L^2_\rho}$, $\langle v,w\rangle_{\mathcal{L},\rho}$ are defined as

$$\begin{aligned}
\langle v,w\rangle_{H,L^2_\rho} &= \int_\Omega \int_D v\, w \;\mathrm{d}x \;\mathrm{d}\rho \\
\langle v,w\rangle_{V,L^2_\rho} &= \int_\Omega \int_D \nabla v \cdot \nabla w \;\mathrm{d}x \;\mathrm{d}\rho \\
\langle v,w\rangle_{\mathcal{L},\rho} &= \int_\Omega \int_D a\nabla v \cdot \nabla w \;\mathrm{d}x \;\mathrm{d}\rho.
\end{aligned}$$

For the coercivity constant $C_{\mathcal{L}}$, it holds $C_{\mathcal{L}} \geq a_{\min}$; for the continuity constant $C_{\mathcal{B}}$, we have $C_{\mathcal{B}} \leq a_{\max}$; C_{P} is the Poincaré constant and the problem states: Given $f \in L^2(0,T;L^2_\rho(\Omega;L^2))$ and $u_0 \in L^2_\rho(\Omega;L^2)$, find $u_{\mathrm{true}} \in L^2(0,T;L^2_\rho(\Omega;H^1_0))$ with $\dot{u}_{\mathrm{true}} \in L^2(0,T;L^2_\rho(\Omega;H^{-1}))$ such that

$$\begin{aligned}
\int_\Omega \int_D \dot{u}_{\mathrm{true}} v \;\mathrm{d}x \;\mathrm{d}\rho + \int_\Omega \int_D a\nabla u_{\mathrm{true}} \cdot \nabla v \;\mathrm{d}x \;\mathrm{d}\rho &= \int_\Omega \int_D f v \;\mathrm{d}x \;\mathrm{d}\rho, \\
&\qquad \forall v \in L^2_\rho(\Omega;H^1_0) \\
u_{\mathrm{true}} = 0 \qquad &\text{a.e. on } (0,T] \times \partial D \times \Omega \\
u_{\mathrm{true}}(0,\cdot,\cdot) = u_0 \qquad &\text{a.e. in } D \times \Omega.
\end{aligned} \tag{3.24}$$

The discretization is performed as described in Section 2.2. To address the condition (2.2) we can consider a triangulation $\mathcal{T}_h$ of the domain D specified by the discretization parameter h and a corresponding finite element space V_h of continuous piece-wise polynomials of degree $\leq r$. Under the condition that the family of meshes $\{\mathcal{T}_h\}_h$ is quasi-uniform (see [EG04a, Def. 1.140] for definition), we have the inverse inequality (see [EG04a, Cor. 1.141])

$$\|\nabla v\|_H^2 \leq \frac{C_{\mathrm{I}}^2}{h^2}\|v\|_H^2, \qquad \forall v \in V_h$$

for some $C_{\mathrm{I}} > 0$. Integrating over Ω results in

$$\|v\|_{V,L^2_{\hat{\rho}}}^2 \leq \frac{C_{\mathrm{I}}^2}{h^2}\|v\|_{H,L^2_{\hat{\rho}}}^2, \qquad \forall v \in V_h \otimes L^2_{\hat{\rho}}, \tag{3.25}$$

i.e. we have the condition (2.2) with $p = 1$.

3.5.1 Explicit Euler scheme

Applying the explicit Euler scheme in the operator evaluation for a random heat equation, i.e.

$$\mathcal{L}(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}}) = -\nabla \cdot (a \nabla u^n_{h,\hat{\rho}}),$$

results in the following system of equations

$$\langle \bar{u}^{n+1}, v_h\rangle_H = \langle \bar{u}^n, v_h\rangle_H - \triangle t\, \langle \mathbb{E}_{\hat{\rho}}[a\nabla u^n_{h,\hat{\rho}}], \nabla v_h\rangle_H + \triangle t \langle \mathbb{E}_{\hat{\rho}}[f(t_n)], v_h\rangle_H, \quad \forall v_h \in V_h$$

$$\langle \tilde{U}_j^{n+1}, v_h\rangle_H = \langle U_j^n, v_h\rangle_H - \triangle t\, \langle \mathbb{E}_{\hat{\rho}}[a\nabla u^n_{h,\hat{\rho}} Y_j^n], \nabla v_h\rangle_H + \triangle t \langle \mathbb{E}_{\hat{\rho}}[f(t_n) Y_j^n], v_h\rangle_H \qquad \forall j,\ \forall v_h \in V_h$$

$$\tilde{M}^{n+1}(\tilde{Y}^{n+1} - Y^n)^{\mathsf{T}} = -\triangle t\, \mathcal{P}^{\perp}_{\hat{\rho},\mathcal{Y}^n}\left[\langle a\nabla u^n_{h,\hat{\rho}}, \nabla \tilde{U}^{n+1}\rangle_H - \langle f(t_n), \tilde{U}^{n+1}\rangle_H\right]^{\mathsf{T}} \qquad \text{in } L^2_{\hat{\rho}}.$$

The stability properties stated in Theorem 3.4.3 part 2. and 4. hold under the condition

$$\frac{\triangle t}{h^2} \leq \frac{2-\kappa}{C_{\mathrm{I}}^2 C_{\mathcal{B}}}.$$

3.5.2 Semi-implicit scheme

Let us consider the decomposition

$$a = \bar{a} + a_{\mathrm{stoch}}, \quad \text{with} \quad \bar{a} = \mathbb{E}_{\hat{\rho}}[a] \quad \text{and} \quad \mathbb{E}_{\hat{\rho}}[a_{\mathrm{stoch}}] = 0, \tag{3.26}$$

i.e.

$$\mathcal{L}(u) = \underbrace{-\nabla \cdot (\bar{a}\nabla u)}_{\mathcal{L}_{\mathrm{det}}} \underbrace{-\nabla \cdot (a_{\mathrm{stoch}}\nabla u)}_{\mathcal{L}_{\mathrm{stoch}}}.$$

The condition (2.14) is satisfied, since $\bar{a}$ is positive everywhere in D as assumed in (3.23). Hence,

$$\langle u, v\rangle_{\mathcal{L}_{\mathrm{det}}} = \int_D \bar{a}\, \nabla u \cdot \nabla v \,\mathrm{d}x, \qquad u, v \in V$$

is a scalar product on $V = H_0^1(D)$. The semi-implicit time integration is realized by

$$\mathcal{L}(u_{h,\hat{\rho}}^n, u_{h,\hat{\rho}}^{n+1}) = -\nabla \cdot (\bar{a} \nabla u_{h,\hat{\rho}}^{n+1}) - \nabla \cdot (a_{\mathrm{stoch}} \nabla u_{h,\hat{\rho}}^n). \tag{3.27}$$

Note that the condition (3.14) is automatically satisfied for a random heat equation, since we have

$$\begin{aligned} \|u\|_{\mathcal{L}_{\mathrm{det}},\rho}^2 = \int_\Omega \int_D \bar{a} \nabla u \cdot \nabla u \,\mathrm{d}x \,\mathrm{d}\rho &\geq \inf_{x \in D, \xi \in \Omega} \frac{\bar{a}}{a} \int_\Omega \int_D a \nabla u \cdot \nabla u \,\mathrm{d}x \,\mathrm{d}\rho \\ &= \inf_{x \in D, \xi \in \Omega} \frac{\bar{a}}{a} \|u\|_{\mathcal{L},\rho}^2 \qquad \forall u \in L_\rho^2(\Omega; V), \end{aligned}$$

and $\inf_{x \in D, \xi \in \Omega} \frac{\bar{a}}{a} \geq \frac{a_{\min}}{a_{\max}} > 0$.

The system of equations (2.16)–(2.18) can be rewritten as

$$\begin{aligned} &\langle \bar{u}^{n+1}, v_h\rangle_H + \triangle t \langle \bar{a} \nabla \bar{u}^{n+1}, \nabla v_h\rangle_H \\ &\qquad = \langle \bar{u}^n, v_h\rangle_H - \triangle t \langle \mathbb{E}_{\hat{\rho}}[a_{\mathrm{stoch}} \nabla u_{h,\hat{\rho}}^n], \nabla v_h\rangle_{H^d} + \triangle t \langle \mathbb{E}_{\hat{\rho}}[f^{n,n+1}], v_h\rangle_H \\ &\langle \tilde{U}_j^{n+1}, v_h\rangle_H + \triangle t \langle \bar{a} \nabla \tilde{U}_j^{n+1}, \nabla v_h\rangle_H \\ &\qquad = \langle \tilde{U}_j^n, v_h\rangle_H - \triangle t \langle \mathbb{E}_{\hat{\rho}}[a_{\mathrm{stoch}} \nabla u_{h,\hat{\rho}}^n Y_j^n], \nabla v_h\rangle_{H^d} + \triangle t \langle \mathbb{E}_{\hat{\rho}}[f^{n,n+1} Y_j^n], v_h\rangle_H \\ &\left(\tilde{Y}^{n+1} - Y^n\right)\left(\tilde{M}^{n+1} + \triangle t \langle \bar{a} \nabla \tilde{U}^{n+1\intercal}, \nabla \tilde{U}^{n+1}\rangle_H\right) \\ &\qquad = -\triangle t \mathcal{P}_{\hat{\rho},\mathcal{Y}^n}^{\perp}[\langle a_{\mathrm{stoch}} \nabla u_{h,\hat{\rho}}^n, \nabla \tilde{U}^{n+1}\rangle_{H^d} - \langle f^{n,n+1^*}, \tilde{U}^{n+1}\rangle_H]. \end{aligned}$$

For a further specified diffusion coefficient we can state the following stability properties.

Proposition 3.5.1. For the case

$$\bar{a}(x) \geq a_{\mathrm{stoch}}(x, \xi), \qquad \forall x \in D, \xi \in \Omega$$

which is satisfied in particular if

$$\begin{gathered} a(x, \xi) = \bar{a}(x) + \sum_{j=1}^M a_j(x) \xi_j, \\ \Omega \subset \mathbb{R}^M \text{ and } \Omega \text{ is symmetric, i.e. } \xi \in \Omega \implies -\xi \in \Omega, \end{gathered} \tag{3.28}$$

we have the stability properties (3.17) and (3.18) for any $\triangle t, h$.

Proof. The condition $\bar{a}(x) \geq a_{\text{stoch}}(x, \xi)$ for every $x \in D, \xi \in \Omega$ implies

$$\frac{\bar{a}(x)}{a(x,\xi)} \geq \frac{1}{2},$$

i.e. $C_{\text{det}} \geq \inf_{x \in D, \xi \in \Omega} \frac{\bar{a}}{a} \geq \frac{1}{2}$. Together with Theorem 3.4.4 we conclude the result. ☐

Proposition 3.5.1 tells us that applying a semi-implicit scheme to solve a heat equation with diffusion coefficient as described in (3.28) results in an unconditionally stable scheme. This result as well as some of the previous estimates will be numerically verified in the following section.

3.6 Numerical results

This section is dedicated to numerically study the stability estimates derived for a discrete DLR approximation in Section 3.4. In particular, we will be concerned with a random heat equation, as introduced in (3.24), with zero forcing term and diffusion coefficient of the form (3.28). We will look at the behavior of suitable norms of the solutions of the discretization schemes introduced in Section 2.2.1. We will as well look at a discretization scheme in which the projection is performed explicitly to see how important it is to project on the new computed basis $\widetilde{U}^{n+1}$ in (2.10). As a last result we provide a comparison with the projector-splitting scheme from [LO14].

Let us consider problem (3.24) set in a unit square $D = [0,1]^2$ and sample space $\Omega = [-1,1]^M$ with M specified below, and an uncertain diffusion coefficient

$$a(x,\xi) = a_0 + \sum_{m=1}^{M} \frac{\cos(2\pi m x_1) + \cos(2\pi m x_2)}{m^2 \pi^2} \xi_m, \tag{3.29}$$

where $x = (x_1, x_2) \in D,\ \xi = (\xi_1, \ldots, \xi_M) \in \Omega$. We let $a_0 = 0.3$, and equip $([-1,1]^M, \mathcal{B}([-1,1]^M))$ with the uniform measure $\rho(\mathrm{d}\xi) = \bigotimes_{i=1}^{M} \frac{\lambda(\mathrm{d}\xi_i)}{2}$ with λ the Lebesgue measure restricted to the Borel σ-algebra $\mathcal{B}([-1,1])$. In this case the conditions (3.23), (2.14) and (3.14) are satisfied with $a_{\min} > 0.04, C_{\text{det}} > \frac{1}{2}$. The initial condition is chosen as

$$\begin{aligned} u_0(x,\xi) &= 10\sin(\pi x_1)\sin(\pi x_2) + 2\sin(2\pi x_1)\sin(2\pi x_2)\xi_1 \\ &\qquad + 2\sin(4\pi x_1)\sin(4\pi x_2)\xi_2 + 2\sin(6\pi x_1)\sin(6\pi x_2)\xi_1^2. \\ &= 10\sin(\pi x_1)\sin(\pi x_2) + \frac{4}{3}\sin(6\pi x_1)\sin(6\pi x_2) + 2\sin(2\pi x_1)\sin(2\pi x_2)\xi_1 \\ &\qquad + 2\sin(4\pi x_1)\sin(4\pi x_2)\xi_2 + 2\sin(6\pi x_1)\sin(6\pi x_2)\Big(\xi_1^2 - \mathbb{E}[\xi_1^2]\Big). \end{aligned} \tag{3.30}$$

The spatial discretization is performed by the finite element (FE) method with $P1$ finite elements over a uniform mesh. The dimension of the corresponding FE space is determined by h—the element size. For this type of spatial discretization we have the inverse inequality (3.25):

$$\|v\|^2_{V,\hat{\rho}} \leq \frac{C_{\mathrm{I}}^2}{h^2}\|v\|^2_{H,L^2_{\hat{\rho}}}, \qquad \forall v \in V_h \otimes L^2_{\hat{\rho}}.$$

Concerning the stochastic discretization we will consider a tensor grid quadrature with Gauss-Legendre points for the case of a low-dimensional stochastic space $M = 2$ and a Monte-Carlo quadrature for the case $M = 10$. The time integration implements the explicit scheme and the semi-implicit scheme described in subsection 2.2.1. We will consider the forcing term $f = 0$, i.e. a dissipative problem and time T such that the energy norm ($\|\cdot\|_{\mathcal{L},\hat{\rho}}$) of the solution attains a value smaller than 10^{-10}. Our simulations were performed using the Fenics library [Aln+15a].

3.6.1 Explicit scheme

Since $f = 0$, the result in Theorem 3.4.3 predicts a decay of the norm of the solution

$$\|u^{n+1}_{h,\hat{\rho}}\|_{\mathcal{L},\hat{\rho}} \leq \|u^{n}_{h,\hat{\rho}}\|_{\mathcal{L},\hat{\rho}}, \qquad \|u^{n+1}_{h,\hat{\rho}}\|_{H,L^2_{\hat{\rho}}} \leq \|u^{n}_{h,\hat{\rho}}\|_{H,L^2_{\hat{\rho}}} \qquad \forall n = 0,\ldots,N-1$$

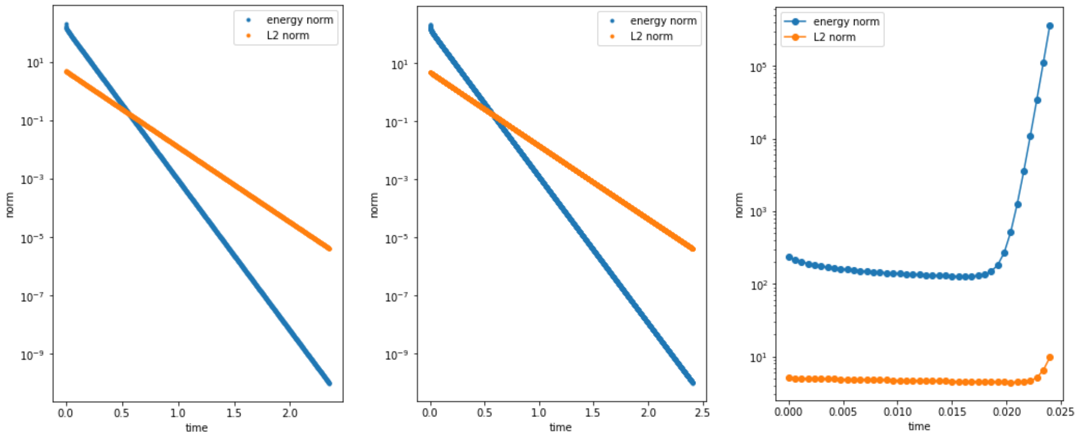
under the stability condition

$$\frac{\triangle t}{h^2} \leq \frac{2}{C_{\mathrm{I}}^2 C_{\mathcal{B}}} =: K. \tag{3.31}$$

We aim at verifying such result numerically. We set a rank $R = 3$ and consider a sample space $[-1,1]^M$ of dimension $M = 2$ or $M = 10$ with either Gauss-Legendre or Monte-Carlo (MC) stochastic discretization.

$M = 2$

First we consider the sample space $[-1,1]^M$ of dimension $M = 2$ and Gauss-Legendre quadrature with $9 \times 9 = 81$ collocation points. From what we observed in our simulations, for this test case we have $K \approx 0.085$. Figure 3.1 shows the behavior of the energy norm ($\|\cdot\|_{\mathcal{L},\hat{\rho}}$) and the L^2 norm ($\|\cdot\|_{H,L^2_{\hat{\rho}}}$) in 3 different scenarios: in the first scenario we set $h_1 = 0.142, \triangle t_1 = 0.0018$, i.e. the condition $\triangle t_1/h_1^2 \leq K$ is satisfied and observe that both the energy norm and the L^2 norm of the solution decrease in time (see Figure 3.1(a)); in the second scenario, we halved the element size $h_2 = h_1/2$ and divided by 4 the time step $\triangle t_2 = \triangle t_1/4$ so that the condition (3.31) is still satisfied. The norms again decreased in time (Figure 3.1(b)); in the third scenario we violated the condition (3.31) by setting $h_3 = h_1/2$ and $\triangle t_3 = \triangle t_1/3$. After a certain time the norms exploded (Figure 3.1(c)).



(a) $h_1 = 0.142$
$\triangle t_1 = 0.0017$

(b) $h_1 = 0.142/2$
$\triangle t_1 = 0.0017/4$

(c) $h_1 = 0.142/2$
$\triangle t_1 = 0.0017/3$

Figure 3.1 – Behaviour of the energy norm ($\|\cdot\|_{\mathcal{L},\rho}$—blue) and the L^2 norm ($\|\cdot\|_{H,L^2_\rho}$—orange) when applying the explicit time integration scheme with $M = 2$ and 81 Gauss-Legendre collocation points for three different pairs of the discretization parameters $h, \triangle t$. When the condition (3.31) is satisfied the solution is stable [(a)–(b)], whereas violating the condition results in instability [(c)].

To numerically demonstrate the sharpness of the condition (3.31), we ran the simulation with 72 different pairs of discretization parameters $h, \triangle t$. The results are shown in Figure 3.2, where we depict whether the energy norm at time T is bellow 10^{-10}, in which case the norm was consistently decreasing; or more than 10^4, in which case the solution blew up. We observe that a stable $\triangle t$ has to be chosen to satisfy $\triangle t \leq Kh^2$, which confirms the sharpness of our theoretical derivations.

$M = 10$

In our second example we will consider a higher-dimensional problem: $M = 10$ for which we use a standard Monte-Carlo technique with 50 points. We observe a very similar behaviour as in the small dimensional case. Figure 3.3 shows that satisfying the condition $\triangle t_1/h_1^2 \leq K$ with $K = 0.085$ results in a stable scheme while violating it makes the solution blow up.

3.6.2 Semi-implicit scheme

We proceed with the same test-case with $M = 10$, same spatial and stochastic discretization, i.e. Monte-Carlo method with 50 samples and employ a semi-implicit scheme in the operator evaluation. Since the diffusion coefficient considered is of the form (3.28) and

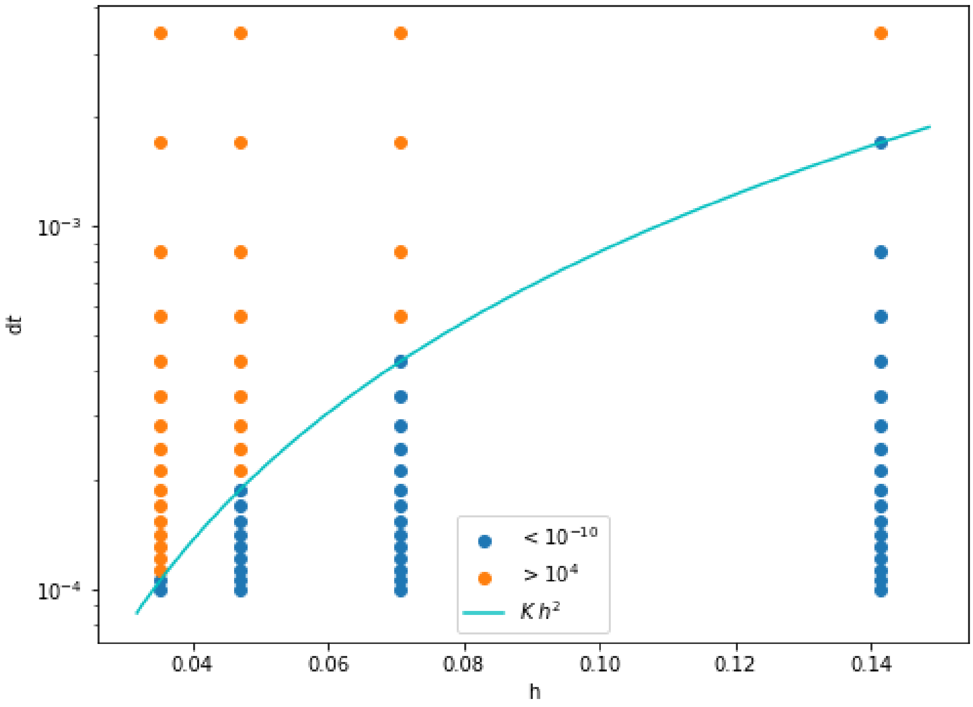
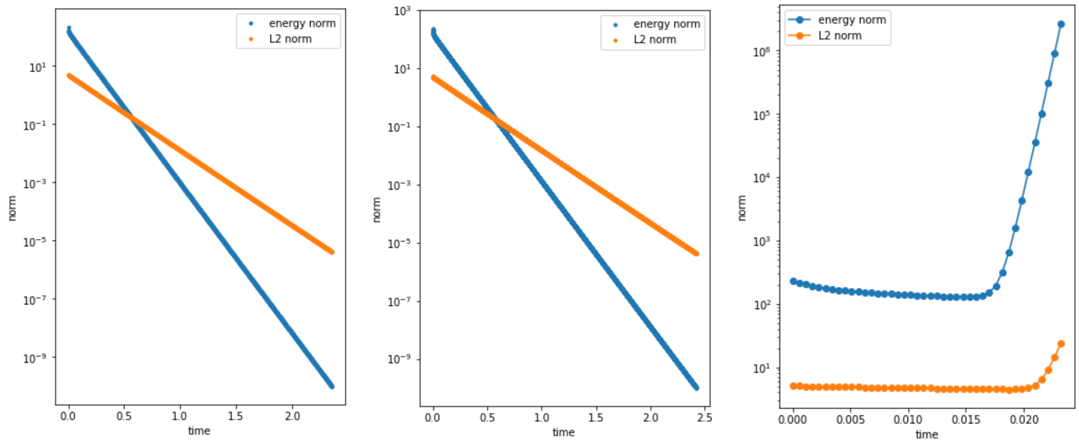


Figure 3.2 – This figure shows whether the energy norm $\| \cdot \|_{\mathcal{L},\rho}$ of the solution was monotonously decreasing till 10^{-10} (blue) or has blown up (orange) for different choices of time step $\triangle t$ and discretization parameter h when applying the explicit scheme for the operator evaluation. We observe a clear quadratic dependence of $\triangle t$ on h. K was set to 0.085.



(a) $h_1 = 0.142$
$\triangle t_1 = 0.0017$

(b) $h_1 = 0.142/2$
$\triangle t_1 = 0.0017/4$

(c) $h_1 = 0.142/2$
$\triangle t_1 = 0.0017/3$

Figure 3.3 – Behaviour of the energy norm ($\| \cdot \|_{\mathcal{L},\rho}$—blue) and the L^2 norm ($\| \cdot \|_{H,L^2_\rho}$—orange) when applying the explicit time integration scheme with $M = 10$ and 50 Monte Carlo points for three different pairs of the discretization parameters $h, \triangle t$. We see, again, that satisfying the condition (3.31) ((a) and (b)) results in stable behaviour while when violating the condition (c) the solution blows up.

$f = 0$, Theorem 3.4.4 predicts

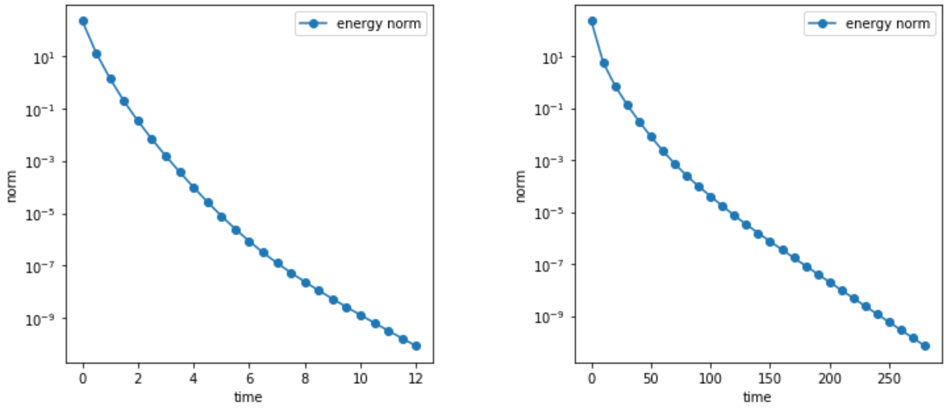
$$\|u^{n+1}_{h,\hat{\rho}}\|_{\mathcal{L},\hat{\rho}} \leq \|u^{n}_{h,\hat{\rho}}\|_{\mathcal{L},\hat{\rho}} \qquad \forall h,\ \triangle t,\ \forall n = 0, \ldots, N-1.$$

We set the spatial discretization $h = 0.142$ and vary the time step $\triangle t$. We observe a stable behaviour no matter what $\triangle t$ is used, which confirms the theoretical result (see Figure 3.4).

We report that the results for $M = 2$ with 81 Gauss-Legendre collocation points exhibited a similar unconditionally-stable behaviour.

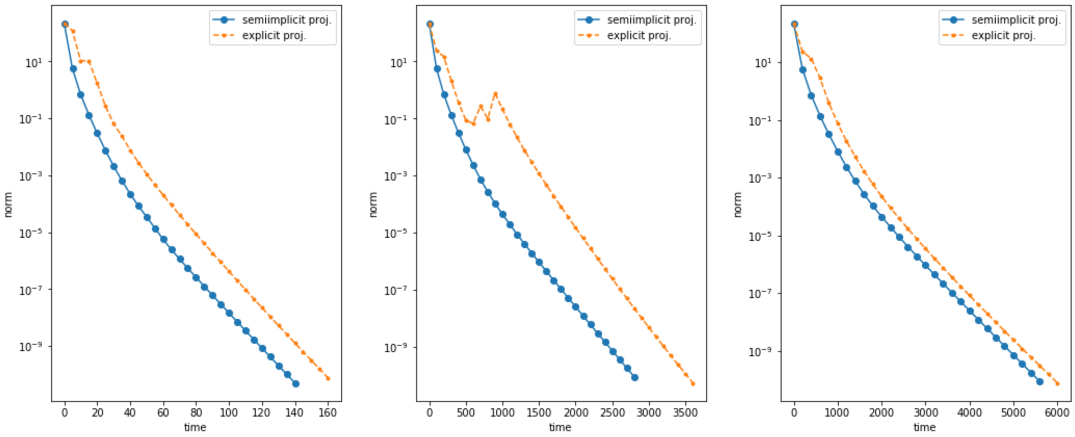
Explicit projection

The following results give an insight into the importance of performing the projection in a ‘Gauss-Seidel’ way, i.e. projection on the stochastic basis is done explicitly, Y^n kept from the previous time step, while the projection on the deterministic basis is done implicitly, i.e. we use the new computed $\tilde{U}^{n+1}$ (see Algorithm 2.2.1 for more details). For comparison we consider a fully explicit projection, i.e. Y^n as the stochastic basis and U^n as the deterministic basis. We use a semi-implicit scheme to treat the operator evaluation term as described in subsection 2.2.1. As shown in Figure 3.5, in all 3 cases the solution reaches the zero steady state, however, not in a monotonous way.



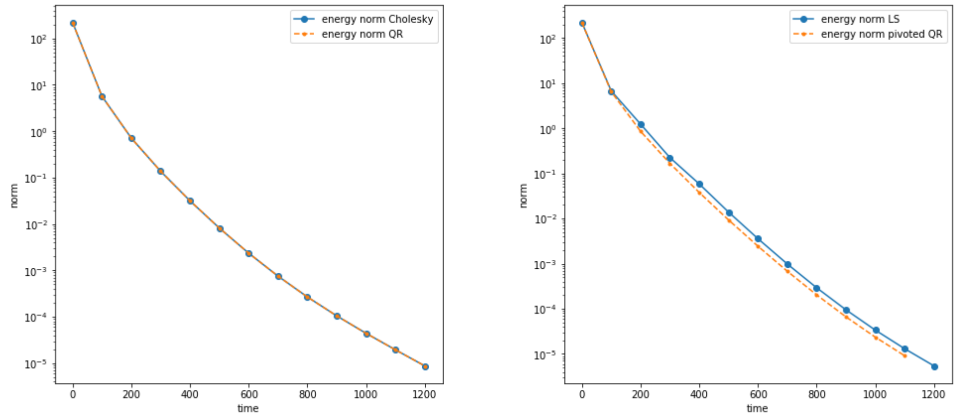
(a) $h = 0.142, \triangle t_1 = 0.5$ (b) $h = 0.142, \triangle t_2 = 10$

Figure 3.4 – Behaviour of the energy norm ($\|\cdot\|_{\mathcal{L},\rho}$) for two different time steps when applying the semi-implicit time integration scheme. We observe a decrease of norms for arbitrarily large time step.



(a) $h = 0.142, \triangle t_1 = 5$ (b) $h = 0.142, \triangle t_2 = 100$ (c) $h = 0.142, \triangle t_3 = 200$

Figure 3.5 – Behaviour of the energy norm ($\|\cdot\|_{\mathcal{L},\rho}$) for 3 different time steps when treating the projection in an explicit way (orange) and in a semi-implicit way (blue). We used the semi-implicit scheme for the operator evaluation term. We see that, as opposed to a semi-implicit projection, with an explicit projection we do not obtain an unconditional norm decrease.



(a) $R = 3, \triangle t_1 = 100$

(b) $R = 20, \triangle t_2 = 100$

Figure 3.6 – Energy norm ($\|\cdot\|_{\mathcal{L},\rho}$) for 2 different ranks $R = 3, 20$ and 2 different time discretization schemes: Algorithm 2.1.2 with (pivoted) QR decomposition (orange) and Algorithm 2.2.1 with Cholesky factorization or least squares. Both methods in both cases exhibit a monotonous decrease of the energy norm.

3.6.3 Comparison with the DDO projector-splitting scheme

We now compare the performance of the discretization scheme from Algorithm 2.2.1 with the projector-splitting scheme from Algorithm 2.1.2.

We proceed with setting $h = 0.142, M = 10, \triangle t = 100$, stochastic discretization is performed again by Monte-Carlo method with 50 points and we implemented the semi-implicit scheme in the operator evaluation for both the Algorithm 2.2.1 and the projector-splitting Algorithm 2.1.2. We expect that the energy norm decreases on every step independently of the time step size.

We fix $R = 3$. Throughout the whole simulation, the computed solution stays full rank, in which case the two schemes have been shown to be equivalent (see subsection 2.2.4). In Figure 3.6(a) this can be well observed. Steps 2. and 5. from Algorithm 2.1.2 are performed by a QR decomposition, whereas the linear system in (2.10) is solved by the Cholesky factorization (with a help of the SciPy library [JOP+01], version 0.19.1).

We now investigate the behavior of the two algorithms in presence of a rank deficient solution. We fix $R = 20$. The initial condition (3.30) is of rank 3. For the first couple of steps the discrete DLR solution therefore stays of rank lower than $R = 20$. The matrix $\tilde{M}^{n+1}$ from (2.10) is singular and the solution of the system (2.10) is obtained as a least squares solution implemented via an SVD decomposition. The threshold to detect the effective rank of $\tilde{M}^{n+1}$ is set to $\varepsilon\,\sigma_1 R$ where ε is the machine precision and σ_1 is the largest singular value of $\tilde{M}^{n+1}$. Steps 2. and 5. from Algorithm 2.1.2 are performed by a pivoted QR decomposition. The solution obtained by both algorithms are proved to be stable in this scenario. The two proposed schemes exhibit minor differences, however both of them are stable (see Figure 3.6(b)).

4 A-priori error estimation

This chapter witnesses another use of the variational formulation (2.37) for the DLR scheme proposed in Chapter 2. We present an a-priori error estimation for a fully discrete DLR solution of a random parabolic equation obtained by the scheme described in Algorithm 2.2.1. The spatial discretization is assumed to be performed by the finite element method and the stochastic discretization by the Monte Carlo method. The algorithm was derived applying a first-order-in-time approximation of the DLR equations (1.8)–(1.10). However, the usual error bounds break down when the DLR approximation has small singular values. On the other hand, we showed that the scheme is exact when the true solution is of rank $\leq R$ and we obtained stability estimates for the DLR solution that do not depend on the smallest singular value. In this work, we derive an a-priori error bound w.r.t. the spatial, time and stochastic discretization without dependence on the smallest singular value. Such a result does not generally hold for a different temporal discretization of (1.8)–(1.10). We point the reader to [KLW16], where the authors considered a DLR approximation for time-dependent matrices or tensors applying a continuous-in-time projector-splitting integrator (see Algorithm 2.1.1 in Section 2.1). As recalled in Theorem 2.1.4, they proved a first-order convergence w.r.t. the time step $\triangle t$ which, as well, does not depend on the smallest singular value. More on the comparison of these two results can be found at the end of Section 4.2. In this work, we restrict ourselves to the case of $\mathcal{L}$ being a random elliptic differential operator. We point the reader to [Con20] for an a-priori error analysis for a continuous DLR approximation for parabolic problems. We start with specifying the problem and the discretization in Section 4.1. In Section 4.2, we state and prove the error estimation without taking into account the stochastic discretization by expressing the error in the discrete $\|\cdot\|_{L^2_{\hat{\rho}}}$-norm. This allows us to reuse the well-established results for a-priori error estimations for a deterministic parabolic equation. Finally, in Section 4.3 we derive an error estimation that includes the stochastic discretization contribution.

4.1 Problem specification

In this work, we consider the same problem setting as in Chapter 3. The considered physical spaces are, however, further specified and described in the this section.

Let $D \subset \mathbb{R}^d$, $1 \leq d \leq 3$ be a polygonal domain with Lipschitz boundary, and let $V = H_0^1(D) =: H_0^1$, $H = L^2(D) =: L^2$, $V' = H^{-1}(D) =: H^{-1}$ and $H^p := H^p(D)$, $p \in \mathbb{N}, p \geq 1$. Given a final time $T > 0$, a random forcing term $f \in L^2(0,T; L^2_\rho(\Omega; H))$ and a random initial condition $u_0 \in L^2_\rho(\Omega; V)$, we assume that there exists a solution $u_{\text{true}} \in L^2(0,T; L^2_\rho(\Omega; V))$ with $\dot{u}_{\text{true}} \in L^2(0,T; L^2_\rho(\Omega; V'))$ satisfying

$$\begin{aligned} \Big(\dot{u}_{\text{true}}(\omega), v\Big)_{V'V} + \Big(\mathcal{L}(u_{\text{true}})(\omega), v\Big)_{V'V, L^2_\rho} &= \langle f(\omega), v\rangle_{H, L^2_\rho}, \\ &\forall v \in V,\ \forall \omega \in \Omega,\ \text{a.e. } t \in (0,T] \\ u_{\text{true}}(0,\omega) = u_0(\omega), &\quad \forall \omega \in \Omega. \end{aligned} \tag{4.1}$$

Note that such solution satisfies the weak formulation (3.3) as well. The considered operator $\mathcal{L}$ is random and elliptic, as defined in Section 3.1. In addition, we assume that $\mathcal{L}$ is a second order differential operator.

The spatial discretization is performed via the finite element method (see e.g. [QV08]). We consider a triangulation $\mathcal{T}_h$ of the domain D specified by the discretization parameter h and a corresponding finite element space V_h of continuous piece-wise polynomials of degree $\leq r$, i.e.

$$V_h = \{v_h \in C^0(\bar{D}) : \quad v_h|_K \in \mathbb{P}_r,\ \forall K \in \mathcal{T}_h\}.$$

Under the condition that the family of meshes $\{\mathcal{T}_h\}_h$ is quasi-uniform (see [EG04a, Def. 1.140] for definition), we have the inverse inequality (see [EG04a, Cor. 1.141])

$$\|\nabla v\|_H^2 \leq \frac{C_{\text{I}}^2}{h^2}\|v\|_H^2, \qquad \forall v \in V_h$$

for some $C_{\text{I}} > 0$. We now follow by introducing an operator which will be later used when deriving the a-priori estimates. Let us consider a (random) 'projection' operator $P^r_{1,h}(\omega) : V \to V_h$ for all $\omega \in \Omega$ defined as

$$\forall u \in V,\ \Big(L(P^r_{1,h}(\omega)u - u),\ v_h\Big)_{V'V} = 0 \qquad \forall v_h \in V_h,$$

where L is the operator introduced in Section 3.1. The existence of such operator for $\forall \omega \in \Omega$ is ensured by the coercivity condition (3.1) and the Lax-Milgram lemma. If L is symmetric, then $P^r_{1,h}$ is simply an orthogonal projection operator onto V_h w.r.t. the scalar product $(L(\omega)\cdot,\cdot)_{V'V}$. We further assume that $\mathcal{T}_h$ is a regular family of triangulations and that for $\forall \omega \in \Omega$ the solution $\phi(\omega)(f)$ of the adjoint problem

$$\phi(\omega)(f) \in V : \quad (\mathcal{L}(v)(\omega), \phi(f))_{V'V} = \langle f, v \rangle_H, \quad \forall v \in V$$

satisfies $\phi(f) \in H^2$ when $f \in H = L^2$. By proceeding as in [QV08, Sec. 3.5], we have for $\forall \omega \in \Omega$

$$\|P^r_{1,h}(\omega)w - w\|_V + h^{-1}\|P^r_{1,h}(\omega)w - w\|_H \leq C_{P^r} h^p \|w\|_{H^{p+1}}, \quad 0 \leq p \leq r, \ w \in V \cap H^{r+1}, \tag{4.2}$$

where $C_{P^r} > 0$ is independent of w and h. In this work, we assume that C_{P^r} is independent of ω. In addition, we have

$$\|P^r_{1,h} w\|_{H,L^2_{\hat{\rho}}} \leq c_r \|w\|_{H,L^2_{\hat{\rho}}}. \tag{4.3}$$

The stochastic discretization applies the Monte Carlo method, where the sample points $\{\omega_k\}_{k=1}^{\hat{N}}$ are taken as iid samples from ρ and the weights satisfy $\lambda_k = \frac{1}{\hat{N}}$, $\forall k = 1, \dots, \hat{N}$. The empirical measure as well as the computation of the expectation value are detailed in Chapter 2.

Note that the true solution u_{true} satisfies the following weak formulation w.r.t. the empirical measure

$$\Big(\dot{u}_{\text{true}}, v_{\hat{\rho}}\Big)_{V'V,L^2_{\hat{\rho}}} + \Big(\mathcal{L}(u_{\text{true}}), v_{\hat{\rho}}\Big)_{V'V,L^2_{\hat{\rho}}} = \langle f, v_{\hat{\rho}} \rangle_{H,L^2_{\hat{\rho}}}, \quad \forall v_{\hat{\rho}} \in L^2_{\hat{\rho}}(\Omega; V), \text{ a.e. } t \in (0, T]. \tag{4.4}$$

We recall that the dicrete DLR solution satisfies

$$\begin{gathered}\Big\langle \frac{u^{n+1}_{h,\hat{\rho}} - u^n_{h,\hat{\rho}}}{\triangle t}, v_{h,\hat{\rho}} \Big\rangle_{H,L^2_{\hat{\rho}}} + \Big(\mathcal{L}(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}}), v_{h,\hat{\rho}}\Big)_{V'V,L^2_{\hat{\rho}}} = \Big\langle f^{n,n+1}, v_{h,\hat{\rho}} \Big\rangle_{H,L^2_{\hat{\rho}}}, \\ \forall v_{h,\hat{\rho}} = \bar{v}_h + v^*_{h,\hat{\rho}} \text{ with } \bar{v}_h \in V_h \text{ and } v^*_{h,\hat{\rho}} \in \mathcal{T}_{\bar{U}^{n+1} Y^{n\intercal}} \mathcal{M}^{h,\hat{\rho}}_R.\end{gathered} \tag{4.5}$$

Both of these variational formulations will play a crucial role when deriving the a-priori estimates.

Let $u^R_{\text{true}}(0)$ denote the truncated Karhunen-Loève expansion (1.4) of the initial condition u^0, for which we assume

$$||u^R_{\text{true}}(0)||_{H^r,L^2_{\hat{\rho}}} \leq c||u_{\text{true}}(0)||_{H^r,L^2_{\hat{\rho}}}. \tag{4.6}$$

Then the initial condition is defined by taking independent samples from $u^R_{\text{true}}(0)$ and applying an operator $P^r_h : V \to V_h$

$$u^0_{h,\hat{\rho}} = \{P^r_h u^R_{\text{true}}(0, \omega_j)\}_{j=1}^{\hat{N}}. \tag{4.7}$$

We assume that this operator satisfies

$$\|v - P_h^r[v]\|_{H,L^2_{\hat{\rho}}} \leq Ch^r \|v\|_{H^r,L^2_{\hat{\rho}}}.$$

The choices for P_h^r include projection on V_h w.r.t. the scalar product $\langle\cdot,\cdot\rangle_V$; if the triangulation is quasi-uniform, projection on V_h w.r.t. the scalar product $\langle\cdot,\cdot\rangle_H$; or if the initial condition $u^0(\omega_i) \in H^2$ for every sample point ω_i, we can apply the finite element interpolation operator (see [QV08, Sec. 3.4]).

In the error estimation for the semi-implicit scheme, we will use a constant $C_{\mathrm{det},\mathcal{B}}$ that bounds the operator $\mathcal{L}_{\mathrm{det}}$

$$|(\mathcal{L}_{\mathrm{det}}(u), v)_{V'V,L^2_{\hat{\rho}}}| \leq C_{\mathrm{det},\mathcal{B}} \|u\|_{V,L^2_{\hat{\rho}}} \|v\|_{V,L^2_{\hat{\rho}}}. \tag{4.8}$$

To proceed with the a-priori error estimation, we need to state the following assumptions.

Assumptions 1.

We assume that the following inequalities hold for $\forall n = 0, \ldots, N-1$

1. $$\left\|\mathcal{P}_{\tilde{\mathcal{U}}^{n+1}}[\mathcal{K}]\right\|_V \leq C_{\mathcal{P}} \|\mathcal{K}\|_V, \quad \forall \mathcal{K} \in V \tag{4.9}$$
2. $$\left\|\mathcal{P}_{\tilde{\mathcal{U}}^{n+1}}[\mathcal{K}]\right\|_V \leq C_{\mathcal{P}} \|\mathcal{K}\|_{V'}, \quad \forall \mathcal{K} \in V' \tag{4.10}$$
3. $$\|\Pi^{\perp}_{\tilde{U}^{n+1}Y^{n\intercal}}[f^{n+1^*} - \mathcal{L}^*(\tilde{U}^{n+1}Y^{n\intercal})]\|_{V',L^2_{\hat{\rho}}} \leq \varepsilon \tag{4.11}$$

We recall that $\mathcal{P}_{\tilde{\mathcal{U}}^{n+1}}$ denotes the H-orthogonal projection onto the subspace $\tilde{\mathcal{U}}^{n+1}$ and was first introduced in (1.14). Concerning the first two inequalities, note that we can bound $\left\|\mathcal{P}_{\tilde{\mathcal{U}}^{n+1}}[\mathcal{K}]\right\|_V$ in the following way

$$\left\|\mathcal{P}_{\tilde{\mathcal{U}}^{n+1}}[\mathcal{K}]\right\|_V \leq \frac{C_{\mathrm{I}}}{h^p}\left\|\mathcal{P}_{\tilde{\mathcal{U}}^{n+1}}[\mathcal{K}]\right\|_H \leq \frac{C_{\mathrm{I}}}{h^p}\left\|\mathcal{K}\right\|_H \leq \frac{C_{\mathrm{I}}C_{\mathrm{P}}}{h^p}\left\|\mathcal{K}\right\|_V$$

(and analogously for $\left\|\mathcal{P}_{\tilde{\mathcal{U}}^{n+1}}[\mathcal{K}]\right\|_{V'}$). This, however, yields the constant $C_{\mathcal{P}}$ dependent on h, which results in suboptimal bounds. Inequalities (4.10)–(4.9) assume the constant $C_{\mathcal{P}}$ to be independent of h for functions in a subspace $\tilde{\mathcal{U}}^{n+1}$. This assumption gets further simplified for the DLR approximation in the DO form - with orthonormal determinic modes $\{U_j\}_{j=1}^R$ and linearly independent stochastic modes $\{Y_j\}_{j=1}^R$ (see Lemma 4.2.2 for further details). The third condition assumes that the operator is in the tangent space up to a small remainder. We need this condition in order to obtain a low-rank approximation error of order $O(\varepsilon)$. Note that this is analogous to the low-rank assumption (2.7) used in the work [KLW16], where such condition is required to hold in a neighbourhood of the trajectory of the approximate solution. In our case, we assume it to hold at all 'intermediate' solutions $\tilde{U}^{n+1}Y^{n\intercal}$. The following lemma highlights the significance of assumptions (4.9) – (4.10).

Lemma 4.1.1. *Assumptions* (4.9) – (4.10) *imply*

1.
$$\left\|\Pi_{\tilde{U}^{n+1}Y^{n\intercal}}[\mathcal{K}]\right\|_{V',L^2_{\hat\rho}} \leq (1+C_{\mathcal{P}})\|\mathcal{K}\|_{V',L^2_{\hat\rho}}, \quad \forall \mathcal{K} \in L^2_{\hat\rho} \otimes V'_h$$

2.
$$\left\|\Pi_{\tilde{U}^{n+1}Y^{n\intercal}}[\mathcal{K}]\right\|_{V_h,L^2_{\hat\rho}} \leq (1+C_{\mathcal{P}})\|\mathcal{K}\|_{V,L^2_{\hat\rho}}, \quad \forall \mathcal{K} \in L^2_{\hat\rho} \otimes V_h$$

Proof. We start with the first property.

$$\begin{aligned}
\left\|\Pi_{\tilde{U}^{n+1}Y^{n\intercal}}[\mathcal{K}]\right\|_{V',L^2_{\hat\rho}} &\leq \left\|\mathcal{P}_{\mathcal{Y}^n}[\mathcal{K}]\right\|_{V',L^2_{\hat\rho}} + \left\|\mathcal{P}^{\perp}_{\mathcal{Y}^n}[\mathcal{P}_{\tilde{\mathcal{U}}^{n+1}}[\mathcal{K}]]\right\|_{V',L^2_{\hat\rho}} \\
&\leq \left\|\mathcal{K}\right\|_{V',L^2_{\hat\rho}} + \left\|\mathcal{P}_{\tilde{\mathcal{U}}^{n+1}}[\mathcal{K}]\right\|_{V',L^2_{\hat\rho}} \\
&\leq (1+C_{\mathcal{P}})\left\|\mathcal{K}\right\|_{V',L^2_{\hat\rho}}.
\end{aligned}$$

The second property can be proved in an analogous way. □

4.2 Error estimates without stochastic error contribution

In this section we derive an a-priori error estimation for a fully discrete DLR solution obtained by Algorithm 2.2.1. The error will be measured in a discrete stochastic norm $\|\cdot\|_{L^2_{\hat\rho}}$ which allows us to reuse some of the well-established results concerning a-priori error estimation for a deterministic parabolic equation. In particular, we followed the work presented in [QV08; Qua09].

Theorem 4.2.1. *Let us assume that* $u_{\text{true}} \in L^\infty(0,T;L^2_{\hat\rho}(\hat\Omega;H^{r+1}))$, $\dot{u}_{\text{true}} \in L^1(0,T;L^2_{\hat\rho}(\hat\Omega;H^r))$, $u^0(\omega_i) \in H^r$, $\forall i = 1,\dots,\hat{N}$, $f \in L^\infty(0,T;L^2_{\hat\rho}(\hat\Omega;H))$, *and* $\frac{\partial^2 u_{\text{true}}}{\partial t^2} \in L^2(0,T;L^2_{\hat\rho}(\hat\Omega;H))$. *Then, for the explicit scheme under the condition*

$$\frac{\triangle t}{h^2} \leq \frac{C_{\mathcal{L}}}{2C_{\mathrm{I}}^2 C_{\mathcal{B}}^2}, \tag{4.12}$$

for the semi-implicit scheme under the condition

$$\frac{\triangle t}{h^2} \leq \frac{C_{\mathcal{L}}}{2C_{\mathrm{I}}^2 C_{\text{stoch}}^2}, \tag{4.13}$$

and for the implicit scheme without any condition, the following a-priori error estimate holds

$$\begin{aligned}\|u_{\text{true}}(T) - u^N_{h,\hat{\rho}}\|^2_{H,L^2_{\hat{\rho}}} + \triangle t C_{\mathcal{L}} \sum_{n=1}^{N} \|u_{\text{true}}(t^n) - u^n_{h,\hat{\rho}}\|^2_{V,L^2_{\hat{\rho}}} \\ \leq c_0 \|u^R_{\text{true}}(0) - u_{\text{true}}(0)\|^2_{H,L^2_{\hat{\rho}}} + c_1\varepsilon^2 + c_2\triangle t^2 + c_3 h^{2r},\end{aligned}$$

where $u^R_{\text{true}}(0)$ *is the truncated Karhunen-Loève expansion* (1.4) *of the initial condition* $u_{\text{true}}(0)$*. The constants* c_0, c_1, c_2, c_3 *do not depend on* $\triangle t, h, \varepsilon$*. Their dependence on the choice of the stochastic discretization points* $\{\omega_i\}_{i=1}^{\hat{N}}$ *is specified in Remark 5.*

Proof. We will start with the proof for the implicit scheme obtained by considering $\mathcal{L}(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}}) = \mathcal{L}(u^{n+1}_{h,\hat{\rho}})$ and $f^{n,n+1} = f^{n+1}$. We split the error at time $t = t^n$ into two terms:

$$\|u_{\text{true}}(t^n) - u^n_{h,\hat{\rho}}\|_{H,L^2_{\hat{\rho}}} \leq \|u_{\text{true}}(t^n) - P^r_{1,h} u_{\text{true}}(t^n)\|_{H,L^2_{\hat{\rho}}} + \|P^r_{1,h} u_{\text{true}}(t^n) - u^n_{h,\hat{\rho}}\|_{H,L^2_{\hat{\rho}}}. \tag{4.14}$$

The first term can be estimated by referring to (4.2):

$$\begin{aligned}\|u_{\text{true}}(t^n) - P^r_{1,h} u_{\text{true}}(t^n)\|^2_{H,L^2_{\hat{\rho}}} &= \sum_{i=1}^{\hat{N}} \frac{1}{\hat{N}} \|u_{\text{true}}(t^n,\omega_i) - P^r_{1,h} u_{\text{true}}(t^n,\omega_i)\|^2_H \\ \leq C h^{2r} \sum_{i=1}^{\hat{N}} \frac{1}{\hat{N}} \|u_{\text{true}}(t^n,\omega_i)\|^2_{H^r} &\leq C h^{2r} \|u_{\text{true}}\|_{L^\infty(0,T;L^2_{\hat{\rho}}(\hat{\Omega};H^r))}\end{aligned}$$

and analogously for the $\|\cdot\|_{V,L^2_{\hat{\rho}}}$ -norm

$$\begin{aligned}\|u_{\text{true}}(t^n) - P^r_{1,h} u_{\text{true}}(t^n)\|^2_{V,L^2_{\hat{\rho}}} &\leq C h^{2r} \|u_{\text{true}}(t^n)\|^2_{H^{r+1},L^2_{\hat{\rho}}} \\ &\leq C h^{2r} \|u_{\text{true}}\|_{L^\infty(0,T;L^2_{\hat{\rho}}(\hat{\Omega};H^{r+1}))}\end{aligned}$$

Now, let us focus on the second term. Setting $e^n_{h,\hat{\rho}} := u^n_{h,\hat{\rho}} - P^r_{1,h} u_{\text{true}}(t^n)$, we obtain

$$\Big(\frac{e^{n+1}_{h,\hat{\rho}} - e^n_{h,\hat{\rho}}}{\triangle t} + \mathcal{L}(e^{n+1}_{h,\hat{\rho}}),\, v_{h,\hat{\rho}}\Big)_{V'V,L^2_{\hat{\rho}}} = \Big(\delta^{n+1},\, v_{h,\hat{\rho}}\Big)_{V'V,L^2_{\hat{\rho}}} \quad \forall v_{h,\hat{\rho}} \in L^2_{\hat{\rho}}(\Omega; V_h),$$

where for $\forall v_{h,\hat{\rho}} \in L^2_{\hat{\rho}}(\Omega; V_h)$ it holds

$$\begin{aligned}\Big(\delta^{n+1}, v_{h,\hat{\rho}}\Big)_{V'V,L^2_{\hat{\rho}}} &= \Big(\frac{u^{n+1}_{h,\hat{\rho}} - u^n_{h,\hat{\rho}}}{\triangle t} + \mathcal{L}(u^{n+1}_{h,\hat{\rho}}), v_{h,\hat{\rho}}\Big)_{V'V,L^2_{\hat{\rho}}} \\ &- \Big(\frac{P^r_{1,h}(u_{\text{true}}(t^{n+1}) - u_{\text{true}}(t^n))}{\triangle t} + \mathcal{L}(P^r_{1,h} u_{\text{true}}(t^{n+1})), v_{h,\hat{\rho}}\Big)_{V'V,L^2_{\hat{\rho}}}\end{aligned}$$

$$
\begin{aligned}
&= \Big(\frac{u^{n+1}_{h,\hat\rho} - u^{n}_{h,\hat\rho}}{\triangle t} + \mathbb{E}_{\hat N}[\mathcal{L}(u^{n+1}_{h,\hat\rho})] + \Pi^{h,\hat\rho}_{\tilde U^{n+1}Y^{n\intercal}}[\mathcal{L}^*(u^{n+1}_{h,\hat\rho})] \\
&\qquad\qquad\qquad + {\Pi^{h,\hat\rho}_{\tilde U^{n+1}Y^{n\intercal}}}^{\perp}[\mathcal{L}^*(u^{n+1}_{h,\hat\rho})], v_{h,\hat\rho}\Big)_{V'V,L^2_{\hat\rho}} \\
&\quad - \Big(\frac{P^r_{1,h}(u_{\text{true}}(t^{n+1}) - u_{\text{true}}(t^n))}{\triangle t} + \mathcal{L}(P^r_{1,h}u_{\text{true}}(t^{n+1})), v_{h,\hat\rho}\Big)_{V'V,L^2_{\hat\rho}} \\
&= \Bigg(\mathbb{E}_{\hat N}[f^{n+1}] + \Pi^{h,\hat\rho}_{\tilde U^{n+1}Y^{n\intercal}}[f^{n+1^*}] + {\Pi^{h,\hat\rho}_{\tilde U^{n+1}Y^{n\intercal}}}^{\perp}[\mathcal{L}^*(u^{n+1}_{h,\hat\rho})], v_{h,\hat\rho}\Bigg)_{V'V,L^2_{\hat\rho}} \\
&\quad - \Big(\frac{P^r_{1,h}(u_{\text{true}}(t^{n+1}) - u_{\text{true}}(t^n))}{\triangle t} + \mathcal{L}(u_{\text{true}}(t^{n+1})), v_{h,\hat\rho}\Big)_{V'V,L^2_{\hat\rho}} \\
&= \Bigg(f^{n+1} - \mathcal{L}(u_{\text{true}}(t^{n+1})) - \frac{P^r_{1,h}(u_{\text{true}}(t^{n+1}) - u_{\text{true}}(u_{t^n}))}{\triangle t}, v_{h,\hat\rho}\Bigg)_{V'V,L^2_{\hat\rho}} \\
&\quad + \Bigg({\Pi^{h,\hat\rho}_{\tilde U^{n+1}Y^{n\intercal}}}^{\perp}[\mathcal{L}^*(u^{n+1}_{h,\hat\rho}) - f^{n+1^*}], v_{h,\hat\rho}\Bigg)_{V'V,L^2_{\hat\rho}} \\
&= \Bigg(\dot u_{\text{true}}(t^{n+1}) - \frac{u_{\text{true}}(t^{n+1}) - u_{\text{true}}(t^n)}{\triangle t}, v_{h,\hat\rho}\Bigg)_{V'V,L^2_{\hat\rho}} \\
&\quad + \Bigg((I - P^r_{1,h})\frac{u_{\text{true}}(t^{n+1}) - u_{\text{true}}(t^n)}{\triangle t}, v_{h,\hat\rho}\Bigg)_{V'V,L^2_{\hat\rho}} \\
&\quad + \Bigg({\Pi^{h,\hat\rho}_{\tilde U^{n+1}Y^{n\intercal}}}^{\perp}[\mathcal{L}^*(u^{n+1}_{h,\hat\rho}) - f^{n+1^*}], v_{h,\hat\rho}\Bigg)_{V'V,L^2_{\hat\rho}} \\
&= \Bigg(\frac{1}{\triangle t}\int_{t^n}^{t^{n+1}} (s - t^n)\frac{\partial^2 u_{\text{true}}}{\partial t^2}(s)\,\mathrm{d}s, v_{h,\hat\rho}\Bigg)_{V'V,L^2_{\hat\rho}} \\
&\quad + \Bigg(\frac{1}{\triangle t}\int_{t^n}^{t^{n+1}} (I - P^r_{1,h})(\dot u_{\text{true}})(s)\,\mathrm{d}s, v_{h,\hat\rho}\Bigg)_{V'V,L^2_{\hat\rho}} \\
&\quad + \Bigg({\Pi^{h,\hat\rho}_{\tilde U^{n+1}Y^{n\intercal}}}^{\perp}[\mathcal{L}^*(u^{n+1}_{h,\hat\rho}) - f^{n+1^*}], v_{h,\hat\rho}\Bigg)_{V'V,L^2_{\hat\rho}}.
\end{aligned}
$$

Concerning the last term we split it into

$$
\begin{aligned}
&\Bigg({\Pi^{h,\hat\rho}_{\tilde U^{n+1}Y^{n\intercal}}}^{\perp}[f^{n+1^*} - \mathcal{L}^*(u^{n+1}_{h,\hat\rho})],\, v_{h,\hat\rho}\Bigg)_{V'V,L^2_{\hat\rho}} \\
&\qquad = \underbrace{\Bigg({\Pi^{h,\hat\rho}_{\tilde U^{n+1}Y^{n\intercal}}}^{\perp}[f^{n+1^*} - \mathcal{L}^*(\tilde U^{n+1}Y^{n\intercal})],\, v_{h,\hat\rho}\Bigg)_{V'V,L^2_{\hat\rho}}}_{I} \\
&\qquad\qquad\qquad + \underbrace{\Bigg({\Pi^{h,\hat\rho}_{\tilde U^{n+1}Y^{n\intercal}}}^{\perp}[\mathcal{L}^*(\tilde U^{n+1}Y^{n\intercal} - u^{n+1}_{h,\hat\rho})],\, v_{h,\hat\rho}\Bigg)_{V'V,L^2_{\hat\rho}}}_{II},
\end{aligned}
$$

where, for the estimation of the first term, we apply the low-rank assumption (4.11), i.e.

$$
|I| \leq \varepsilon \|v_{h,\hat\rho}\|_{L^2_{\hat\rho},V}.
$$

Concerning the second term, we proceed as

$$
\begin{aligned}
|II| &= \left|\Big(\mathcal{L}^*(\tilde{U}^{n+1}Y^{n^\intercal} - u_{h,\hat{\rho}}^{n+1}),\, \Pi_{\tilde{U}^{n+1}Y^{n^\intercal}}^{h,\hat{\rho}}{}^{\perp}[v_{h,\hat{\rho}}]\Big)_{V'V,L^2_{\hat{\rho}}}\right| \\
&\leq C_{\mathcal{B}}\|\tilde{U}^{n+1}Y^{n^\intercal} - \tilde{U}^{n+1}\tilde{Y}^{n+1^\intercal}\|_{V,L^2_{\hat{\rho}}}\|\Pi_{\tilde{U}^{n+1}Y^{n^\intercal}}^{h,\hat{\rho}}{}^{\perp}[v_{h,\hat{\rho}}]\|_{V,L^2_{\hat{\rho}}} \\
&\leq C_{\mathcal{B}}\|\tilde{U}^{n+1}Y^{n^\intercal} - \tilde{U}^{n+1}\tilde{Y}^{n+1^\intercal}\|_{V,L^2_{\hat{\rho}}}\|v_{h,\hat{\rho}} - \Pi_{\tilde{U}^{n+1}Y^{n^\intercal}}^{h,\hat{\rho}}[v_{h,\hat{\rho}}]\|_{V,L^2_{\hat{\rho}}} \\
&\leq C_{\mathcal{B}}\|\tilde{U}^{n+1}(Y^{n} - \tilde{Y}^{n+1})^\intercal\|_{V,L^2_{\hat{\rho}}}\,(2+C_{\mathcal{P}})\|v_{h,\hat{\rho}}\|_{V,L^2_{\hat{\rho}}} \\
&= \triangle t\,(2+C_{\mathcal{P}})C_{\mathcal{B}}\left\|\mathcal{P}_{\tilde{\mathcal{U}}^{n+1}}[\mathcal{P}^{\perp}_{\mathcal{Y}^{n^\intercal}}[\mathcal{L}^*(u_{h,\hat{\rho}}^{n+1}) - f^{n+1^*}]]\right\|_{V,L^2_{\hat{\rho}}}\|v_{h,\hat{\rho}}\|_{V,L^2_{\hat{\rho}}} \\
&\leq \triangle t\,C_{\mathcal{P}}(2+C_{\mathcal{P}})C_{\mathcal{B}}\left\|\mathcal{L}^*(u_{h,\hat{\rho}}^{n+1}) - f^{n+1^*}\right\|_{V',L^2_{\hat{\rho}}}\|v_{h,\hat{\rho}}\|_{V,L^2_{\hat{\rho}}} \\
&\leq \triangle t\,C_{\mathcal{P}}(2+C_{\mathcal{P}})C_{\mathcal{B}}\left(\left\|\mathcal{L}^*(u_{h,\hat{\rho}}^{n+1})\right\|_{V',L^2_{\hat{\rho}}} + \left\|f^{n+1^*}\right\|_{L^2_{\hat{\rho}},V'}\right)\|v_{h,\hat{\rho}}\|_{L^2_{\hat{\rho}},V} \\
&\leq \triangle t\,C_{\mathcal{P}}(2+C_{\mathcal{P}})C_{\mathcal{B}}\Big(C_{\mathcal{B}}\|u_{h,\hat{\rho}}^{n+1}\|_{V,L^2_{\hat{\rho}}} + C_{\mathrm{P}}\|f\|_{L^\infty(0,T;L^2_{\hat{\rho}}(\Omega;H))}\Big)\|v_{h,\hat{\rho}}\|_{L^2_{\hat{\rho}},V}
\end{aligned}
\tag{4.15}
$$

In the third and fifth step we applied Lemma 4.1.1.

We introduce the notations

$$
K_1 := C_{\mathcal{P}}(2+C_{\mathcal{P}})C_{\mathcal{B}}, \tag{4.16}
$$

$$
K_2(\|u_{h,\hat{\rho}}^{n+1}\|_{V,L^2_{\hat{\rho}}}, \|f\|_{L^\infty(0,T;L^2_{\hat{\rho}}(\Omega;H))}) := \Big(C_{\mathcal{B}}\|u_{h,\hat{\rho}}^{n+1}\|_{V,L^2_{\hat{\rho}}} + C_{\mathrm{P}}\|f\|_{L^\infty(0,T;L^2_{\hat{\rho}}(\Omega;H))}\Big). \tag{4.17}
$$

Now, let us take $v_{h,\hat{\rho}} = e_{h,\hat{\rho}}^{n+1} \in L^2_{\hat{\rho}}(\Omega; V_h)$ and proceed by

$$
\begin{aligned}
&\frac{1}{2\triangle t}\Big(\|e_{h,\hat{\rho}}^{n+1}\|^2_{H,L^2_{\hat{\rho}}} - \|e_{h,\hat{\rho}}^{n}\|^2_{H,L^2_{\hat{\rho}}}\Big) + C_{\mathcal{L}}\|e_{h,\hat{\rho}}^{n+1}\|^2_{V,L^2_{\hat{\rho}}} \\
&\quad\leq C_{\mathrm{P}}\left\|\frac{1}{\triangle t}\int_{t^n}^{t^{n+1}}(s-t^n)\frac{\partial^2 u_{\mathrm{true}}}{\partial t^2}(s)\,\mathrm{d}s\right\|_{H,L^2_{\hat{\rho}}}\|e_{h,\hat{\rho}}^{n+1}\|_{V,L^2_{\hat{\rho}}} \\
&\qquad + C_{\mathrm{P}}\left\|\frac{1}{\triangle t}\int_{t^n}^{t^{n+1}}(I - P^r_{1,h})(\dot{u}_{\mathrm{true}})(s)\,\mathrm{d}s\right\|_{H,L^2_{\hat{\rho}}}\|e_{h,\hat{\rho}}^{n+1}\|_{V,L^2_{\hat{\rho}}} + \varepsilon\|e_{h,\hat{\rho}}^{n+1}\|_{L^2_{\hat{\rho}},V} \\
&\qquad + \triangle t K_1 K_2(\|u_{h,\hat{\rho}}^{n+1}\|_{V,L^2_{\hat{\rho}}}, \|f\|_{L^\infty(0,T;L^2_{\hat{\rho}}(\Omega;H))})\|e_{h,\hat{\rho}}^{n+1}\|_{L^2_{\hat{\rho}},V} \\
&\quad\leq \frac{2C_{\mathrm{P}}^2}{C_{\mathcal{L}}}\bigg(\int_{t^n}^{t^{n+1}}\|\frac{\partial^2 u_{\mathrm{true}}}{\partial t^2}(s)\|_{H,L^2_{\hat{\rho}}}\mathrm{d}s\bigg)^2 + \frac{C_{\mathcal{L}}}{8}\|e_{h,\hat{\rho}}^{n+1}\|^2_{V,L^2_{\hat{\rho}}} \\
&\qquad + \frac{C_{\mathrm{P}}^2}{C_{\mathcal{L}}}\frac{2}{\triangle t^2}\bigg(\int_{t^n}^{t^{n+1}}\|(I - P^r_{1,h})(\dot{u}_{\mathrm{true}})(s)\|_{H,L^2_{\hat{\rho}}}\,\mathrm{d}s\bigg)^2 + \frac{C_{\mathcal{L}}}{8}\|e_{h,\hat{\rho}}^{n+1}\|^2_{V,L^2_{\hat{\rho}}} \\
&\qquad + \frac{2\varepsilon^2}{C_{\mathcal{L}}} + \frac{C_{\mathcal{L}}}{8}\|e_{h,\hat{\rho}}^{n+1}\|^2_{V,L^2_{\hat{\rho}}} \\
&\qquad + \frac{2\triangle t^2 K_1^2}{C_{\mathcal{L}}}K_2^2(\|u_{h,\hat{\rho}}^{n+1}\|_{V,L^2_{\hat{\rho}}}, \|f\|_{L^\infty(0,T;L^2_{\hat{\rho}}(\Omega;H))}) + \frac{C_{\mathcal{L}}}{8}\|e_{h,\hat{\rho}}^{n+1}\|^2_{V,L^2_{\hat{\rho}}}
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{2C_{\mathrm{P}}^2}{C_{\mathcal{L}}}\triangle t \int_{t^n}^{t^{n+1}} \|\frac{\partial^2 u_{\text{true}}}{\partial t^2}(s)\|^2_{H,L^2_{\hat\rho}} \mathrm{d}s + \frac{C_{\mathrm{P}}^2}{C_{\mathcal{L}}}\frac{2}{\triangle t} h^{2r} \int_{t^n}^{t^{n+1}} |\dot u_{\text{true}}(s)|^2_{H^r,L^2_{\hat\rho}}\,\mathrm{d}s + \frac{2\varepsilon^2}{C_{\mathcal{L}}} \\
&\quad + \frac{2\triangle t^2 K_1^2}{C_{\mathcal{L}}} K_2^2(\|u^{n+1}_{h,\hat\rho}\|_{V,L^2_{\hat\rho}}, \|f\|_{L^\infty(0,T;L^2_{\hat\rho}(\Omega;H))}) + \frac{C_{\mathcal{L}}}{2}\|e^{n+1}_{h,\hat\rho}\|^2_{V,L^2_{\hat\rho}}.
\end{aligned}
$$

Rearranging the terms and summing over $n = 0, \ldots, N-1$ we obtain

$$
\begin{aligned}
\|e^N_{h,\hat\rho}\|^2_{H,L^2_{\hat\rho}} &+ C_{\mathcal{L}} \sum_{n=1}^{N} \triangle t \|e^n_{h,\hat\rho}\|^2_{V,L^2_{\hat\rho}} \leq \|e^0_{h,\hat\rho}\|^2_{H,L^2_{\hat\rho}} + \frac{4C_{\mathrm{P}}^2}{C_{\mathcal{L}}}\triangle t^2 \int_0^T \|\frac{\partial^2 u_{\text{true}}}{\partial t^2}(s)\|^2_{H,L^2_{\hat\rho}} \mathrm{d}s \\
&+ \frac{4C_{\mathrm{P}}^2}{C_{\mathcal{L}}} h^{2r} \int_0^T |\dot u_{\text{true}}(s)|^2_{H^r,L^2_{\hat\rho}}\,\mathrm{d}s + \frac{4T}{C_{\mathcal{L}}}\varepsilon^2 \\
&\qquad + \triangle t^2 \frac{4K_1^2}{C_{\mathcal{L}}} \sum_{n=0}^{N-1} \triangle t K_2^2(\|u^{n+1}_{h,\hat\rho}\|_{V,L^2_{\hat\rho}}, \|f\|_{L^\infty(0,T;L^2_{\hat\rho}(\Omega;H))}) \\
&\leq \|e^0_{h,\hat\rho}\|^2_{H,L^2_{\hat\rho}} + \frac{4C_{\mathrm{P}}^2}{C_{\mathcal{L}}}\triangle t^2 \int_0^T \|\frac{\partial^2 u_{\text{true}}}{\partial t^2}(s)\|^2_{H,L^2_{\hat\rho}} \mathrm{d}s + \frac{4C_{\mathrm{P}}^2}{C_{\mathcal{L}}} h^{2r} \int_0^T |\dot u_{\text{true}}(s)|^2_{H^r,L^2_{\hat\rho}}\,\mathrm{d}s \\
&\quad + \frac{4T}{C_{\mathcal{L}}}\varepsilon^2 + \triangle t^2 \frac{4K_1^2}{C_{\mathcal{L}}}\Big(2C_{\mathcal{B}}^2 \triangle t \sum_{n=0}^{N-1} \|u^{n+1}_{h,\hat\rho}\|^2_{V,L^2_{\hat\rho}} + 2C_{\mathrm{P}}^2 T \|f\|^2_{L^\infty(0,T;L^2_{\hat\rho}(\Omega;H))}\Big) \\
&\leq \|e^0_{h,\hat\rho}\|^2_{H,L^2_{\hat\rho}} + h^{2r}\frac{4C_{\mathrm{P}}^2}{C_{\mathcal{L}}} |\dot u_{\text{true}}|^2_{L^2(0,T;L^2_{\hat\rho}(\Omega;H^r))} + \varepsilon^2 \frac{4T}{C_{\mathcal{L}}} \\
&\quad + \triangle t^2 \bigg(\frac{4C_{\mathrm{P}}^2}{C_{\mathcal{L}}} \|\frac{\partial^2 u_{\text{true}}}{\partial t^2}\|^2_{L^2(0,T;L^2_{\hat\rho}(\Omega;H))} + \frac{4K_1^2}{C_{\mathcal{L}}^2}\Big(C_{\mathcal{B}}^2 \|u^0_{h,\hat\rho}\|^2_{H,L^2_{\hat\rho}} \\
&\qquad + \frac{C_{\mathcal{B}}^2 C_{\mathrm{P}}^2}{C_{\mathcal{L}}} T \|f\|^2_{L^\infty(0,T;L^2_{\hat\rho}(\Omega;H))} + C_{\mathrm{P}}^2 T \|f\|^2_{L^\infty(0,T;L^2_{\hat\rho}(\Omega;H))}\Big)\bigg) \\
&\leq \|e^0_{h,\hat\rho}\|^2_{H,L^2_{\hat\rho}} + h^{2r}\frac{4C_{\mathrm{P}}^2}{C_{\mathcal{L}}} |\dot u_{\text{true}}|^2_{L^2(0,T;L^2_{\hat\rho}(\Omega;H^r))} + \varepsilon^2 \frac{4T}{C_{\mathcal{L}}} \\
&\quad + \triangle t^2 \bigg(\frac{4C_{\mathrm{P}}^2}{C_{\mathcal{L}}} \|\frac{\partial^2 u_{\text{true}}}{\partial t^2}\|^2_{L^2(0,T;L^2_{\hat\rho}(\Omega;H))} + \frac{4K_1^2}{C_{\mathcal{L}}^2}\Big(C_{\mathcal{B}}^2 c_r^2 \|u_{\text{true}}(0)\|^2_{H,L^2_{\hat\rho}} \\
&\qquad + \frac{C_{\mathcal{B}}^2 C_{\mathrm{P}}^2}{C_{\mathcal{L}}} T \|f\|^2_{L^\infty(0,T;L^2_{\hat\rho}(\Omega;H))} + C_{\mathrm{P}}^2 T \|f\|^2_{L^\infty(0,T;L^2_{\hat\rho}(\Omega;H))}\Big)\bigg)
\end{aligned}
$$

In the second step we applied our stability estimate stated in Theorem 3.4.2 to bound the term $\triangle t \sum_{n=1}^{N} \|u^n_{h,\hat\rho}\|^2_{V,L^2_{\hat\rho}}$. In the last step we applied (4.3) to bound the norm $\|u^0_{h,\hat\rho}\|^2_{H,L^2_{\hat\rho}}$ by $c_r^2 \|u_{\text{true}}(0)\|^2_{H,L^2_{\hat\rho}}$.

We follow by bounding the initial error $\|e^0_{h,\hat\rho}\|^2_{H,L^2_{\hat\rho}}$. The initial condition is taken as explained in (4.7), i.e. $u^0_{h,\hat\rho} = \{P^r_h u^R_{\text{true}}(0,\omega_j)\}_{j=1}^{\hat N}$. We can then split the error as

$$
\begin{aligned}
\|e^0_{h,\hat\rho}\|_{H,L^2_{\hat\rho}} &= \|u^0_{h,\hat\rho} - P^r_{1,h} u_{\text{true}}(0)\|_{H,L^2_{\hat\rho}} \\
&= \|P^r_h u^R_{\text{true}}(0) - u^R_{\text{true}}(0)\|_{H,L^2_{\hat\rho}} + \|u^R_{\text{true}}(0) - u_{\text{true}}(0)\|_{H,L^2_{\hat\rho}}
\end{aligned}
$$

$$
\begin{aligned}
&\qquad\qquad + \|u_{\text{true}}(0) - P^r_{1,h} u_{\text{true}}(0)\|_{H,L^2_{\hat\rho}} \\
&\leq Ch^r ||u_{\text{true}}(0)||_{H^r,L^2_{\hat\rho}} + \|u^R_{\text{true}}(0) - u_{\text{true}}(0)\|_{H,L^2_{\hat\rho}},
\end{aligned}
$$

where we used the assumption (4.6). The statement then follows from (4.2) and (4.14).

As for the explicit scheme, we proceed as follows:

Setting as before $e^n_{h,\hat\rho} := u^n_{h,\hat\rho} - P^r_{1,h} u_{\text{true}}(t^n)$, we obtain

$$
\Big(\frac{e^{n+1}_{h,\hat\rho} - e^n_{h,\hat\rho}}{\triangle t} + \mathcal{L}(e^n_{h,\hat\rho}),\, v_{h,\hat\rho}\Big)_{V'V,L^2_{\hat\rho}} = \Big(\delta^n,\, v_{h,\hat\rho}\Big)_{V'V,L^2_{\hat\rho}} \quad \forall v_{h,\hat\rho} \in L^2_{\hat\rho}(\Omega; V_h),
$$

where, following analogous steps as in the derivation for the implicit scheme, for any $v_{h,\hat\rho} \in L^2_{\hat\rho}(\Omega; V_h)$ it holds

$$
\begin{aligned}
\Big(\delta^n,\, v_{h,\hat\rho}\Big)_{V'V,L^2_{\hat\rho}} &= \Big(\frac{u^{n+1}_{h,\hat\rho} - u^n_{h,\hat\rho}}{\triangle t} + \mathcal{L}(u^n_{h,\hat\rho}), v_{h,\hat\rho}\Big)_{V'V,L^2_{\hat\rho}} \\
&\quad - \Big(\frac{P^r_{1,h}(u_{\text{true}}(t^{n+1}) - u_{\text{true}}(t^n))}{\triangle t} + \mathcal{L}(P^r_{1,h} u_{\text{true}}(t^n)), v_{h,\hat\rho}\Big)_{V'V,L^2_{\hat\rho}} \\
&= \Big(\dot u_{\text{true}}(t^n) - \frac{u_{\text{true}}(t^{n+1}) - u_{\text{true}}(t^n)}{\triangle t}, v_{h,\hat\rho}\Big)_{V'V,L^2_{\hat\rho}} \\
&\quad + \Big((I - P^r_{1,h})\frac{u_{\text{true}}(t^{n+1}) - u_{\text{true}}(t^n)}{\triangle t}, v_{h,\hat\rho}\Big)_{V'V,L^2_{\hat\rho}} \\
&\quad + \Big({\Pi^{h,\hat\rho}_{\tilde U^{n+1} Y^{n\intercal}}}^{\perp}[\mathcal{L}^*(u^n_{h,\hat\rho}) - f^{n^*}], v_{h,\hat\rho}\Big)_{V'V,L^2_{\hat\rho}} \\
&= \Big(\frac{1}{\triangle t}\int_{t^n}^{t^{n+1}} (t^{n+1} - s)\frac{\partial^2 u_{\text{true}}}{\partial t^2}(s)\,\mathrm{d}s, v_{h,\hat\rho}\Big)_{V'V,L^2_{\hat\rho}} \\
&\quad + \Big(\frac{1}{\triangle t}\int_{t^n}^{t^{n+1}} (I - P^r_{1,h})(\dot u_{\text{true}})(s)\,\mathrm{d}s, v_{h,\hat\rho}\Big)_{V'V,L^2_{\hat\rho}} \\
&\quad + \Big({\Pi^{h,\hat\rho}_{\tilde U^{n+1} Y^{n\intercal}}}^{\perp}[\mathcal{L}^*(u^n_{h,\hat\rho}) - f^{n^*}], v_{h,\hat\rho}\Big)_{V'V,L^2_{\hat\rho}}.
\end{aligned}
$$

Concerning the last term, we split it as before into

$$
\begin{aligned}
&\Big({\Pi^{h,\hat\rho}_{\tilde U^{n+1} Y^{n\intercal}}}^{\perp}[f^{n^*} - \mathcal{L}^*(u^n_{h,\hat\rho})],\, v_{h,\hat\rho}\Big)_{V'V,L^2_{\hat\rho}} \\
&\qquad = \underbrace{\Big({\Pi^{h,\hat\rho}_{\tilde U^{n+1} Y^{n\intercal}}}^{\perp}[f^{n^*} - \mathcal{L}^*(\tilde U^{n+1} Y^{n^\intercal})],\, v_{h,\hat\rho}\Big)_{V'V,L^2_{\hat\rho}}}_{I} \\
&\qquad\qquad + \underbrace{\Big({\Pi^{h,\hat\rho}_{\tilde U^{n+1} Y^{n\intercal}}}^{\perp}[\mathcal{L}^*(\tilde U^{n+1} Y^{n^\intercal} - u^n_{h,\hat\rho})],\, v_{h,\hat\rho}\Big)_{V'V,L^2_{\hat\rho}}}_{II},
\end{aligned}
$$

where, for the estimation of the first term, we apply the low-rank assumption (4.11), i.e.

$$|I| \leq \varepsilon \|v_{h,\hat{\rho}}\|_{L^2_{\hat{\rho}},V}. \tag{4.18}$$

Concerning the second term, we proceed analogously to the implicit case and derive

$$\begin{aligned}
|II| &= \left| \left(\mathcal{L}^*(\tilde{U}^{n+1} Y^{n^\intercal} - u^n_{h,\hat{\rho}}),\, {\Pi^{h,\hat{\rho}}_{\tilde{U}^{n+1}Y^{n^\intercal}}}^{\perp}[v_{h,\hat{\rho}}] \right)_{V'V,L^2_{\hat{\rho}}} \right| \\
&\leq C_{\mathcal{B}} \|\tilde{U}^{n+1} Y^{n^\intercal} - U^n Y^{n^\intercal}\|_{V,L^2_{\hat{\rho}}} \|v_{h,\hat{\rho}} - \Pi^{h,\hat{\rho}}_{\tilde{U}^{n+1}Y^{n^\intercal}}[v_{h,\hat{\rho}}]\|_{V,L^2_{\hat{\rho}}} \\
&\leq C_{\mathcal{B}} \|(\tilde{U}^{n+1} - U^n) Y^n)^\intercal\|_{V,L^2_{\hat{\rho}}} \, (2 + C_{\mathcal{P}}) \|v_{h,\hat{\rho}}\|_{V,L^2_{\hat{\rho}}} \\
&= \triangle t\, (2 + C_{\mathcal{P}}) C_{\mathcal{B}} \left\| \mathcal{P}_{\mathcal{Y}^n} [\mathcal{L}^*(u^n_{h,\hat{\rho}}) - f^{n^*}] \right\|_{V',L^2_{\hat{\rho}}} \|v_{h,\hat{\rho}}\|_{V,L^2_{\hat{\rho}}} \\
&\leq \triangle t\, (2 + C_{\mathcal{P}}) C_{\mathcal{B}} \left\| \mathcal{L}^*(u^n_{h,\hat{\rho}}) - f^{n^*} \right\|_{V',L^2_{\hat{\rho}}} \|v_{h,\hat{\rho}}\|_{V,L^2_{\hat{\rho}}} \\
&\leq \triangle t\, (2 + C_{\mathcal{P}}) C_{\mathcal{B}} \left(\left\| \mathcal{L}^*(u^n_{h,\hat{\rho}}) \right\|_{V',L^2_{\hat{\rho}}} + \left\| f^{n^*} \right\|_{L^2_{\hat{\rho}},V'} \right) \|v_{h,\hat{\rho}}\|_{L^2_{\hat{\rho}},V} \\
&\leq \triangle t\, (2 + C_{\mathcal{P}}) C_{\mathcal{B}} \Big(C_{\mathcal{B}} \|u^n_{h,\hat{\rho}}\|_{V,L^2_{\hat{\rho}}} + C_{\mathrm{P}} \|f\|_{L^\infty(0,T;L^2_{\hat{\rho}}(\Omega;H))} \Big) \|v_{h,\hat{\rho}}\|_{L^2_{\hat{\rho}},V}.
\end{aligned} \tag{4.19}$$

In the following computation, we employ $K_3 := (2 + C_{\mathcal{P}}) C_{\mathcal{B}}$

Now, let us take $v_{h,\hat{\rho}} = e^{n+1}_{h,\hat{\rho}} \in L^2_{\hat{\rho}}(\Omega; V_h)$ and proceed by

$$\begin{aligned}
&\frac{1}{2\triangle t} \Big(\|e^{n+1}_{h,\hat{\rho}}\|^2_{H,L^2_{\hat{\rho}}} - \|e^n_{h,\hat{\rho}}\|^2_{H,L^2_{\hat{\rho}}} + ||e^{n+1}_{h,\hat{\rho}} - e^n_{h,\hat{\rho}}||^2_{H,L^2_{\hat{\rho}}} \Big) + \langle e^n_{h,\hat{\rho}}, e^{n+1}_{h,\hat{\rho}} \rangle_{\mathcal{L},\hat{\rho}} \\
&\quad \leq C_{\mathrm{P}} \left\| \frac{1}{\triangle t} \int_{t^n}^{t^{n+1}} (t^{n+1} - s) \frac{\partial^2 u_{\mathrm{true}}}{\partial t^2}(s)\, \mathrm{d}s \right\|_{H,L^2_{\hat{\rho}}} \|e^{n+1}_{h,\hat{\rho}}\|_{V,L^2_{\hat{\rho}}} \\
&\qquad + C_{\mathrm{P}} \left\| \frac{1}{\triangle t} \int_{t^n}^{t^{n+1}} (I - P^r_{1,h})(\dot{u}_{\mathrm{true}})(s)\, \mathrm{d}s \right\|_{H,L^2_{\hat{\rho}}} \|e^{n+1}_{h,\hat{\rho}}\|_{V,L^2_{\hat{\rho}}} + \varepsilon \|e^{n+1}_{h,\hat{\rho}}\|_{L^2_{\hat{\rho}},V} \\
&\qquad + \triangle t K_3 K_2 (\|u^n_{h,\hat{\rho}}\|_{V,L^2_{\hat{\rho}}}, \|f\|_{L^\infty(0,T;L^2_{\hat{\rho}}(\Omega;H))}) \|e^{n+1}_{h,\hat{\rho}}\|_{L^2_{\hat{\rho}},V} \\
&\quad \leq \frac{4C^2_{\mathrm{P}}}{C_{\mathcal{L}}} \triangle t \int_{t^n}^{t^{n+1}} \| \frac{\partial^2 u_{\mathrm{true}}}{\partial t^2}(s) \|^2_{H,L^2_{\hat{\rho}}} \mathrm{d}s + \frac{C^2_{\mathrm{P}}}{C_{\mathcal{L}}} \frac{4}{\triangle t} h^{2r} \int_{t^n}^{t^{n+1}} |\dot{u}_{\mathrm{true}}(s)|^2_{H^r,L^2_{\hat{\rho}}}\, \mathrm{d}s + \frac{4\varepsilon^2}{C_{\mathcal{L}}} \\
&\qquad + \frac{4\triangle t^2 K^2_3}{C_{\mathcal{L}}} K^2_2 (\|u^n_{h,\hat{\rho}}\|_{V,L^2_{\hat{\rho}}}, \|f\|_{L^\infty(0,T;L^2_{\hat{\rho}}(\Omega;H))}) + \frac{C_{\mathcal{L}}}{4} \|e^{n+1}_{h,\hat{\rho}}\|^2_{V,L^2_{\hat{\rho}}}
\end{aligned}$$

Applying the same calculations as in (3.12) with $\kappa = 1/2$, we derive

$$\frac{1}{2\triangle t} ||e^{n+1}_{h,\hat{\rho}} - e^n_{h,\hat{\rho}}||^2_{H,L^2_{\hat{\rho}}} + \langle e^n_{h,\hat{\rho}}, e^{n+1}_{h,\hat{\rho}} \rangle_{\mathcal{L},\hat{\rho}} \geq C_{\mathcal{L}} \frac{3}{4} \|e^{n+1}_{h,\hat{\rho}}\|^2_{V,L^2_{\hat{\rho}}},$$

which holds true thanks to the condition on the time step (4.12). The rest of the proof is the analogous to the implicit case, employing the stability estimate for the explicit scheme proved in Theorem 3.4.3.

The last scheme to analyze is the semi-implicit scheme with $\mathcal{L}(u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho}) = \mathcal{L}_{\text{det}}(u^{n+1}_{h,\hat\rho}) + \mathcal{L}_{\text{stoch}}(u^n_{h,\hat\rho})$ and $f^{n,n+1} = f^{n+1}$. The error term $e^n_{h,\hat\rho}$ satisfies

$$\Big(\frac{e^{n+1}_{h,\hat\rho} - e^n_{h,\hat\rho}}{\triangle t} + \mathcal{L}(e^{n+1}_{h,\hat\rho}) - \mathcal{L}_{\text{stoch}}(e^{n+1}_{h,\hat\rho} - e^n_{h,\hat\rho}),\, v_{h,\hat\rho}\Big)_{V'V,L^2_{\hat\rho}} = \Big(\delta^{n+1},\, v_{h,\hat\rho}\Big)_{V'V,L^2_{\hat\rho}} \quad \forall v_{h,\hat\rho} \in L^2_{\hat\rho}(\Omega; V_h), \tag{4.20}$$

where for any $v_{h,\hat\rho} \in L^2_{\hat\rho}(\Omega; V_h)$ it holds

$$\begin{aligned}
\Big(\delta^{n+1},\, v_{h,\hat\rho}\Big)_{V'V,L^2_{\hat\rho}} &= \Big(\frac{u^{n+1}_{h,\hat\rho} - u^n_{h,\hat\rho}}{\triangle t} + \mathcal{L}_{\text{det}}(u^{n+1}_{h,\hat\rho}) + \mathcal{L}_{\text{stoch}}(u^n_{h,\hat\rho}), v_{h,\hat\rho}\Big)_{V'V,L^2_{\hat\rho}} \\
&\quad - \Big(\frac{P^r_{1,h}(u_{\text{true}}(t^{n+1}) - u_{\text{true}}(t^n))}{\triangle t} + \mathcal{L}(P^r_{1,h} u_{\text{true}}(t^{n+1})) \\
&\quad + \mathcal{L}_{\text{stoch}}(P^r_{1,h}(u_{\text{true}}(t^{n+1}) - u_{\text{true}}(t^n))), v_{h,\hat\rho}\Big)_{V'V,L^2_{\hat\rho}} \\
&= \Bigg(\mathbb{E}_{\hat N}[f^{n+1}] + \Pi^{h,\hat\rho}_{\tilde U^{n+1} Y^{n\mathsf{T}}}[f^{n+1^*}] \\
&\quad + \Pi^{h,\hat\rho}_{\tilde U^{n+1} Y^{n\mathsf{T}}}{}^{\perp}[\mathcal{L}^*_{\text{det}}(u^{n+1}_{h,\hat\rho}) + \mathcal{L}^*_{\text{stoch}}(u^n_{h,\hat\rho})], v_{h,\hat\rho}\Big)_{V'V,L^2_{\hat\rho}} \\
&\quad - \Big(\frac{P^r_{1,h}(u_{\text{true}}(t^{n+1}) - u_{\text{true}}(t^n))}{\triangle t} + \mathcal{L}(u_{\text{true}}(t^n)) \\
&\quad + \mathcal{L}_{\text{stoch}}(P^r_{1,h}(u_{\text{true}}(t^{n+1}) - u_{\text{true}}(t^n))), v_{h,\hat\rho}\Big)_{V'V,L^2_{\hat\rho}} \\
&= \Bigg(f^{n+1} - \mathcal{L}(u_{\text{true}}(t^n)) - \frac{P^r_{1,h}(u_{\text{true}}(t^{n+1}) - u_{\text{true}}(u_{t^n}))}{\triangle t}, v_{h,\hat\rho}\Bigg)_{V'V,L^2_{\hat\rho}} \\
&\quad + \Bigg(\Pi^{h,\hat\rho}_{\tilde U^{n+1} Y^{n\mathsf{T}}}{}^{\perp}[\mathcal{L}^*_{\text{det}}(u^{n+1}_{h,\hat\rho}) + \mathcal{L}^*_{\text{stoch}}(u^n_{h,\hat\rho}) - f^{n+1^*}], v_{h,\hat\rho}\Bigg)_{V'V,L^2_{\hat\rho}} \\
&\quad - \Bigg(\mathcal{L}_{\text{stoch}}(P^r_{1,h}(u_{\text{true}}(t^{n+1}) - u_{\text{true}}(t^n))), v_{h,\hat\rho}\Bigg)_{V'V,L^2_{\hat\rho}} \\
&= \Bigg(\dot u_{\text{true}}(t^n) - \frac{u_{\text{true}}(t^{n+1}) - u_{\text{true}}(t^n)}{\triangle t}, v_{h,\hat\rho}\Bigg)_{V'V,L^2_{\hat\rho}} \\
&\quad + \Bigg((I - P^r_{1,h})\frac{u_{\text{true}}(t^{n+1}) - u_{\text{true}}(t^n)}{\triangle t}, v_{h,\hat\rho}\Bigg)_{V'V,L^2_{\hat\rho}} \\
&\quad + \Bigg(\Pi^{h,\hat\rho}_{\tilde U^{n+1} Y^{n\mathsf{T}}}{}^{\perp}[\mathcal{L}^*_{\text{det}}(u^{n+1}_{h,\hat\rho}) + \mathcal{L}^*_{\text{stoch}}(u^n_{h,\hat\rho}) - f^{n+1^*}], v_{h,\hat\rho}\Bigg)_{V'V,L^2_{\hat\rho}} \\
&\quad - \Bigg(\mathcal{L}_{\text{stoch}}(P^r_{1,h}(u_{\text{true}}(t^{n+1}) - u_{\text{true}}(t^n))), v_{h,\hat\rho}\Bigg)_{V'V,L^2_{\hat\rho}} \\
&= \Bigg(\frac{1}{\triangle t}\int_{t^n}^{t^{n+1}} (t^{n+1} - s)\frac{\partial^2 u_{\text{true}}}{\partial t^2}(s)\,\mathrm{d}s, v_{h,\hat\rho}\Bigg)_{V'V,L^2_{\hat\rho}} \\
&\quad + \Bigg(\frac{1}{\triangle t}\int_{t^n}^{t^{n+1}} (I - P^r_{1,h})(\dot u_{\text{true}})(s)\,\mathrm{d}s, v_{h,\hat\rho}\Bigg)_{V'V,L^2_{\hat\rho}} \\
&\quad + \Bigg(\Pi^{h,\hat\rho}_{\tilde U^{n+1} Y^{n\mathsf{T}}}{}^{\perp}[\mathcal{L}^*_{\text{det}}(u^{n+1}_{h,\hat\rho}) + \mathcal{L}^*_{\text{stoch}}(u^n_{h,\hat\rho}) - f^{n+1^*}], v_{h,\hat\rho}\Bigg)_{V'V,L^2_{\hat\rho}}
\end{aligned}$$

$$- \Big(\mathcal{L}_{\text{stoch}}\Big(\int_{t^n}^{t^{n+1}} P^r_{1,h}(\dot{u}_{\text{true}})(s)\,\mathrm{d}s\Big), v_{h,\hat{\rho}}\Big)_{V'V,L^2_{\hat{\rho}}}$$

Concerning the second to last term, we split it into

$$\begin{aligned}
&\Big({\Pi^{h,\hat{\rho}}_{\tilde{U}^{n+1}Y^{n\intercal}}}^{\perp}[f^{n+1^*} - \mathcal{L}^*_{\text{det}}(u^{n+1}_{h,\hat{\rho}}) - \mathcal{L}^*_{\text{stoch}}(u^n_{h,\hat{\rho}})],\, v_{h,\hat{\rho}}\Big)_{V'V,L^2_{\hat{\rho}}} \\
&\quad = \underbrace{\Big({\Pi^{h,\hat{\rho}}_{\tilde{U}^{n+1}Y^{n\intercal}}}^{\perp}[f^{n+1^*} - \mathcal{L}^*(\tilde{U}^{n+1}Y^{n\intercal})],\, v_{h,\hat{\rho}}\Big)_{V'V,L^2_{\hat{\rho}}}}_{I} \\
&\qquad + \underbrace{\Big({\Pi^{h,\hat{\rho}}_{\tilde{U}^{n+1}Y^{n\intercal}}}^{\perp}[\mathcal{L}^*_{\text{det}}(\tilde{U}^{n+1}Y^{n\intercal} - u^{n+1}_{h,\hat{\rho}})],\, v_{h,\hat{\rho}}\Big)_{V'V,L^2_{\hat{\rho}}}}_{II} \\
&\qquad\quad + \underbrace{\Big({\Pi^{h,\hat{\rho}}_{\tilde{U}^{n+1}Y^{n\intercal}}}^{\perp}[\mathcal{L}^*_{\text{stoch}}(\tilde{U}^{n+1}Y^{n\intercal} - u^n_{h,\hat{\rho}}))],\, v_{h,\hat{\rho}}\Big)_{V'V,L^2_{\hat{\rho}}}}_{III}.
\end{aligned}$$

In the estimation of I, we proceed in the same way as in (4.18). The estimation of II is performed equivalently to (4.15), where instead of $C_{\mathcal{B}}$ we use the constant $C_{\text{det},\mathcal{B}}$. Lastly, the term III can be bound equivalently to (4.19), using C_{stoch} instead of $C_{\mathcal{B}}$.

Now, let us consider $v_{h,\hat{\rho}} = e^{n+1}_{h,\hat{\rho}} \in L^2_{\hat{\rho}}(\Omega; V_h)$. Applying the same computations as in the proof of Theorem 3.4.4 and using the time-step condition (4.13), we see that the left hand side of (4.20) can be bounded from below by

$$\begin{aligned}
&\frac{1}{2\triangle t}\Big(\|e^{n+1}_{h,\hat{\rho}}\|^2_{H,L^2_{\hat{\rho}}} - \|e^n_{h,\hat{\rho}}\|^2_{H,L^2_{\hat{\rho}}}\Big) + \frac{3}{4}C_{\mathcal{L}}\|e^{n+1}_{h,\hat{\rho}}\|^2_{V,L^2_{\hat{\rho}}} \\
&\qquad \le \Big(\frac{e^{n+1}_{h,\hat{\rho}} - e^n_{h,\hat{\rho}}}{\triangle t} + \mathcal{L}(e^{n+1}_{h,\hat{\rho}}) - \mathcal{L}_{\text{stoch}}(e^{n+1}_{h,\hat{\rho}} - e^n_{h,\hat{\rho}}),\, e^{n+1}_{h,\hat{\rho}}\Big)_{V'V,L^2_{\hat{\rho}}}
\end{aligned}$$

Concerning the right-hand side, we then proceed as

$$\begin{aligned}
&\Big(\delta^{n+1},\, e^{n+1}_{h,\hat{\rho}}\Big)_{V'V,L^2_{\hat{\rho}}} \\
&\quad \le C_{\text{P}}\Big\|\frac{1}{\triangle t}\int_{t^n}^{t^{n+1}}(t^{n+1}-s)\frac{\partial^2 u_{\text{true}}}{\partial t^2}(s)\,\mathrm{d}s\Big\|_{H,L^2_{\hat{\rho}}}\|e^{n+1}_{h,\hat{\rho}}\|_{V,L^2_{\hat{\rho}}} \\
&\qquad + C_{\text{P}}\Big\|\frac{1}{\triangle t}\int_{t^n}^{t^{n+1}}(I - P^r_{1,h})(\dot{u}_{\text{true}})(s)\,\mathrm{d}s\Big\|_{H,L^2_{\hat{\rho}}}\|e^{n+1}_{h,\hat{\rho}}\|_{V,L^2_{\hat{\rho}}} + \varepsilon\|e^{n+1}_{h,\hat{\rho}}\|_{L^2_{\hat{\rho}},V} \\
&\qquad + 2\triangle t\tilde{K}_1\tilde{K}_2(\|u^n_{h,\hat{\rho}}\|_{V,L^2_{\hat{\rho}}}, \|f\|_{L^\infty(0,T;L^2_{\hat{\rho}}(\Omega;H))})\|e^{n+1}_{h,\hat{\rho}}\|_{L^2_{\hat{\rho}},V} \\
&\qquad + C_{\text{stoch}}c_r C_{\text{P}}\Big\|\int_{t^n}^{t^{n+1}}(P^r_{1,h})(\dot{u}_{\text{true}})(s)\,\mathrm{d}s\Big\|_{H,L^2_{\hat{\rho}}}\|e^{n+1}_{h,\hat{\rho}}\|_{V,L^2_{\hat{\rho}}}
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{8C_{\mathrm{P}}^2}{C_{\mathcal{L}}}\triangle t \int_{t^n}^{t^{n+1}} \|\frac{\partial^2 u_{\mathrm{true}}}{\partial t^2}(s)\|^2_{H,L^2_{\hat{\rho}}} \mathrm{d}s + \frac{C_{\mathrm{P}}^2}{C_{\mathcal{L}}}\frac{8}{\triangle t} h^{2r} \int_{t^n}^{t^{n+1}} |\dot{u}_{\mathrm{true}}(s)|^2_{H^r,L^2_{\hat{\rho}}}\, \mathrm{d}s + \frac{8\varepsilon^2}{C_{\mathcal{L}}} \\
&\quad + \frac{16\triangle t^2 K_1^2}{C_{\mathcal{L}}} K_2^2(\|u^n_{h,\hat{\rho}}\|_{V,L^2_{\hat{\rho}}}, \|f\|_{L^\infty(0,T;L^2_{\hat{\rho}}(\Omega;H))}) + \frac{8C_{\mathrm{stoch}}^2 c_r^2 C_{\mathrm{P}}^2}{C_{\mathcal{L}}}\triangle t \int_{t^n}^{t^{n+1}} \|\dot{u}_{\mathrm{true}}(s)\|^2_{V,L^2_{\hat{\rho}}}\, \mathrm{d}s \\
&\quad + \frac{C_{\mathcal{L}}}{4}\|e^{n+1}_{h,\hat{\rho}}\|^2_{V,L^2_{\hat{\rho}}},
\end{aligned}
$$

where $\tilde{K}_1, \tilde{K}_2$ are defined analogously to (4.16), stemming from the estimation of II and III. The rest of the proof is analogous to the implicit or explicit case, employing the stability estimate for the semi-implicit scheme proved in Theorem 3.4.4. $\square$

Remark 5. In this remark we detail how the final constants c_0, c_1, c_2, c_3 appearing in Theorem 4.2.1 depend on the choice and number of the samples $\{\omega_j\}_{j=1}^{\hat{N}}$. As a matter of fact, they all depend on some discrete norm $\|\cdot\|_{H,L^2_{\hat{\rho}}}, \|\cdot\|_{V,L^2_{\hat{\rho}}}$ of the true solution or forcing term. In particular,

$$
\begin{aligned}
c_2 &= C_1\|\frac{\partial^2 u_{\mathrm{true}}}{\partial t^2}\|^2_{L^2(0,T;L^2_{\hat{\rho}}(\Omega;H))} + C_2\|u_{\mathrm{true}}(0)\|^2_{H,L^2_{\hat{\rho}}} + C_3\|f\|^2_{L^\infty(0,T;L^2_{\hat{\rho}}(\Omega;H))} \\
&\qquad + C_4\|\dot{u}_{\mathrm{true}}\|^2_{L^2(0,T;L^2_{\hat{\rho}}(\Omega;V))} \\
c_3 &= C_5\|\dot{u}_{\mathrm{true}}\|^2_{L^2(0,T;L^2_{\hat{\rho}}(\Omega;H^r))} + C_6\|u_{\mathrm{true}}(0)\|_{H^r,L^2_{\hat{\rho}}}.
\end{aligned}
$$

The constants c_0, c_1 and the new introduced constants $C_i, i = 1, \ldots, 6$ do not depend on the choice of the sampling points.

Let us comment on the comparison of Theorem 4.2.1 and Theorem 2.1.4, published in [KLW16]. Both of these results show that the approximation error is of $O(\varepsilon + \triangle t)$ with constants independent of the smallest singular value. As pointed out by their authors, a limitation of their theoretical result is that it requires a (local) Lipschitz condition on $\mathcal{F}$, and is applicable to stiff differential equations such as discretized PDEs only under a severe CFL condition $\triangle t L \ll 1$, where L is the Lipschitz constant for $\mathcal{F}$. Such a restriction is not present in our analysis. Furthermore, in our setting, the operator $\mathcal{F}$ does not need to be uniformly bounded. On the other hand, we assume our operator to be elliptic. By restricting ourselves to a parabolic problem, we have managed to bound the error in a stronger norm. We have considered a problem set in an infinite-dimensional setting, and provided convergence w.r.t. the spatial discretization parameter h (and the number of collocation points $\hat{N}$ in the next section).

Let us consider now an approximation of the solution in a DO format (as opposed to the Dual DO format used so far), i.e.

$$
u(t) = \bar{u}(t) + \sum_{j=1}^{R} U_j(t) Y_j(t)
$$

with the deterministic basis orthonormal in H, $\langle U_i, U_j \rangle_H = \mathrm{Id}$, Y_i with zero mean and the covariance matrix $\langle Y^\intercal, Y\rangle_{L^2_\rho}$ full rank. In this case, we should use a variant of the projector-splitting scheme summarized in Algorithm 2.2.1, which updates first the stochastic basis and then the deterministic basis. The discrete DLR approximation in the DO format satisfies then a variational formulation analogous to (2.27)

$$\Big\langle \frac{u^{n+1}_{h,\hat\rho} - u^n_{h,\hat\rho}}{\triangle t}, v_{h,\hat\rho} \Big\rangle_{H,L^2_{\hat\rho}} + \Big(\mathcal{L}(u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho}), v_{h,\hat\rho}\Big)_{V'V,L^2_{\hat\rho}} = \Big\langle f^{n,n+1}, v_{h,\hat\rho}\Big\rangle_{H,L^2_{\hat\rho}},$$
$$\forall v_{h,\hat\rho} = \bar v_h + {v_{h,\hat\rho}}^* \text{ with } \bar v_h \in V_h \text{ and } {v_{h,\hat\rho}}^* \in \mathcal{T}_{U^n \tilde Y^{n+1\intercal}} \mathcal{M}^{h,\hat\rho}_R.$$

The only change w.r.t. (2.27) is in the definition of the 'intermediate' point $U^n \tilde Y^{n+1\intercal}$ characterising the tangent space. Employing the DO format instead of the dual DO format allows us to simplify assumptions (4.9) – (4.10), as they follow from the simpler assumption that the deterministic basis remains bounded in the V-norm at all times.

Lemma 4.2.2. *Let us assume*

$$\|U^n\|_V \le C_{\mathcal{P}}, \qquad \forall n = 0, \dots, N. \tag{4.21}$$

Then the following holds

$$\Big\|\Pi_{U^n\tilde Y^{n+1\intercal}}[\mathcal{K}]\Big\|_{V',L^2_{\hat\rho}} \le k_1 \|\mathcal{K}\|_{V',L^2_{\hat\rho}}, \quad \forall \mathcal{K} \in V'_h \otimes L^2_{\hat\rho}$$
$$\Big\|\Pi_{U^n\tilde Y^{n+1\intercal}}[\mathcal{K}]\Big\|_{V,L^2_{\hat\rho}} \le k_2 \|\mathcal{K}\|_{V,L^2_{\hat\rho}}, \quad \forall \mathcal{K} \in V_h \otimes L^2_{\hat\rho}$$

for some k_1, k_2 independent of n.

Proof.

$$\begin{aligned}
\Big\|\Pi_{U^n\tilde Y^{n+1\intercal}}[\mathcal{K}]\Big\|_{V',L^2_{\hat\rho}} &\le \Big\|\mathcal{P}_{\tilde{\mathcal{Y}}^{n+1}}[\mathcal{K}] + \mathcal{P}^\perp_{\tilde{\mathcal{Y}}^{n+1}}[\mathcal{P}_{\mathcal{U}^n}[\mathcal{K}]]\Big\|_{V',L^2_{\hat\rho}} \\
&\le \|\mathcal{P}_{\tilde{\mathcal{Y}}^{n+1}}[\mathcal{K}]\|_{V',L^2_{\hat\rho}} + \|\mathcal{P}^\perp_{\tilde{\mathcal{Y}}^{n+1}}[\mathcal{P}_{\mathcal{U}^n}[\mathcal{K}]]\|_{V',L^2_{\hat\rho}} \\
&\le \|\mathcal{K}\|_{V',L^2_{\hat\rho}} + \|\mathcal{P}_{\mathcal{U}^n}[\mathcal{K}]\|_{V',L^2_{\hat\rho}} \\
&= \|\mathcal{K}\|_{V',L^2_{\hat\rho}} + \Big\|\sum_{i=1}^R \langle \mathcal{K}, U^n\rangle_{V',V} U^n\Big\|_{V',L^2_{\hat\rho}} \\
&\le \|\mathcal{K}\|_{V',L^2_{\hat\rho}} + \sum_{i=1}^R \|\mathcal{K}\|_{V',L^2_{\hat\rho}} \underbrace{\|U^n\|_V}_{\le C_{\mathcal{P}}} c \underbrace{\|U^n\|_H}_{=1} \\
&\le k_1 \|\mathcal{K}\|_{V',L^2_{\hat\rho}},
\end{aligned}$$

where the constant c is the constant from the continuous embedding of $H \hookrightarrow V'$. Analogous result holds for the second inequality, using the fact that $V \hookrightarrow H$. □

We see that, for the DLR solution in the DO format, we can bound the projection on the tangent space by bounding the V- norm of the deterministic modes. Note that the constant $C_{\mathcal{P}}$ in (4.21) can be again bounded using the inverse inequality (2.2)

$$\|U^n\|_V \leq \frac{c}{h^p}\|U^n\|_H \leq \frac{c}{h^p},$$

which, however, leads to a suboptimal result.

4.3 Error estimate with stochastic error contribution

This subsection is dedicated to studying the error contribution caused by the stochastic discretization which applies the Monte-Carlo method. The points $\{\omega_j\}_{j=1}^{\hat{N}} \subset \Omega$ are chosen as i.i.d. samples from ρ. The fully discrete DLR solution obtained by Algorithm 2.2.1 is a collection of $\hat{N}$ functions

$$u^n_{h,\hat{\rho}} = \Big((u^n_{h,\hat{\rho}})_{(1)}, \ldots, (u^n_{h,\hat{\rho}})_{(\hat{N})}\Big) \in V_h \times V_h \times \cdots \times V_h, \qquad \forall n \in \mathbb{N}.$$

By $u_{\text{true},\hat{\rho}} = ((u_{\text{true},\hat{\rho}})_{(1)}, \ldots, (u_{\text{true},\hat{\rho}})_{(\hat{N})})$ let us denote the $\hat{N}$-tuple of paths

$$\begin{aligned}((u_{\text{true},\hat{\rho}})_{(1)}, \ldots, (u_{\text{true},\hat{\rho}})_{(\hat{N})}) &= (u_{\text{true}}(\cdot,\omega_1), \ldots, u_{\text{true}}(\cdot,\omega_{\hat{N}})) \\ &\qquad \in L^2(0,T;V) \times \ldots, \times L^2(0,T;V).\end{aligned}$$

We express the stochastic discretization error w.r.t. a Lipschitz functional $\Phi : H \to \mathbb{R}$

$$|\Phi(u) - \Phi(v)| \leq C_{lip,1}\|u - v\|_H. \tag{4.22}$$

Note that consequently Φ satisfies

$$|\Phi(u)| \leq C_{lip,2}(1 + \|u\|_H).$$

As functions of the sampling points $\{\omega_i\}_{i=1}^{\hat{N}}$, both $u_{\text{true},\hat{\rho}}$ and $\{u^n_{h,\hat{\rho}}\}_{n=1}^{N}$ are random variables with an underlying probability space

$$(\Omega \times \cdots \times \Omega, \mathcal{F} \otimes \cdots \otimes \mathcal{F}, \rho \otimes \cdots \otimes \rho).$$

By $\mathbb{E}_{\otimes\rho}$ we denote the expectation with respect to $\rho \otimes \cdots \otimes \rho$ on $\Omega \times \cdots \times \Omega$.

Theorem 4.3.1. *Consider a Lipschitz functional $\Phi : H \to \mathbb{R}$ satisfying* (4.22). *Let $m^n_{h,\hat{\rho}} := \mathbb{E}_{\hat{\rho}}[\Phi(u^n_{h,\hat{\rho}})]$ denote the sample mean of Φ evaluated in the discrete DLR solution at time $t = t^n$ obtained by Algorithm 2.2.1, and $m^n := \mathbb{E}_\rho[\Phi(u_{\text{true}}(t^n))]$ the mean value of Φ evaluated at the true solution u_{true} at time $t = t^n$. Then for the quantities $m^n_{h,\hat{\rho}}$ and m^n, the following mean-square error estimate holds*

$$\sqrt{\mathbb{E}_{\otimes\rho}\left[\left|m^n - m^n_{h,\hat{\rho}}\right|^2\right]} \leq \frac{C_{lip,2}}{\sqrt{\hat{N}}}\sqrt{2(1 + \|u_{\text{true}}(0)\|^2_{H,L^2_\rho} + \frac{1}{C_{\mathcal{L}}}\|f\|^2_{L^2(0,t^n;L^2_\rho(\Omega;H))}}$$
$$+ C_{lip,1}\left(c_0 \, \|u^R_{\text{true}}(0) - u_{\text{true}}(0)\|^2_{H,L^2_\rho} + c_1\varepsilon^2 + \tilde{c}_2\triangle t^2 + \tilde{c}_3 h^{2r}\right)^{1/2},$$

where $\|u^R_{\text{true}}(0) - u_{\text{true}}(0)\|^2_{H,L^2_\rho}$ *is the error of the Karhunen-Loève expansion truncation on the initial datum. The constants* $c_0, c_1, \tilde{c}_2, \tilde{c}_3$ *do not depend on any of the discretization parameters* $\triangle t, h, \{\omega_i\}_{i=1}^{\hat{N}}$.

Proof. The mean-square error of $m^n_{h,\hat{\rho}}$ can be split into

$$\sqrt{\mathbb{E}_{\otimes\rho}\left[\left|m^n - m^n_{h,\hat{\rho}}\right|^2\right]} \leq \sqrt{\mathbb{E}_{\otimes\rho}\left[\left|\mathbb{E}_\rho[\Phi(u_{\text{true}}(t^n))] - \mathbb{E}_{\hat{\rho}}[\Phi(u_{\text{true}}(t^n))]\right|^2\right]}$$
$$+ \sqrt{\mathbb{E}_{\otimes\rho}\left[\left|\mathbb{E}_{\hat{\rho}}[\Phi(u_{\text{true}}(t^n))] - \mathbb{E}_{\hat{\rho}}[\Phi(u^n_{h,\hat{\rho}})]\right|^2\right]}$$

The first term $\sqrt{\mathbb{E}_{\otimes\rho}\left[\left|\mathbb{E}_\rho[\Phi(u_{\text{true}}(t^n))] - \mathbb{E}_{\hat{\rho}}[\Phi(u_{\text{true}}(t^n))]\right|^2\right]}$ expresses the standard error caused by approximating the mean value m^n with the sample mean. It can be bounded by

$$\begin{aligned}
\sqrt{\mathbb{E}_{\otimes\rho}\left[\left|\mathbb{E}_\rho[\Phi(u_{\text{true}}(t^n))] - \mathbb{E}_{\hat{\rho}}[\Phi(u_{\text{true}}(t^n))]\right|^2\right]} &= \sqrt{\frac{\mathbb{E}_\rho[\Phi(u_{\text{true}}(t^n))^2] - \mathbb{E}_\rho[\Phi(u_{\text{true}}(t^n))]^2}{\hat{N}}} \\
&\leq \sqrt{\frac{\mathbb{E}_\rho[\Phi(u_{\text{true}}(t^n))^2]}{\hat{N}}} \\
&\leq \sqrt{\frac{\mathbb{E}_\rho[C^2_{lip,2}(\|u_{\text{true}}(t^n)\|_H + 1)^2]}{\hat{N}}} \\
&\leq \sqrt{\frac{2C^2_{lip,2} + 2C^2_{lip,2}\mathbb{E}_\rho[\|u_{\text{true}}(t^n)\|^2_H]}{\hat{N}}} \\
&\leq \sqrt{\frac{2C^2_{lip,2} + 2C^2_{lip,2}\mathbb{E}_\rho[\|u_{\text{true}}(0)\|^2_H + \frac{1}{C_{\mathcal{L}}}\|f\|^2_{L^2(0,t^n;H)}]}{\hat{N}}} \\
&\leq \frac{C_{lip,2}}{\sqrt{\hat{N}}}\sqrt{2(1 + \|u_{\text{true}}(0)\|^2_{H,L^2_\rho} + \frac{1}{C_{\mathcal{L}}}\|f\|^2_{L^2(0,t^n;L^2_\rho(\Omega;H))}},
\end{aligned}$$

where in the fifth step we applied a stability result

$$\|u_{\text{true}}(t^n)\|^2_H \leq \|u_{\text{true}}(0)\|^2_H + \frac{1}{C_{\mathcal{L}}}\|f\|^2_{L^2(0,t^n;H)} \qquad \forall\omega \in \Omega,$$

which holds for the true solution. The second term can be bounded as

$$
\begin{aligned}
&\sqrt{\mathbb{E}_{\otimes\rho}\Big[\Big|\mathbb{E}_{\hat\rho}[\Phi(u_{\text{true}}(t^n))] - \mathbb{E}_{\hat\rho}[\Phi(u^n_{h,\hat\rho})]\Big|^2\Big]} \\
&\quad= \sqrt{\mathbb{E}_{\otimes\rho}\Big[\Big(\sum_{j=1}^{\hat N}\frac{1}{\hat N}\Big(\Phi(u_{\text{true}}(t^n,\omega_j)) - \Phi(u^n_{h,\hat\rho}(\omega_j))\Big)\Big)^2\Big]} \\
&\quad\le \sqrt{\mathbb{E}_{\otimes\rho}\Big[\sum_{j=1}^{\hat N}\frac{1}{\hat N}|\Phi(u_{\text{true}}(t^n,\omega_j)) - \Phi(u^n_{h,\hat\rho}(\omega_j))|^2\Big]} \\
&\quad\le \sqrt{\mathbb{E}_{\otimes\rho}\Big[\sum_{j=1}^{\hat N}\frac{1}{\hat N}C^2_{lip,1}\|u_{\text{true}}(t^n,\omega_j) - u^n_{h,\hat\rho}(\omega_j)\|^2_H\Big]} \\
&\quad\le \sqrt{C^2_{lip,1}\mathbb{E}_{\otimes\rho}\Big[\|u_{\text{true}}(t^n) - u^n_{h,\hat\rho}\|^2_{H,L^2_{\hat\rho}}\Big]} \\
&\quad\le \sqrt{C^2_{lip,1}\mathbb{E}_{\otimes\rho}\Big[c_0\|u^R_{\text{true}}(0) - u_{\text{true}}(0)\|^2_{H,L^2_{\hat\rho}} + c_1\varepsilon^2 + c_2\triangle t^2 + c_3h^{2r}\Big]} \\
&\quad\le C_{lip,1}\Big(c_0\,\mathbb{E}_{\otimes\rho}[\|u^R_{\text{true}}(0) - u_{\text{true}}(0)\|^2_{H,L^2_{\hat\rho}}] + c_1\varepsilon^2 + \mathbb{E}_{\otimes\rho}[c_2]\triangle t^2 + \mathbb{E}_{\otimes\rho}[c_3]h^{2r}\Big)^{1/2} \\
&\quad= C_{lip,1}\Big(c_0\,\mathbb{E}_{\otimes\rho}[\|u^R_{\text{true}}(0) - u_{\text{true}}(0)\|^2_{H,L^2_{\hat\rho}}] + c_1\varepsilon^2 + \tilde c_2\triangle t^2 + \tilde c_3h^{2r}\Big)^{1/2},
\end{aligned}
$$

where in the forth step we applied Theorem 4.2.1 and in the fifth step we applied the observation that c_0 and c_1 do not depend on the choice of the sampling points. In the last step we defined new constants $\tilde c_2, \tilde c_3$, which can be expressed as

$$
\begin{aligned}
\tilde c_2 = \mathbb{E}_{\otimes\rho}[c_2] &= C_1\mathbb{E}_{\otimes\rho}\Big[\|\frac{\partial^2 u_{\text{true}}}{\partial t^2}\|^2_{L^2(0,T;L^2_{\hat\rho}(\Omega;H))}\Big] + C_2\mathbb{E}_{\otimes\rho}\Big[\|u_{\text{true}}(0)\|^2_{H,L^2_{\hat\rho}}\Big] \\
&\qquad + C_3\mathbb{E}_{\otimes\rho}\Big[\|f\|^2_{L^\infty(0,T;L^2_{\hat\rho}(\Omega;H))}\Big] + C_4\mathbb{E}_{\otimes\rho}\Big[\|\dot u_{\text{true}}\|^2_{L^2(0,T;L^2_{\hat\rho}(\Omega;V))}\Big] \\
&= C_1\|\frac{\partial^2 u_{\text{true}}}{\partial t^2}\|^2_{L^2(0,T;L^2_\rho(\Omega;H))} + C_2\|u_{\text{true}}(0)\|^2_{H,L^2_\rho} + C_3\|f\|^2_{L^\infty(0,T;L^2_{\hat\rho}(\Omega;H))} \\
&\qquad + C_4\|\dot u_{\text{true}}\|^2_{L^2(0,T;L^2_\rho(\Omega;V))} \\
\tilde c_3 = \mathbb{E}_{\otimes\rho}[c_3] &= C_5\mathbb{E}_{\otimes\rho}\Big[\|\dot u_{\text{true}}\|^2_{L^2(0,T;L^2_{\hat\rho}(\Omega;H^r))}\Big] + C_6\mathbb{E}_{\otimes\rho}\Big[\|u_{\text{true}}(0)\|_{H^r,L^2_{\hat\rho}}\Big] \\
&= C_5\|\dot u_{\text{true}}\|^2_{L^2(0,T;L^2_\rho(\Omega;H^r))} + C_6\|u_{\text{true}}(0)\|_{H^r,L^2_\rho}.
\end{aligned}
$$

As for the initial error, we proceed analogously

$$
\mathbb{E}_{\otimes\rho}[\|u^R_{\text{true}}(0) - u_{\text{true}}(0)\|^2_{H,L^2_{\hat\rho}}] = \|u^R_{\text{true}}(0) - u_{\text{true}}(0)\|^2_{H,L^2_\rho}.
$$

□

5 A-posteriori error estimation

The goal of this work is to derive a residual based a-posteriori error estimation for a DLR approximation of a random parabolic equation, with a special focus on a random heat equation with diffusion coefficient affine w.r.t. the random variables. The problem is discretized by the finite element method (FEM) in physical space and a stochastic collocation (SC) method in the random variables. Before tackling this problem, we direct our attention to an a-posteriori error estimation of a random heat equation, without any DLR error contribution. The results of this work are available in Section 5.1, which follows very closely the publication [NV19]. Section 5.2 is then dedicated to deriving an error estimation including a DLRA error contribution.

5.1 A posteriori error estimation for a random heat equation

In this section, we present a residual based a posteriori error estimation for a random heat equation. The problem is discretized by a stochastic collocation finite element method and advanced in time by the θ-scheme. Concerning a reliable estimation of the discretization error, the work [GN18] derives a residual based a posteriori error estimation for an elliptic problem discretized by the stochastic collocation finite element method. There, the authors propose an algorithm that adaptively builds the sparse grid based on the a posteriori estimation of the SC error.

This work extends the results obtained in [GN18] to a heat equation with random right hand side and random diffusion coefficient that depends affinely on a finite number of random variables. The provided estimates bound the norm of the error in L^2 in stochastic space, L^2 in time and H^1 in physical space. The estimator naturally splits into a spatial discretization estimator, time discretization estimator and stochastic discretization estimator, which are then used to drive the adaptivity with respect to all three types of discretizations. We then propose an adaptive algorithm to build a suitable time discretization, as well as an FE mesh and a sparse grid common to all time steps, so as to achieve a prescribed tolerance on a global norm of the error.

We start with introducing the problem, namely a heat equation with a random diffusion coefficient and right hand side. We follow with defining the spatial, time and stochastic discretization. We derive two residual based a posteriori error estimations. The first one concerns the general case of sparse grids and spatial meshes that change in time. The second one is simpler and concerns the case of spatial mesh and sparse grid kept fixed over the time iterations. After, we propose an adaptive algorithm to build nonuniform time discretizations, as well as nonuniform meshes and anisotropic sparse grids that are fixed in time for the case of a deterministic right hand side. In the last part, we first study on a specific example the behaviour and sharpness of all three components of the estimator (spatial, temporal and stochastic) and then apply our adaptive algorithm and assess its performance. The results presented in this section are summed up in the paper [NV19].

5.1.1 Problem statement

Let $D \subset \mathbb{R}^d$ be an open polygonal domain with Lipschitz boundary ∂D and $(\Omega, \mathcal{F}, P)$ be a complete probability space. Given a final time T, random forcing term $f : D \times \Omega \times (0,T) \to \mathbb{R}$, initial condition $u_0 : D \times \Omega \to \mathbb{R}$ and a diffusion coefficient $a : D \times \Omega \to \mathbb{R}$, the problem states: find a solution $u : D \times \Omega \times (0,T] \to \mathbb{R}$ satisfying $P-$almost everywhere in Ω

$$\begin{aligned} \frac{\partial u}{\partial t} - \nabla \cdot (a \nabla u) = f & \qquad \text{in } D \times \Omega \times (0,T], \\ u = 0 & \qquad \text{on } \partial D \times \Omega \times (0,T], \\ u(\cdot,\cdot,0) = u_0 & \qquad \text{in } D \times \Omega. \end{aligned} \tag{5.1}$$

Suppose that $f \in L^2(0,T; L^2(\Omega, H^{-1}(D)))$, $u_0 \in L^2(\Omega, H_0^1(D))$ and a is a random variable on $(\Omega, \mathcal{F}, P)$ taking values in $W^{1,\infty}(D)$ (random field) satisfying

$$\exists\, a_{min}, a_{max} : P(\omega \in \Omega : 0 < a_{min} \leq a(x,\omega) \leq a_{max} < \infty \quad \forall x \in D) = 1. \tag{5.2}$$

In addition we require that the diffusion coefficient as well as the forcing term and the initial condition can be parametrized by a finite number of independent, real-valued random variables $\{Y_m\}_{m=1}^M$ defined on Ω, i.e. $f(x,\omega,t) = f(x, Y_1(\omega), \ldots, Y_M(\omega), t)$, $u_0(x,\omega) = u_0(x, Y_1(\omega), \ldots, Y_M(\omega))$ and the dependence of a on $\{Y_m\}_{m=1}^M$ is affine, i.e.

$$a(x,\omega) = a_0(x) + \sum_{m=1}^{M} a_m(x) Y_m(\omega). \tag{5.3}$$

The solution u then depends on the same random variables as well, i.e. $u(x,\omega,t) = u(x,Y_1(\omega),\dots,Y_M(\omega),t)$, and we can recast the probability space $(\Omega,\mathcal{F},P)$ into $(\Gamma,B(\Gamma),\rho(y)\mathrm{d}y)$ by introducing $\Gamma=\Gamma_1\times\cdots\times\Gamma_M$ with $\Gamma_m=Y_m(\Omega)$ for $m=1,\dots,M$. The expression $B(\Gamma)$ denotes the Borel $\sigma-$algebra defined over Γ. The joint probability density function of the random vector $Y=(Y_1,\dots,Y_M)$ is denoted by $\rho:\Gamma\to\mathbb{R}_+$ and factorizes as $\rho(y)=\Pi_{m=1}^M\rho_m(y_m)$ for all $y=(y_1,\dots,y_M)\in\Gamma$.

In what follows we consider the following two Bochner spaces: for a given Banach space $(V,\|\cdot\|_V)$ and for any $t_1,t_2\in[0,T]$, $t_1<t_2$ we define

$$L^2(t_1,t_2;V)=\{v:(t_1,t_2)\to V|v \text{ is strongly measurable and } \|v\|_{L^2(t_1,t_2;V)}<\infty\}$$

where $\|v\|^2_{L^2(t_1,t_2;V)}=\int_{t_1}^{t_2}\|v(t)\|^2_V\,\mathrm{d}t$ and

$$L^2_\rho(\Gamma;L^2(t_1,t_2;V))=\{v:\Gamma\to L^2(t_1,t_2;V)|\, v \text{ is strongly measurable and } \|v\|_{L^2_\rho(\Gamma;L^2(t_1,t_2;V))}<\infty\}$$

with $\|v\|^2_{L^2_\rho(\Gamma,L^2(t_1,t_2;V))}=\int_\Gamma\|v(y)\|^2_{L^2(t_1,t_2;V)}\rho(y)\mathrm{d}y$. It holds

$$L^2_\rho(\Gamma;L^2(t_1,t_2;V))\cong L^2(t_1,t_2;L^2_\rho(\Gamma;V)),$$

i.e. this Bochner space is isometrically isomorphic to the Bochner space $L^2(t_1,t_2;L^2_\rho(\Gamma;V))$ [Nee08, p. 12].

The (pointwise in Γ) weak formulation of problem (5.1) then reads: Find $u\in W$ where

$$W=\left\{w\in L^2_\rho\Big(\Gamma;L^2\Big(0,T;H^1_0(D)\Big)\Big) \text{ and } \frac{\partial w}{\partial t}\in L^2_\rho\Big(\Gamma;L^2\Big(0,T,H^{-1}(D)\Big)\Big)\right\}$$

s.t.

$$\int_D\frac{\partial u(x,y,t)}{\partial t}v(x)\,\mathrm{d}x+\int_D a(x,y)\nabla u(x,y,t)\cdot\nabla v(x)\,\mathrm{d}x=\int_D f(x,y,t)v(x)\,\mathrm{d}x$$
$$\forall v\in H^1_0(D),\ \rho-\text{a.e. } y\in\Gamma,\ \text{and a.e. } t\in(0,T] \qquad (5.4)$$

with initial and boundary conditions:

$$\begin{aligned} u(x,y,0)&=u_0(x,y) && \rho\text{-a.e. } y\in\Gamma\\ u(x,y,t)&=0 && x\in\partial D,\ \rho\text{-a.e. } y\in\Gamma,\text{a.e. } t\in(0,T).\end{aligned}$$

We endow the Sobolev space $H^1_0(D)$ with the gradient norm $\|v\|_{H^1_0}=\|\nabla v\|_{L^2(D)}$. Based on the existence result of the deterministic problem [Dau+99, p.513], the assumption (5.2) ensures the well-posedness of problem (5.4), i.e. there exists a unique solution $u \in W$ which moreover satisfies

$$\|u\|_{L^2_\rho(\Gamma;L^2(0,T;H^1_0(D)))} \leq \frac{C}{\sqrt{a_{min}}}\left[\|u_0\|^2_{L^2_\rho(\Gamma;H^1_0(D))} + \frac{1}{a_{min}}\|f\|^2_{L^2_\rho(\Gamma;L^2(0,T;L^2(D)))}\right]^{1/2}.$$

5.1.2 Discretization aspects

In the following sub-sections we describe the techniques used for the discretization of problem (5.4) and corresponding assumptions necessary for a rigorous a posteriori estimation. We will closely follow the techniques used in [Ver13; Ver03] for the time and space discretization and [GN18; NTW08a; NTW08b; BNT10] for the stochastic discretization by the stochastic collocation method.

Time discretization

For the time discretization we divide the time interval into N subintervals $0 = t_0 < t_1 < \cdots < t_N = T$. By τ we will denote the discretization $\tau = \{t_n\}_{n=1}^N$ and τ_{n+1} will denote the length of the $(n+1)$-th interval $\tau_{n+1} = t_{n+1} - t_n$. We will also assume that f is continuous w.r.t. time. We will use the abbreviations

$$g^n(x,y) = g(x,y,t_n), \quad g^{n\theta} = (1-\theta)g^n + \theta g^{n+1}.$$

The numerical scheme considered here for the time discretization is the $\theta-$scheme with $\theta \in [0,1]$.

Space discretization

The spatial discretization will be performed by the finite element method. To each time instant t_n, $0 \leq n \leq N$, we associate a triangulation $\mathcal{T}_{h_n}$ of D which satisfies $\bigcup_{K\in\mathcal{T}_{h_n}} K = D$ and a corresponding conforming finite element space V_{h_n}. For a rigorous estimation we require the following conditions to be satisfied, which are taken from [Ver03].

1. *Affine equivalence:* there is an invertible affine mapping for every element $K \in \mathcal{T}_{h_n}$ onto the standard reference d-simplex or the standard unit cube in $\mathbb{R}^d$.
2. *Admissibility:* any two elements either share a vertex or a complete edge ($d = 2$) or a complete face ($d = 3$) or are disjoint.
3. *Shape regularity:* the ratio of the diameter of any element to the diameter of its

largest inscribed ball is bounded uniformly with respect to all partitions $\mathcal{T}_{h_n}$ and to N.

4. *Transition condition:* for every $n = 1, \dots, N$ there is a refinement of both $\mathcal{T}_{h_n}$ and $\mathcal{T}_{h_{n-1}}$, denoted by $\tilde{\mathcal{T}}_{h_n}$, which is an affinely equivalent, admissable and shape-regular triangulation and such that

$$\sup_{1 \le n \le N} \sup_{K \in \tilde{\mathcal{T}}_{h_n}} \sup_{K' \in \mathcal{T}_{h_n}; K' \supset K} \frac{h_{K'}}{h_K} < \infty. \tag{5.5}$$

5. For every $n = 1, \dots, N$, V_{h_n} consists of continuous functions which are piecewise polynomials of degree $\le p_n$, $p_n \ge 1$ where p_n is uniformly bounded with respect to N.

Stochastic discretization

The stochastic discretization is performed by a sparse grid collocation method, first introduced in [Smo63]. We will briefly recall this method and refer the reader to [NTW08a; NTW08b; BNT10] for more details.

Let us define a sequence of univariate polynomial interpolant operators

$$\mathcal{U}_j^{m(i_j)} : C^0(\Gamma_j) \to \mathbb{P}_{m(i_j)-1}(\Gamma_j), \quad j = 1, \dots, M,$$

where $m(i_j)$, called a level function, denotes the number of collocation points for level i_j and $\mathbb{P}_q(\Gamma_j)$ is the space of polynomials over Γ_j with degree at most q. The function m is a strictly increasing function satisfying $m(0) = 0$, $m(1) = 1$. For a multi-index $q = (q_1, \dots, q_M) \in \mathbb{N}^M$, we denote by $\mathbb{P}_q(\Gamma)$ the tensor product polynomial space $\mathbb{P}_q(\Gamma) = \bigotimes_{j=1}^M \mathbb{P}_{q_j}(\Gamma_j)$.

A sparse grid is built over a multi-index set $I \subset \mathbb{N}_+^M$ with the only assumption being that I is downward-closed (called also admissibility condition), i.e.

$$\forall i \in I,\ i - e_j \in I \quad \forall j \in \{1, 2, \dots, M\} \quad \text{s.t.} \quad i_j > 1,$$

where e_j is the $j-$th canonical unit vector.

By setting $\mathcal{U}_j^0 = 0$ for $j = \{1, \dots, M\}$ we can define the sparse grid interpolant S_I :

$L^2_\rho(\Gamma) \cap C^0(\Gamma) \to \mathbb{P}_I := \bigoplus_{i \in I} \mathbb{P}_{m(i)-1}(\Gamma)$ of a continuous function $f : \Gamma \to \mathbb{R}$ by

$$S_I[f](y) = \sum_{i \in I} \triangle^{m(i)}[f](y), \tag{5.6}$$

where

$$\triangle^{m(i)} = \bigotimes_{j=1}^{M} \Big(\mathcal{U}_j^{m(i_j)} - \mathcal{U}_j^{m(i_j - 1)} \Big).$$

The operator S_I can be equivalently expressed as a linear combination of tensor grid interpolations (see [NTW08b])

$$S_I[f](y) = \sum_{i \in I} c_i \bigotimes_{j=1}^{M} \mathcal{U}_j^{m(i_j)}(f)(y), \qquad c_i = \sum_{\substack{k \in \{0,1\}^M \\ (i+k) \in I}} (-1)^{|k|} \tag{5.7}$$

with $|k| = \sum_{j=1}^M k_j$. We then call a sparse grid the collection of $N_c(I)$ points $\mathcal{X}(I) = \{y_1, \ldots, y_{N_c(I)}\}$ that are used in (5.7) to build the interpolant $S_I[f]$. The collocation points are called nested if we have $\mathcal{X}(I) \subset \mathcal{X}(J)$ whenever $I \subset J$. Since $S_I[f]$ is linear in the point evaluations $\{f(y_k),\ y_k \in \mathcal{X}(I)\}$, it can be written in the form

$$S_I[f](y) = \sum_{k=1}^{N_c(I)} f(y_k) L_k(y) \tag{5.8}$$

for suitable functions L_k. Finally, we introduce the notion of margin M_I of the index set I defined by

$$M_I = \{ i \in \mathbb{N}_+^M \setminus I : \quad i - e_j \in I \text{ for some } j \in \{1, \ldots, M\} \}. \tag{5.9}$$

Equation (5.4) will be collocated on the grid $\mathcal{X}(I_n) = \{y_1, \ldots, y_{N_c(I_n)}\}$ defined by an index set I_n that is allowed to change between the time steps. In particular, we allow for both refinement and coarsening of the index set. The collocation points are assumed to be nested. This condition implies, in particular, that S_{I_n} is interpolatory, i.e.

$$S_{I_n}[f](y_k) = f(y_k), \quad k = 1, \ldots, N_c(I_n), \quad n = 0, \ldots, N, \tag{5.10}$$

see [BNR00, p. 277]. By $\tilde{I}_{n+1}$ we will denote the index set

$$\tilde{I}_{n+1} = I_n \cup I_{n+1}. \tag{5.11}$$

The following proposition will be useful for the derivation of the error estimates.

Proposition 5.1.1. Let S_I be an interpolatory sparse grid interpolant, as defined in (5.6). Then

1. $\forall f, g \in C^0(\Gamma)$: $S_I[f\, g] = S_I[f\, S_I[g]]$,
2. $\forall f \in C^0(\Gamma)$: $S_I[f] \in \mathbb{P}_I$,
3. $\forall p \in \mathbb{P}_I(\Gamma)$: $S_I[p] = p$.

A proof can be found in [GN18, p.3126] for part 1 and in [Bäc+11, p.52] for part 2 and part 3.

If S_I is interpolatory, then the functions L_k in (5.8) are Lagrangian, i.e. $L_k(y_j) = \delta_{jk}$ and form a basis of $\mathbb{P}_I$.

5.1.3 Fully discrete problem

We allow the spatial and the stochastic grid to change over time and we define the discrete solution for each $n = 0, \dots, N$ as a function belonging to $V_{h_n} \otimes \mathbb{P}_{I_n}$:

$$u^n_{h_n,I_n} = \sum_{k=1}^{N_c(I_n)} u^n_{h_n,I_n,k}\, L_k(y),$$

where $u^n_{h_n,I_n,k} = u^n_{h_n,I_n}(y_k) \in V_{h_n}$ and $u^{n+1}_{h_{n+1},I_{n+1}}$ satisfies for all $v_{h_{n+1}} \in V_{h_{n+1}}$ and for all $k = 1, \dots, N_c(I_{n+1})$ the equation

$$\begin{aligned}\int_D \frac{u^{n+1}_{h_{n+1},I_{n+1},k}(x) - u^n_{h_n,I_n}(x, y_k)}{\tau_{n+1}} v_{h_{n+1}}(x)\mathrm{d}x \\ + \int_D a(x, y_k)\Big(\theta \nabla u^{n+1}_{h_{n+1},I_{n+1},k}(x) + (1-\theta)\nabla u^n_{h_n,I_n}(x, y_k)\Big)\nabla v_{h_{n+1}}(x)\mathrm{d}x \\ = \int_D f^{n\theta}(x, y_k)\, v_{h_{n+1}}(x)\mathrm{d}x\end{aligned} \tag{5.12}$$

with initial condition

$$u^0_{h_0,I_0}(x, y) = \sum_{k=1}^{N_c(I_0)} \Pi_{h_0} u_0(x, y_k) L_k(y) \tag{5.13}$$

where Π_{h_0} is a Lagrange interpolation operator into V_{h_0}. The Lax Milgram lemma implies the existence of a unique sequence of solutions $\{u^n_{h_n,I_n}\}^N_{n=0}$. Based on this sequence we build a piecewise affine function $\tilde{u}$ on $[0, T]$ which equals $u^n_{h_n,I_n}$ at times t_n, $n = 0, \dots, N$, i.e.

$$\tilde{u}(t) = \frac{t_{n+1} - t}{\tau_{n+1}} u^n_{h_n,I_n} + \frac{t - t_n}{\tau_{n+1}} u^{n+1}_{h_{n+1},I_{n+1}}, \qquad t \in [t_n, t_{n+1}]. \tag{5.14}$$

Note that
$$\frac{\partial \tilde{u}}{\partial t} = \frac{1}{\tau_{n+1}}\Big(u^{n+1}_{h_{n+1},I_{n+1}} - u^{n}_{h_n,I_n}\Big) \qquad \text{on } (t_n, t_{n+1}].$$
With this construction, for every $n = 0, \ldots, N-1$, the discretized solution belongs to the space

$$\tilde{u} \in L^2(t_n, t_{n+1}; \mathbb{P}_{\tilde{I}_{n+1}} \otimes \tilde{V}_{h_{n+1}}) \subset L^2(t_n, t_{n+1}; L^2_\rho(\Gamma; H^1_0(D)))$$

where $\tilde{V}_{h_{n+1}}$ is the FE space corresponding to the refined triangulation $\tilde{\mathcal{T}}_{h_{n+1}}$, see (5.5), and $\tilde{I}_{n+1}$ is the union of the index sets defined in (5.11).

5.1.4 Residual based a posteriori error estimation

In this section we will derive an a posteriori error estimate for $u - \tilde{u}$ which consists of three error contributors: space, time and stochastic. First we shall start by stating the equation satisfied by $\tilde{u}$.

From (5.12) it is easy see that the discretized solution $\tilde{u}$ satisfies the following equation in $(t_n, t_{n+1}]$ and for each $n = 0, \ldots, N-1$

$$\begin{aligned} \int_D S_{I_{n+1}}\left[\frac{\partial \tilde{u}}{\partial t}\right] v_{h_{n+1}} + \int_D S_{I_{n+1}}\Big[a\nabla\tilde{u}\Big]\nabla v_{h_{n+1}} &= \int_D S_{I_{n+1}}\Big[f\Big] v_{h_{n+1}} \\ + \int_D S_{I_{n+1}}\Big[a\nabla\tilde{u} - a\nabla\tilde{u}^{n\theta}\Big]\nabla v_{h_{n+1}} &+ \int_D S_{I_{n+1}}\Big[f^{n\theta} - f\Big] v_{h_{n+1}} \qquad (5.15) \\ &\forall v_{h_{n+1}} \in V_{h_{n+1}}, \text{ everywhere in } \Gamma. \end{aligned}$$

For any element, face or edge S, h_S denotes its diameter. With every edge ($d = 2$) or face ($d = 3$) E, we identify a unit vector η_E orthogonal to it and denote the jump across E in direction η_E by $[\cdot]_E$. The assumption (5.2) ensures that the energy norm and the H^1_0 norm are equivalent for every $y \in \Gamma$, i.e. there exists $0 < c_{min} \leq c_{max}$ s.t.

$$c_{min}\|\nabla v\|_{L^2(D)} \leq \|a^{1/2}(y)\nabla v\|_{L^2(D)} \leq c_{max}\|\nabla v\|_{L^2(D)}, \quad \rho - \text{a.e. in } \Gamma$$

for any $v \in H^1_0(D)$. The constants c_{min}, c_{max} can be bounded by $c_{min} \geq \frac{1}{\sqrt{a_{min}}}$ and $c_{max} \leq \sqrt{a_{max}}$.

Now we can proceed to state the a posteriori error estimate.

Theorem 5.1.2. *Let u be the solution of (5.4) and $\tilde{u}$ be defined as in (5.14). Then there exists a constant $C > 0$ independent of the time step, mesh size, the sparse grid index set such that*

$$\begin{aligned} &\|(u-\tilde{u})(T)\|^2_{L^2_\rho(\Gamma;L^2(D))} + c^2_{min}\,\|u-\tilde{u}\|^2_{L^2(0,T;L^2_\rho(\Gamma;H^1_0(D)))} \\ &\qquad\qquad \leq \|(u-\tilde{u})(0)\|^2_{L^2_\rho(\Gamma;L^2(D))} + \epsilon^2_{spa} + \epsilon^2_{tem} + \epsilon^2_{sto}, \end{aligned}$$

where

$$\begin{aligned}\epsilon_{spa}^2 = \frac{C}{c_{min}^2} \sum_{n=0}^{N-1} \Lambda_{I_{n+1}} & \sum_{k=1}^{N_c(I_{n+1})} \\ \Bigg(\sum_{K \in \tilde{\mathcal{T}}_{h_{n+1}}} & h_K^2 \left\| f(y_k) - \frac{\partial \tilde{u}}{\partial t}(y_k) + \nabla \cdot \Big(a(y_k) \nabla \tilde{u}(y_k) \Big) \right\|^2_{L^2(t_n,t_{n+1};L^2(K))} \\ + \sum_{E \subset \partial K} & h_E \left\| \frac{1}{2} [a(y_k) \nabla \tilde{u}(y_k) \cdot \eta_E]_E \right\|^2_{L^2(t_n,t_{n+1};L^2(E))} \Bigg) \|L_k\|_{L^1_\rho(\Gamma)}\end{aligned} \tag{5.16}$$

and

$$\begin{aligned}\epsilon_{tem}^2 = \frac{C}{c_{min}^2} \sum_{n=0}^{N-1} \Lambda_{I_{n+1}} \sum_{k=1}^{N_c(I_{n+1})} 2 \Big(& \|f(y_k) - f^{n\theta}(y_k)\|^2_{L^2(t_n,t_{n+1};L^2(D))} \\ + \tau_{n+1} \frac{\theta^3 + (1-\theta)^3}{3} & \|a(y_k) \nabla (u^{n+1}_{h_{n+1},I_{n+1}} - u^n_{h_n,I_n})\|^2_{L^2(D)} \Big) \|L_k\|_{L^1_\rho(\Gamma)}\end{aligned} \tag{5.17}$$

and

$$\begin{aligned}\epsilon_{sto}^2 = \frac{C}{c_{min}^2} \sum_{n=0}^{N-1} \tau_{n+1} \Bigg(& \sum_{i \in I_{n+1}^C \cap (I_n \cup M_{I_n})} \left\| \triangle^{m(i)} (a \nabla u^n_{h_n,I_n}) \right\|^2_{L^2_\rho(\Gamma;L^2(D))} \\ + \sum_{i \in M_{I_{n+1}}} & \left\| \triangle^{m(i)} (a \nabla u^{n+1}_{h_{n+1},I_{n+1}}) \right\|^2_{L^2_\rho(\Gamma;L^2(D))} \Bigg) \\ + \sum_{i \in I_{n+1}^C} & \left\| \triangle^{m(i)} (f) \right\|^2_{L^2(t_n,t_{n+1};L^2_\rho(\Gamma,L^2(D)))} \\ + \frac{1}{\tau_{n+1}} & \sum_{i \in \tilde{I}_{n+1} \setminus I_{n+1}} \left\| \triangle^{m(i)} (u^n_{h_n,I_n}) \right\|^2_{L^2_\rho(\Gamma;L^2(D))} .\end{aligned} \tag{5.18}$$

where $\Lambda_{I_{n+1}}$ *denotes the Lebesgue constant corresponding to the index set* I_{n+1}.

Proof. In what follows all equations hold a.e. in (t_n, t_{n+1}), $n = 0, \dots, N-1$ and ρ-a.e. in Γ and we will omit the dependence on the variables x, y, t. We will start by dividing the estimate into a stochastic and a deterministic part. For every $v \in H_0^1(D)$ we have

$$\begin{aligned}\int_D \Big(\frac{\partial u}{\partial t} - \frac{\partial \tilde{u}}{\partial t} \Big) v + \int_D a \nabla \Big(u - \tilde{u} \Big) \nabla v = \int_D f v - \int_D \frac{\partial \tilde{u}}{\partial t} v - \int_D a \nabla \tilde{u} \nabla v = \\ = \underbrace{S_{I_{n+1}} \Big[\int_D f v - \int_D \frac{\partial \tilde{u}}{\partial t} v - \int_D a \nabla \tilde{u} \nabla v \Big] \Big\}}_{=:A_{det}} \\ + \underbrace{S_{I_{n+1}} \Big[\int_D a \nabla \tilde{u} \nabla v + \int_D \frac{\partial \tilde{u}}{\partial t} - \int_D f v \Big] - \Big(\int_D a \nabla \tilde{u} \nabla v + \int_D \frac{\partial \tilde{u}}{\partial t} - \int_D f v \Big)}_{=:A_{sto}} .\end{aligned}$$

We analyze A_{det} and A_{sto} separately. The term A_{det} accounts for both spatial and temporal error contribution and we can use standard techniques for a posteriori error estimation of deterministic heat equations, see [Ver13; Ver03]. For any $v_{h_{n+1}} \in V_{h_{n+1}}$ we have

$$\begin{aligned}
A_{det} = \; & S_{I_{n+1}}\Big[\int_D f(v - v_{h_{n+1}}) - \int_D \frac{\partial \tilde{u}}{\partial t}(v - v_{h_{n+1}}) - \int_D a\nabla\tilde{u}\nabla(v - v_{h_{n+1}})\Big] \\
& + S_{I_{n+1}}\Big[\int_D f v_{h_{n+1}} - \int_D \frac{\partial \tilde{u}}{\partial t} v_{h_{n+1}} - \int_D a\nabla\tilde{u}\nabla v_{h_{n+1}}\Big] \\
= \; & \underbrace{S_{I_{n+1}}\Big[\int_D f(v - v_{h_{n+1}}) - \int_D \frac{\partial \tilde{u}}{\partial t}(v - v_{h_{n+1}}) - \int_D a\nabla\tilde{u}\nabla(v - v_{h_{n+1}})\Big]}_{=:A_{spa}} \\
& + \underbrace{S_{I_{n+1}}\Big[\int_D \Big(a\nabla\tilde{u}^{n\theta} - a\nabla\tilde{u}\Big)\nabla v_{h_{n+1}}\Big] + S_{I_{n+1}}\Big[\int_D \Big(f - f^{n\theta}\Big)v_{h_{n+1}}\Big]}_{=:A_{tem}}
\end{aligned} \tag{5.19}$$

where in the second equality we employed the equation (5.15) for $\tilde{u}$. Now we have divided A_{det} into a spatial A_{spa} and a temporal A_{tem} error contributor.

For the spatial part we will follow the estimation provided in [Ver03]. We denote by J_{h_n} any of the quasi interpolation operators of [Ver99a] defined on $H_0^1(D)$ and with values in the space of continuous, piecewise linear finite element functions corresponding to $\mathcal{T}_{h_n}$. Then, combining the interpolation error estimates of [Ver99a], a standard trace theorem [Ver99a, Lemma 3.2] and the condition 4. stated in (5.5), the following estimates hold for every $v \in H_0^1$ and for any element $K \in \tilde{\mathcal{T}}_{h_n}$ and interior edge/face $E \in \tilde{\mathcal{E}}_{h_n}$

$$\begin{aligned}
& \|\nabla(v - J_{h_n}v)\|_{L^2(K)} \le \|\nabla(v - J_{h_n}v)\|_{L^2(K')} \le c_0 \|\nabla v\|_{L^2(\tilde{\omega}_K)}, \\
& \|v - J_{h_n}v\|_{L^2(K)} \le \|v - J_{h_n}v\|_{L^2(K')} \le c_1 h_{K'} \|\nabla v\|_{L^2(\tilde{\omega}_K)} \le \tilde{c}_1 h_K \|\nabla v\|_{L^2(\tilde{\omega}_K)}, \\
& \|v - J_{h_n}v\|_{L^2(E)} \le c_2\Big\{h_E^{-1/2}\|v - J_{h_n}v\|_{L^2(K)} + h_E^{1/2}\|\nabla(v - J_{h_n}v)\|_{L^2(K)}\Big\} \\
& \qquad\qquad\qquad \le \tilde{c}_2 h_E^{1/2}\|\nabla v\|_{L^2(\tilde{\omega}_K)},
\end{aligned} \tag{5.20}$$

where K' denotes the element of $\mathcal{T}_{h_n}$ that contains K and $\tilde{\omega}_K$ denotes the subset that consists of all elements of $\tilde{\mathcal{T}}_{h_n}$ sharing at least a vertex with K'. The constants c_0, c_1, c_2 only depend on the maximal ratio of the diameter of any element to the diameter of its largest inscribed ball. The constants $\tilde{c}_1, \tilde{c}_2$ in addition depend on the maximal ratio $h_{K'}/h_K$.

With η_K denoting a unit outward pointing normal we further derive

$$\begin{aligned}
A_{spa}(y,t) = \sum_{k=1}^{N_c(I_{n+1})} \Big[&\int_D f(y_k)(v - v_{h_{n+1}}) - \int_D \frac{\partial \tilde{u}}{\partial t}(y_k)\Big(v - v_{h_{n+1}}\Big) \\
&- \int_D a(y_k)\nabla\tilde{u}(y_k)\nabla(v - v_{h_{n+1}})\Big] L_k(y) \\
= \sum_{k=1}^{N_c(I_{n+1})} \Big[&\sum_{K\in\tilde{\mathcal{T}}_{h_{n+1}}} \int_K \Big[f(y_k) - \frac{\partial \tilde{u}}{\partial t}(y_k) + \nabla\cdot\Big(a(y_k)\nabla\tilde{u}(y_k)\Big)\Big](v - v_{h_{n+1}})\Big] \\
&- \Big[\sum_{E\in\tilde{\mathcal{E}}_{h_{n+1}}} \int_E [a(y_k)\nabla\tilde{u}(y_k)\cdot\eta_E]_E(v - v_{h_{n+1}})\Big] L_k(y).
\end{aligned}$$

Considering $v_{h_{n+1}} = J_{h_{n+1}}(v)$ leads us to

$$\begin{aligned}
A_{spa}(y,t) \le \sum_{k=1}^{N_c(I_{n+1})} \Big[&\sum_{K\in\tilde{\mathcal{T}}_{h_{n+1}}} \tilde{c}_1 h_K \left\| f(y_k) - \frac{\partial \tilde{u}}{\partial t}(y_k) + \nabla\cdot\Big(a(y_k)\nabla\tilde{u}(y_k)\Big)\right\|_{L^2(K)} \|\nabla v\|_{L^2(\tilde{\omega}_K)} \\
&+ \sum_{E\in\tilde{\mathcal{E}}_{h_{n+1}}} \tilde{c}_2 h_E^{1/2} \left\|[a(y_k)\nabla\tilde{u}(y_k)\cdot\eta_E]_E\right\|_{L^2(E)} \|\nabla v\|_{L^2(\tilde{\omega}_K)}\Big] \Big|L_k(y)\Big|.
\end{aligned}$$

Now, using the discrete Cauchy–Schwarz inequality and the fact that the domains $\tilde{\omega}_K$ only consist of a finite number of elements, this number being bounded by the maximal ratio of the diameter of any element to the diameter of its largest inscribed ball and on the ratios $h_{K'}/h_K$, we derive

$$\begin{aligned}
A_{spa}(y,t) \le C_1 \sum_{k=1}^{N_c(I_{n+1})} \Bigg[&\Big(\sum_{K\in\tilde{\mathcal{T}}_{h_{n+1}}} h_K^2 \left\| f(y_k) - \frac{\partial \tilde{u}}{\partial t}(y_k) + \nabla\cdot\Big(a(y_k)\nabla\tilde{u}(y_k)\Big)\right\|_{L^2(K)}^2\Big)^{1/2} \\
&+ \Big(\sum_{E\in\tilde{\mathcal{E}}_{h_{n+1}}} h_E \left\|[a(y_k)\nabla\tilde{u}(y_k)\cdot\eta_E]_E\right\|_{L^2(E)}^2\Big)^{1/2}\Bigg] \Big|L_k(y)\Big| \; \|\nabla v\|_{L^2(D)} \\
= C_1 \sum_{k=1}^{N_c(I_{n+1})} &\mathcal{E}_{spa,k}^{n+1}(t) \Big|L_k(y)\Big| \; \|\nabla v\|_{L^2(D)}.
\end{aligned}$$

As for the temporal part A_{tem}, we proceed in a similar manner

$$
\begin{aligned}
A_{tem}(y,t) = &\sum_{k=1}^{N_c(I_{n+1})} \Big[\int_D a(y_k)\Big(\nabla \tilde{u}^{n\theta}(y_k) - \nabla \tilde{u}(y_k)\Big)\nabla v_{h_{n+1}} \\
&+ \int_D \Big(f(y_k) - f^{n\theta}(y_k)\Big) v_{h_{n+1}} \Big] L_k(y) \\
&\le C_2 \sum_{k=1}^{N_c(I_{n+1})} \Big[\left\| a(y_k)\nabla\Big(\tilde{u}^{n\theta}(y_k) - \tilde{u}(y_k)\Big)\right\|_{L^2(D)} \\
&+ \left\| \Big(f(y_k) - f^{n\theta}(y_k)\Big)\right\|_{L^2(D)} \Big] \Big| L_k(y) \Big| \; \|\nabla v\|_{L^2(D)} \\
&= C_2 \sum_{k=1}^{N_c(I_{n+1})} \mathcal{E}^{n+1}_{tem,k}(t) \Big| L_k(y) \Big| \; \|\nabla v\|_{L^2(D)} \,,
\end{aligned}
\tag{5.21}
$$

where C_2 depends on the $J_{h_{n+1}}$ interpolation operator norm and the Poincaré constant.

Now we focus on the term A_{sto} describing the stochastic part of the error. We will use the fact that $S_{I_{n+1}}[a\nabla u_{h_{n+1}}] = S_{I_{n+1}}[a\nabla S_{I_{n+1}}[u_{h_{n+1}}]]$ given by Proposition 5.1.1. We derive

$$
\begin{aligned}
A_{sto}(y,t) &= S_{I_{n+1}}\Big[\int_D a\nabla\tilde{u}\nabla v + \int_D \frac{\partial \tilde{u}}{\partial t} - \int_D fv \Big] - \Big(\int_D a\nabla\tilde{u}\nabla v + \int_D \frac{\partial \tilde{u}}{\partial t} - \int_D fv \Big) \\
&= \int_D \Big(S_{I_{n+1}}[a\nabla\tilde{u}] - a\nabla\tilde{u} \Big)\nabla v + \int_D \Big(f - S_{I_{n+1}}[f] \Big) v \\
&+ \int_D \left(\frac{S_{I_{n+1}}[u^{n+1}_{h_{n+1},I_{n+1}} - u^n_{h_n,I_n}] - (u^{n+1}_{h_{n+1},I_{n+1}} - u^n_{h_n,I_n})}{\tau_{n+1}} \right) v \\
&= \int_D \Big(S_{I_{n+1}}[a\nabla\tilde{u}] - a\nabla\tilde{u} \Big)\nabla v + \int_D \Big(f - S_{I_{n+1}}[f] \Big) v \\
&+ \int_D \frac{S_{\tilde{I}_{n+1}}[u^n_{h_n,I_n}] - S_{I_{n+1}}[u^n_{h_n,I_n}]}{\tau_{n+1}} v \\
&\le \left\| \sum_{i\in I^C_{n+1}} \triangle^{m(i)}(a\nabla\tilde{u}) \right\|_{L^2(D)} \|\nabla v\|_{L^2(D)} + \left\| \sum_{i\in I^C_{n+1}} \triangle^{m(i)}(f) \right\|_{L^2(D)} \|v\|_{L^2(D)} \\
&+ \frac{1}{\tau_{n+1}} \left\| \sum_{i\in \tilde{I}_{n+1}\setminus I_{n+1}} \triangle^{m(i)}(u^n_{h_n,I_n}) \right\|_{L^2(D)} \|v\|_{L^2(D)}
\end{aligned}
$$

$$\begin{aligned}
&\leq C_3 \Bigg(\Bigg\| \sum_{i\in(I_n\setminus I_{n+1})\cup M_{\tilde{I}_{n+1}}} \triangle^{m(i)}(a\nabla\tilde{u}) \Bigg\|_{L^2(D)} + \Bigg\| \sum_{i\in I_{n+1}^C} \triangle^{m(i)}(f) \Bigg\|_{L^2(D)} \\
&\quad + \frac{1}{\tau_{n+1}} \Bigg\| \sum_{i\in\tilde{I}_{n+1}\setminus I_{n+1}} \triangle^{m(i)}(u^n_{h_n,I_n}) \Bigg\|_{L^2(D)} \Bigg) \|\nabla v\|_{L^2(D)} \\
&= C_3\, \mathcal{E}^{n+1}_{sto}\, \|\nabla v\|_{L^2(D)}.
\end{aligned}$$

In the last inequality we used the affine dependence of a on the random variables, stated in the assumption (5.3), which allows us to restrict the sum over I_{n+1}^C to the index set $M_{\tilde{I}_{n+1}} \cup (I_n \setminus I_{n+1}) = I_{n+1}^C \cap (\tilde{I}_{n+1} \cup M_{\tilde{I}_{n+1}})$. This comes from the fact that $\tilde{u} \in \mathbb{P}_{\tilde{I}_{n+1}}$ which implies $a\nabla\tilde{u} \in \mathbb{P}_{\tilde{I}_{n+1}\cup M_{\tilde{I}_{n+1}}}$. Since $S_{\tilde{I}_{n+1}\cup M_{\tilde{I}_{n+1}}}$ is exact on $\mathbb{P}_{\tilde{I}_{n+1}\cup M_{\tilde{I}_{n+1}}}$ (Proposition 5.1.1.3), we obtain

$$\triangle^{m(i)}(a\nabla\tilde{u}) = 0, \quad \forall i \notin \tilde{I}_{n+1} \cup M_{\tilde{I}_{n+1}}.$$

Altogether we obtained for every $n = 0, \dots, N-1$

$$\begin{aligned}
&\int_D \frac{\partial(u-\tilde{u})}{\partial t} v + \int_D a\nabla(u-\tilde{u})\cdot\nabla v \\
&\qquad \leq C_4 \Bigg(\mathcal{E}^{n+1}_{sto} + \sum_{k=1}^{N_c(I_{n+1})} (\mathcal{E}^{n+1}_{tem,k} + \mathcal{E}^{n+1}_{spa,k}) \Big| L_k(y) \Big| \Bigg) \|\nabla v\|_{L^2(D)}.
\end{aligned}$$

Taking $v = u - \tilde{u} \in H^1_0$ and using the Young inequality we have

$$\begin{aligned}
&\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u-\tilde{u}\|^2_{L^2(D)}(y,t) + c^2_{min}\|\nabla(u-\tilde{u})\|^2_{L^2(D)}(y,t) \\
&\qquad \leq \frac{1}{2\,c^2_{min}} C_4^2 \Bigg(\mathcal{E}^{n+1}_{sto}(y,t) + \sum_{k=1}^{N_c(I_{n+1})} \Big(\mathcal{E}^{n+1}_{tem,k}(t) + \mathcal{E}^{n+1}_{spa,k}(t) \Big) \Big| L_k(y) \Big| \Bigg)^2 \\
&\qquad\qquad + \frac{c^2_{min}}{2} \|\nabla(u-\tilde{u})\|^2_{L^2(D)}(y,t)
\end{aligned}$$

which holds for a.e. $t \in (t_n, t_{n+1}]$. The last step is to integrate the last inequality w.r.t t over $(0,T)$ and w.r.t y over Γ. Using the discrete Cauchy-Schwarz inequality we derive

$$\begin{aligned}
&\sum_{n=0}^{N-1} \int_{t_n}^{t_{n+1}} \int_\Gamma \Big(\sum_{k=1}^{N_c(I_{n+1})} \mathcal{E}^{n+1}_{tem,k}(t) \Big| L_k(y) \Big| \Big)^2 \rho(y)\,\mathrm{d}y\,\mathrm{d}t \\
&\quad \leq \sum_{n=0}^{N-1} \int_{t_n}^{t_{n+1}} \int_\Gamma \sum_{k=1}^{N_c(I_{n+1})} \mathcal{E}^{n+1}_{tem,k}(t)^2 |L_k(y)| \sum_{k=1}^{N_c(I_{n+1})} \Big| L_k(y) \Big| \rho(y)\,\mathrm{d}y\,\mathrm{d}t
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{n=0}^{N-1} \Big(\sum_{k=1}^{N_c(I_{n+1})} \int_{t_n}^{t_{n+1}} \mathcal{E}_{tem,k}^{n+1}(t)^2 \,\mathrm{d}t \int_\Gamma \Big| L_k(y) \Big| \rho(y) \,\mathrm{d}y \Big) \Big(\sup_{y\in\Gamma} \sum_{k=1}^{N_c(I_{n+1})} |L_k(y)| \Big) \\
&\leq \sum_{n=0}^{N-1} \Lambda_{I_{n+1}} \Big(\sum_{k=1}^{N_c(I_{n+1})} 2\Big(\|f(y_k) - f^{n\theta}(y_k)\|^2_{L^2(t_n,t_{n+1};L^2(D))} \\
&\quad + \tau_{n+1} \, \frac{\theta^3 + (1-\theta)^3}{3} \, \|a(y_k)\nabla(u^{n+1}_{h_{n+1},I_{n+1}} - u^n_{h_n,I_n})\|^2_{L^2(D)} \Big) \|L_k\|_{L^1_\rho(\Gamma)} \Big),
\end{aligned}
$$

where in the last inequality we employed the observation

$$
\tilde{u}^{n\theta} - \tilde{u} = \Big(\theta - \frac{t - t_n}{\tau_{n+1}} \Big) \, (u^{n+1} - u^n).
$$

Analogously for the spatial part

$$
\begin{aligned}
&\sum_{n=0}^{N-1} \int_{t_n}^{t_{n+1}} \int_\Gamma \Big(\sum_{k=1}^{N_c(I_{n+1})} \mathcal{E}_{spa,k}^{n+1}(t) \Big| L_k(y) \Big| \Big)^2 \rho(y) \,\mathrm{d}y \,\mathrm{d}t \\
&\quad \leq \sum_{n=0}^{N-1} \int_{t_n}^{t_{n+1}} \int_\Gamma \sum_{k=1}^{N_c(I_{n+1})} \mathcal{E}_{spa,k}^{n+1}(t)^2 \Big| L_k(y) \Big| \sum_{k=1}^{N_c(I_{n+1})} \Big| L_k(y) \Big| \rho(y) \,\mathrm{d}y \,\mathrm{d}t \\
&\quad \leq \sum_{n=0}^{N-1} \Lambda_{I_{n+1}} \Big(\sum_{k=1}^{N_c(I_{n+1})} \Big(\sum_{K\in\tilde{\mathcal{T}}_{h_{n+1}}} h_K^2 \Big\| f(y_k) - \frac{\partial \tilde{u}}{\partial t}(y_k) + \nabla \cdot \Big(a(y_k) \nabla \tilde{u}(y_k) \Big) \Big\|^2_{L^2(t_n,t_{n+1};L^2(K))} \\
&\qquad + \sum_{E\subset\partial K} h_E \Big\| \frac{1}{2} [a(y_k)\nabla\tilde{u}(y_k) \cdot \eta_E]_E \Big\|^2_{L^2(t_n,t_{n+1};L^2(E))} \Big) \|L_k\|_{L^1_\rho(\Gamma)} \Big)
\end{aligned}
$$

As for the stochastic part we derive

$$
\begin{aligned}
&\sum_{n=0}^{N-1} \int_{t_n}^{t_{n+1}} \int_\Gamma \mathcal{E}_{sto}^{n+1}(y,t)^2 \rho(y) \,\mathrm{d}y \,\mathrm{d}t \\
&\leq 3 \sum_{n=0}^{N-1} \int_{t_n}^{t_{n+1}} \int_\Gamma \sum_{i\in(I_n\setminus I_{n+1})\cup M_{\tilde{I}_{n+1}}} \Big\| \triangle^{m(i)}(a\nabla\tilde{u}) \Big\|^2_{L^2(D)} \\
&\quad + \sum_{i\in I_{n+1}^C} \Big\| \triangle^{m(i)}(f) \Big\|^2_{L^2(D)} + \sum_{i\in\tilde{I}_{n+1}\setminus I_{n+1}} \frac{1}{\tau_{n+1}^2} \Big\| \triangle^{m(i)}(u^n_{h_n,I_n}) \Big\|^2_{L^2(D)} \rho(y) \,\mathrm{d}y \,\mathrm{d}t \\
&\leq C_5 \sum_{n=0}^{N-1} \Big[\int_{t_n}^{t_{n+1}} \Big(\frac{t - t_n}{\tau_{n+1}} \Big)^2 \mathrm{d}t \sum_{i\in I_{n+1}^C\cap(I_n\cup M_{I_n})} \Big\| \triangle^{m(i)}(a\nabla u^n_{h_n,I_n}) \Big\|^2_{L^2_\rho(\Gamma;L^2(D))} \\
&\quad + \int_{t_n}^{t_{n+1}} \Big(\frac{t_{n+1} - t}{\tau_{n+1}} \Big)^2 \mathrm{d}t \sum_{i\in M_{I_{n+1}}} \Big\| \triangle^{m(i)}(a\nabla u^{n+1}_{h_{n+1},I_{n+1}}) \Big\|^2_{L^2_\rho(\Gamma;L^2(D))} \\
&\quad + \sum_{i\in I_{n+1}^C} \Big\| \triangle^{m(i)}(f) \Big\|^2_{L^2(t_n,t_{n+1};L^2_\rho(\Gamma,L^2(D)))}
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{\tau_{n+1}}{\tau_{n+1}^2} \sum_{i\in \tilde{I}_{n+1}\setminus I_{n+1}} \left\| \triangle^{m(i)}(u^n_{h_n,I_n})\right\|^2_{L^2_\rho(\Gamma;L^2(D))} \Bigg] \\
&= C_5 \sum_{n=0}^{N-1} \Bigg[\frac{\tau_{n+1}}{3} \Bigg(\sum_{i\in I^C_{n+1}\cap(I_n\cup M_{I_n})} \left\| \triangle^{m(i)}(a\nabla u^n_{h_n,I_n})\right\|^2_{L^2_\rho(\Gamma;L^2(D))} \\
&+ \sum_{i\in M_{I_{n+1}}} \left\| \triangle^{m(i)}(a\nabla u^{n+1}_{h_{n+1},I_{n+1}})\right\|^2_{L^2_\rho(\Gamma;L^2(D))} \Bigg) \\
&+ \sum_{i\in I^C_{n+1}} \left\| \triangle^{m(i)}(f)\right\|^2_{L^2(t_n,t_{n+1};L^2_\rho(\Gamma,L^2(D)))} \\
&+ \frac{1}{\tau_{n+1}} \sum_{i\in \tilde{I}_{n+1}\setminus I_{n+1}} \left\| \triangle^{m(i)}(u^n_{h_n,I_n})\right\|^2_{L^2_\rho(\Gamma;L^2(D))} \Bigg]
\end{aligned}
$$

□ □

The spatial and time estimators in (5.16), (5.17) depend on the Lebesgue constant Λ_{n+1}. The growth of the Lebesgue constant depends on the choice of the level function m and the type of the collocation points, which for the nested Clenshaw-Curtis points yields an estimate $\Lambda_I \sim |I|^2$ and for the projected Leja points an estimate $\Lambda_I \sim |I|^{3+\varepsilon}$ for any $\varepsilon > 0$ (see [CCS14]). Such estimation can cause the estimator to be too conservative. This issue was addressed in [GN18, Rem 4.4]. The following theorem provides an alternative way of estimating the spatial and time estimator without involving the Lebesgue constant and is an extension of the results from [GN18, Rem 4.4].

Theorem 5.1.3. *The spatial estimator* ϵ_{spa} *from* (5.16) *and the time estimator* ϵ_{tem} *from* (5.17) *can be alternatively expressed as*

$$
\begin{aligned}
\epsilon^2_{tem} = \frac{C}{c^2_{min}} \sum_{n=0}^{N-1} \Bigg[&\left\| \sum_{k=1}^{N_c(I_{n+1})} \Big[\Big(f(y_k) - f^{n\theta}(y_k)\Big)\Big] L_k(y) \right\|^2_{L^2(t_n,t_{n+1};L^2_\rho(\Gamma;L^2(D)))} \\
&+ \tau_{n+1} \left\| \sum_{k=1}^{N_c(I_{n+1})} \Big[a(y_k)\nabla\Big(u^{n+1}_{h_{n+1},I_{n+1}}(y_k) - u^n_{h_n,I_n}(y_k)\Big)\Big] L_k(y) \right\|^2_{L^2_\rho(\Gamma;L^2(D))} \Bigg]
\end{aligned}
\tag{5.22}
$$

and

$$
\epsilon^2_{spa} = \sum_{n=0}^{N-1} \sum_{K\in\tilde{\mathcal{T}}_{h_{n+1}}} (\epsilon^n_{spa,K})^2 \tag{5.23}
$$

with

$$
(\epsilon^n_{spa,K})^2 = \frac{C}{c^2_{min}} h_K^2 \left\| \sum_{k=1}^{N_c(I_{n+1})} \Big[f(y_k) - \frac{\partial \tilde{u}}{\partial t}(y_k) + \nabla\cdot\Big(a(y_k)\nabla\tilde{u}(y_k)\Big)\Big] L_k(y) \right\|^2_{L^2(t_n,t_{n+1};L^2_\rho(\Gamma;L^2(K)))}
$$

$$+\sum_{E\subset\partial K} h_E \left\| \sum_{k=1}^{N_c(I_{n+1})} \Big(\frac{1}{2}[a(y_k)\nabla\tilde{u}(y_k)\cdot\eta_E]_E\Big) L_k(y) \right\|^2_{L^2(t_n,t_{n+1};L^2_\rho(\Gamma;L^2(E)))}.$$

Proof. We follow by estimating the term A_{tem} from (5.21) by

$$\begin{aligned}
A_{tem}(y,t) &= \int_D \sum_{k=1}^{N_c(I_{n+1})} a(y_k)\Big(\nabla\tilde{u}^{n\theta}(y_k)-\nabla\tilde{u}(y_k)\Big)L_k(y)\,\nabla v_{h_{n+1}} \\
&+\int_D \sum_{k=1}^{N_c(I_{n+1})} \Big(f(y_k)-f^{n\theta}(y_k)\Big)L_k(y)\,v_{h_{n+1}} \\
&\le C\Big(\left\| \sum_{k=1}^{N_c(I_{n+1})} a(y_k)\Big(\nabla\tilde{u}^{n\theta}(y_k)-\nabla\tilde{u}(y_k)\Big)\,L_k(y)\right\|_{L^2(D)} \\
&+\left\| \sum_{k=1}^{N_c(I_{n+1})} \Big(f(y_k)-f^{n\theta}(y_k)\Big)\,L_k(y)\right\|_{L^2(D)}\Big)\,\|\nabla v\|_{L^2(D)} \\
&= C\mathcal{E}^{n+1}_{tem}\,\|\nabla v\|_{L^2(D)}
\end{aligned}$$

where we applied the same interpolation results as proposed in (5.20). Analogously for the spatial estimation we derive

$$\begin{aligned}
A_{spa}(y,t) = &\sum_{K\in\tilde{\mathcal{T}}_{h_{n+1}}}\Big[\int_K \sum_{k=1}^{N_c(I_{n+1})}\Big[f(y_k)-\frac{\partial\tilde{u}}{\partial t}(y_k)+\nabla\cdot\Big(a(y_k)\nabla\tilde{u}(y_k)\Big)\Big]L_k(y) \\
&\qquad (v-v_{h_{n+1}})\Big] \\
&-\sum_{E\in\tilde{\mathcal{E}}_{h_{n+1}}}\int_E \sum_{k=1}^{N_c(I_{n+1})}[a(y_k)\nabla\tilde{u}(y_k)\cdot\eta_E]_E\,L_k(y)(v-v_{h_{n+1}}) \\
\le &\sum_{K\in\tilde{\mathcal{T}}_{h_{n+1}}}\left\|\sum_{k=1}^{N_c(I_{n+1})}\Big[f(y_k)-\frac{\partial\tilde{u}}{\partial t}(y_k)+\nabla\cdot\Big(a(y_k)\nabla\tilde{u}(y_k)\Big)\Big]L_k(y)\right\|_{L_2(K)}\|v-v_{h_{n+1}}\|_{L^2(K)} \\
&+\sum_{E\in\tilde{\mathcal{E}}_{h_{n+1}}}\left\|\sum_{k=1}^{N_c(I_{n+1})}[a(y_k)\nabla\tilde{u}(y_k)\cdot\eta_E]_E\,L_k(y)\right\|_{L^2(E)}\|v-v_{h_{n+1}}\|_{L^2(E)}.
\end{aligned}$$

Using again the interpolation results (5.20) and the discrete Cauchy-Schwarz inequality we obtain

$$A_{spa}(y,t)\le C\Big[\Big(\sum_{K\in\tilde{\mathcal{T}}_{h_{n+1}}} h_K^2\left\|\sum_{k=1}^{N_c(I_{n+1})}\Big[f(y_k)-\frac{\partial\tilde{u}}{\partial t}(y_k)+\nabla\cdot\Big(a(y_k)\nabla\tilde{u}(y_k)\Big)\Big]L_k(y)\right\|^2_{L_2(K)}\Big)^{1/2}$$

$$+\left(\sum_{E\in\tilde{\mathcal{E}}_{h_{n+1}}} h_E \left\|\sum_{k=1}^{N_c(I_{n+1})}[a(y_k)\nabla\tilde{u}(y_k)\cdot\eta_E]_E\, L_k(y)\right\|_{L^2(E)}\right)^{1/2}\Bigg] \|\nabla v\|_{L^2(D)}$$
$$= C\mathcal{E}_{spa}^{n+1}\,\|\nabla v\|_{L^2(D)}$$

The rest of the proof follows the same steps as in the proof of Theorem 5.1.2. □ □

Remark 6. Note that the spatial estimator from Theorem 5.1.2 allows for different FE meshes for different collocation points. This property is sacrificed in the spatial estimator from Theorem 5.1.3. We shall also note that the error estimator derived in this work is of suboptimal order in the case $\theta = 1/2$, which corresponds to the Crank-Nicolson scheme. In order to restore the second order convergence one shall work with a piecewise quadratic polynomial function in time instead of the linear one defined in (5.14). We refer an interested reader to [LPP09; AMN06].

5.1.5 Adaptive algorithm

The estimators from the preceding section provide us with an upper bound of the error that is naturally localized in all variables - time, space and stochastics. There are many possible choices of adaptive algorithms that can be constructed starting from these estimators. One could drive the adaptive choice of time-varying finite element and stochastic grids by a local in time error estimator, as was proposed in [Pic98; BR03] for time varying FE or DG meshes in the case of a deterministic heat equation. Also, the spatial estimator ϵ_{spa} in Theorem 5.1.2 is naturally localized over the collocation points so it allows for different adapted FE meshes in different collocation points. This idea has been explored e.g. in [Eig+14] in the context of a stochastic Galerkin polynomial chaos approximation of an elliptic problem with random coefficients. There are, however, many problems whose behaviour does not require FE meshes and sparse grids that dramatically change in time. Considering fixed in time FE meshes and sparse grids simplifies the estimators and the adaptive process. In this work we will restrict to adapted FE meshes and sparse grids which are fixed in time with the goal to obtain the overall error

$$\varepsilon = \|(u-\tilde{u})(T)\|^2_{L^2_\rho(\Gamma;L^2(D))} + c^2_{min}\,\|u-\tilde{u}\|^2_{L^2(0,T;L^2_\rho(\Gamma;H^1_0(D)))}$$

under a prescribed tolerance TOL. We will apply the global spatial and time estimators from Theorem 5.1.3, i.e. (5.22), (5.23) by localizing the spatial estimator into elements, the time estimator into time steps and the stochastic estimator (5.18) into indices. For a deterministic right hand side the corresponding error estimators become

$$\begin{aligned}\epsilon^2_{spa,K} = \frac{c_1}{c^2_{min}} h_K^2 &\left\| \sum_{k=1}^{N_c(I)} \left[f(y_k) - \frac{\partial \tilde{u}}{\partial t}(y_k) + \nabla \cdot \Big(a(y_k) \nabla \tilde{u}(y_k) \Big) \right] L_k(y) \right\|^2_{L^2(0,T;L^2_\rho(\Gamma;L^2(K)))} \\ &+ \frac{c_2}{c^2_{min}} \sum_{E \subset \partial K} h_E \left\| \sum_{k=1}^{N_c(I)} \left(\frac{1}{2} [a(y_k) \nabla \tilde{u}(y_k) \cdot \eta_E]_E \right) L_k(y) \right\|^2_{L^2(0,T;L^2_\rho(\Gamma;L^2(E)))}\end{aligned} \tag{5.24}$$

for every element $K \in \mathcal{T}_h$,

$$\begin{aligned}\epsilon^2_{tem,n} = \frac{c_3}{c^2_{min}} &\left\| \sum_{k=1}^{N_c(I)} \left[\Big(f(y_k) - f^{n\theta}(y_k) \Big) \right] L_k(y) \right\|^2_{L^2(t_n,t_{n+1};L^2_\rho(\Gamma;L^2(D)))} \\ &+ \frac{c_4}{c^2_{min}} \tau_{n+1} \left\| \sum_{k=1}^{N_c(I)} \left[a(y_k) \nabla \Big(u^{n+1}_{h,I}(y_k) - u^n_{h,I}(y_k) \Big) \right] L_k(y) \right\|^2_{L^2_\rho(\Gamma;L^2(D))}\end{aligned} \tag{5.25}$$

for every subinterval $[t_n, t_{n+1}]$, $n = 0, \ldots, N-1$, and

$$\epsilon^2_{sto,i} = \frac{1}{c^2_{min}} \sum_{n=0}^{N-1} \tau_{n+1} \left(\left\| \triangle^{m(i)} (a \nabla u^n_{h,I}) \right\|^2_{L^2_\rho(\Gamma;L^2(D))} + \left\| \triangle^{m(i)} (a \nabla u^{n+1}_{h,I}) \right\|^2_{L^2_\rho(\Gamma;L^2(D))} \right) \tag{5.26}$$

for every multi index $i \in M_I$.

Then the overall error ε can be bounded by

$$\varepsilon^2 \leq \sum_{K \in \mathcal{T}_h} \epsilon^2_{spa,K} + \sum_{n=0}^{N-1} \epsilon^2_{tem,n} + \sum_{i \in M_I} \epsilon^2_{sto,i}.$$

The algorithm will start with fairly coarse grids and index set $\mathcal{T}_h$, τ, I, compute the numerical solution $\tilde{u}$ and compute the estimators (5.24), (5.25), (5.26) for every cell, time subinterval and index from the margin. Let $\mathcal{N} = |\mathcal{T}_h| + N + |M_I|$ denote the total number of elements in $\{\mathcal{T}_h, \tau, M_I\}$, i.e. number of cells + number of subintervals (N) + number of indices in the margin M_I. Then we will refine a cell K whenever $\epsilon^2_{spa,K} > (\alpha TOL/\mathcal{N})^2$, divide a time interval $[t_n, t_{n+1}]$ into 2 equal subintervals whenever $\epsilon^2_{tem,n} > (\alpha TOL/\mathcal{N})^2$ and add an index $i \in M_I$ into the index set I whenever $\epsilon^2_{sto,i} > (\alpha TOL/\mathcal{N})^2$, where $\alpha > 1$. Note that adding an index i might result in adding more indices since we need to keep the index set I downward closed. With the new refined mesh, time grid and sparse grid we need to compute a new solution $\tilde{u}$ and continue until the stopping criterion

$$\varepsilon^2_{\mathcal{T}_h,\tau,I} := \sum_{K\in\mathcal{T}_h} \epsilon^2_{spa,K} + \sum_{n=0}^{N-1} \epsilon^2_{tem,n} + \sum_{i\in M_I} \epsilon^2_{sto,i} < TOL^2$$

is satisfied. This procedure is described in Algorithm 1. We shall note that a proof of convergence for this algorithm is not available yet.

Algorithm 1: Adaptive algorithm

```
Data: $TOL > 0$
Result: $\tau, I, \mathcal{T}_h$ and $\tilde{u}$ s.t. $\varepsilon_{\mathcal{T}_h,\tau,I} < TOL$
Initialize $\tau, I, \mathcal{T}_h$;
compute $\tilde{u}$ on $\tau, I, \mathcal{T}_h$;
compute $\epsilon_{spa,K}$, $\epsilon_{tem,n}$, $\epsilon_{sto,i}$;
while $\varepsilon_{\mathcal{T}_h,\tau,I} \geq TOL$ do
    set $\mathcal{N} = |\mathcal{T}_h| + N + |M_I|$;
    for $K \in \mathcal{T}_h$ do
        if $\epsilon_{spa,K} > \alpha \frac{TOL}{\mathcal{N}}$ then
            refine $K$
    for $n \in \{0, \dots, N-1\}$ do
        if $\epsilon_{tem,n} > \alpha \frac{TOL}{\mathcal{N}}$ then
            refine $[t_n, t_{n+1}]$
    for $i \in M_I$ do
        if $\epsilon_{sto,i} > \alpha \frac{TOL}{\mathcal{N}}$ then
            $I = I \cup i$;
            add indices s.t. $I$ is downward closed
    update $\tau, I, \mathcal{T}_h$;
    compute $\tilde{u}$ on new $\tau, I, \mathcal{T}_h$;
    compute $\epsilon_{spa,K}$, $\epsilon_{tem,n}$, $\epsilon_{sto,i}$;
```

5.1.6 Numerical results

This section is dedicated to study the effectiveness of the estimators in (5.24), (5.25), (5.26) and the performance of the adaptive algorithm introduced in Section 5.1.5. The practical computation of these estimators requires some estimation of the constants $c_1, \dots, c_4, c_{min}$ as well as an approximate computation of the $L^2_\rho(\Gamma)$ norm. This is discussed hereafter.

Let us consider problem (5.1.1) set in a unit square $D = [0,1]^2$ with time domain $[0,1]$

and an uncertain diffusion coefficient

$$a(x,y) = a_0 + \sum_{m=1}^{2} \frac{cos(2\pi m x_1) + cos(2\pi m x_2)}{(\pi m)^2} y_m \tag{5.27}$$

with $x = (x_1, x_2)$, $y = (y_1, y_2)$ and $a_0 > 0$ set to satisfy

$$\inf_{x \in D,\, y \in \Gamma} a(x,y) = 0.01.$$

The random variables are independent and uniformly distributed $y_m \sim U([-1,1])$ and the forcing term is deterministic and time-independent

$$f(x) = 20\, \mathbb{1}_F(x)$$

with $F = [0.4, 0.6] \times [0.4, 0.6]$ a square in the middle of the domain.

In all of our simulations we used the spatial and time estimators provided in Theorem 5.1.3 which do not require an explicit estimation of the Lebesgue constant. The norm $\|g\|_{L^2_\rho(\Gamma)}$ for $g \in C^0(\Gamma)$ is approximated using a set $\Theta \subset \Gamma$ of finite cardinality by

$$\|g\|_{L^2_\rho(\Gamma)} \approx \left(\frac{1}{|\Theta|} \sum_{y \in \Theta} g(y)^2 \right)^{1/2}.$$

We set Θ to consist of 500 randomly sampled points in $\Gamma = [-1,1]^2$ according to the distribution ρ, uniform on Γ. As suggested in [GN18], instead of setting $c_{min} = \sqrt{a_{min}}$, which may be too conservative, we will rather approximate it by

$$c_{min} := \min_{v \in U \subset L^2_\rho(\Gamma; H^1_0(D))} \min_{y \in \Xi} \frac{\|a^{1/2}(y) \nabla v(y)\|_{L^2(D)}}{\|\nabla v(y)\|_{L^2(D)}},$$

where we take $U = \{u^n_{h,I},\ n = 0, \dots, N\}$ and Ξ is a set of random samples of small cardinality (different from Θ). For the specific diffusion coefficient in (5.27) we estimated $c_{min} \approx 0.41$. The norm $\|g\|_{L^2(0,T)}$ is computed using the trapezoidal rule as suggested in [Pic98]. We have considered $P1$ finite elements without fitting the FE mesh to the subdomain F and $\theta = 1$, namely the implicit Euler method. The sparse grid consists of Leja points built as symmetric Leja sequences within $[-1,1]$ (see e.g. [CCS14]) with level function $m(i) = i$. This combination satisfies the interpolatory condition (5.10). All our simulations were performed using the FEniCS library [Aln+15b].

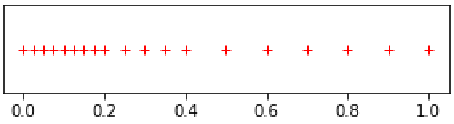
For a sharp behaviour of the estimators and an efficient performance of the adaptive algorithm one needs a good estimation of the constants c_1, c_2, c_3, c_4 in (5.24), (5.25), (5.26). This requires a good estimation of the interpolation constants from (5.20) which is not an easy task and we refer the reader to [Ver99a] for ways to bound the interpolation constants. In our case, we adopted the strategy proposed in [Pic98], i.e. estimated the constants by observing the behaviour of the estimators vs. the behaviour of the error with respect to a reference solution when refining individually uniform spatial grids, uniform time grids and isotropic sparse grids for different solutions u. This is done on relatively coarse FE meshes, sparse grids and time discretizations so that the overall cost of estimating the constants is much smaller than the cost of the adaptive process. We obtained the estimates $c_1 = 0.016, c_2 = 0.023, c_4 = 0.078$. The term including the constant c_3 is in our case equal to 0. Note that a correct estimation of the constants is rather important. Poor estimation directly influences the stopping criterion. Moreover, it may alter the balance between the three sources of discretization error which may lead to over-refining in one variable while insufficiently refining in another one.

Numerical study of the performance of the estimators

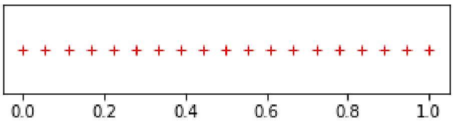
This part is dedicated to study the effectiveness of the error estimator considering different non uniform FE meshes, time discretizations and index sets. We proceed by studying first a “marginalized” error and estimator, where by “marginalized” spatial error we mean an error caused by only spatial discretization, i.e. the numerical solution and the “true” solution are computed using the same (“overkilling”) discretization for time and random variables. Analogously, for the marginalized time and stochastic errors.
In Figure 5.1 we show the convergence results for the marginalized time estimator. The numerical and reference solution were computed on a uniform spatial grid consisting of 6400 triangles, with diameter 0.025 and a sparse grid having 113 collocation points. We considered both uniform and non uniform time discretizations for the numerical solution specified in Figure 5.1 (left). The reference solution was computed on a much finer time grid. Figure 5.1 (right) shows the “true" error as well as the error estimator over the sequence of time grids obtained by refinement of the three grids shown in the left plot. We observe that the estimator provides a good control of the “true" error for all three cases. In the same figure we have also plotted the reference function $2/N$ where N is the number of time steps. We clearly see that the order of convergence is 1 in all cases, which is expected as we used the implicit Euler method for the time discretization. We can as well observe that the smallest error is attained for the non-uniform discretization that is denser at the beginning of the interval. This is related to the fact that the considered problem has a dissipative behavior.

The convergence study of the stochastic estimator was performed on a triangulation with 6400 triangles with diameter 0.025 and with 200 uniform time steps. The results are shown in Figure 5.2. We considered sequences of anisotropic sparse grids with the index sets defined as

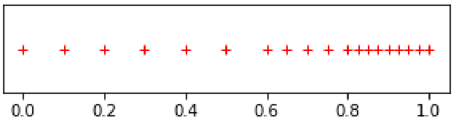
$$I(w) = \{i \in \mathbb{N}_+^M \, : \, \sum_n \beta_n (i_n - 1) \leq w\}$$



(a) Time discretization 1



(b) Time discretization 2



(c) Time discretization 3

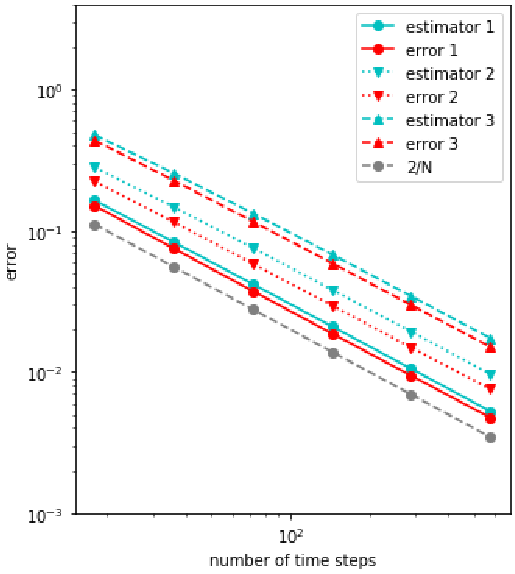
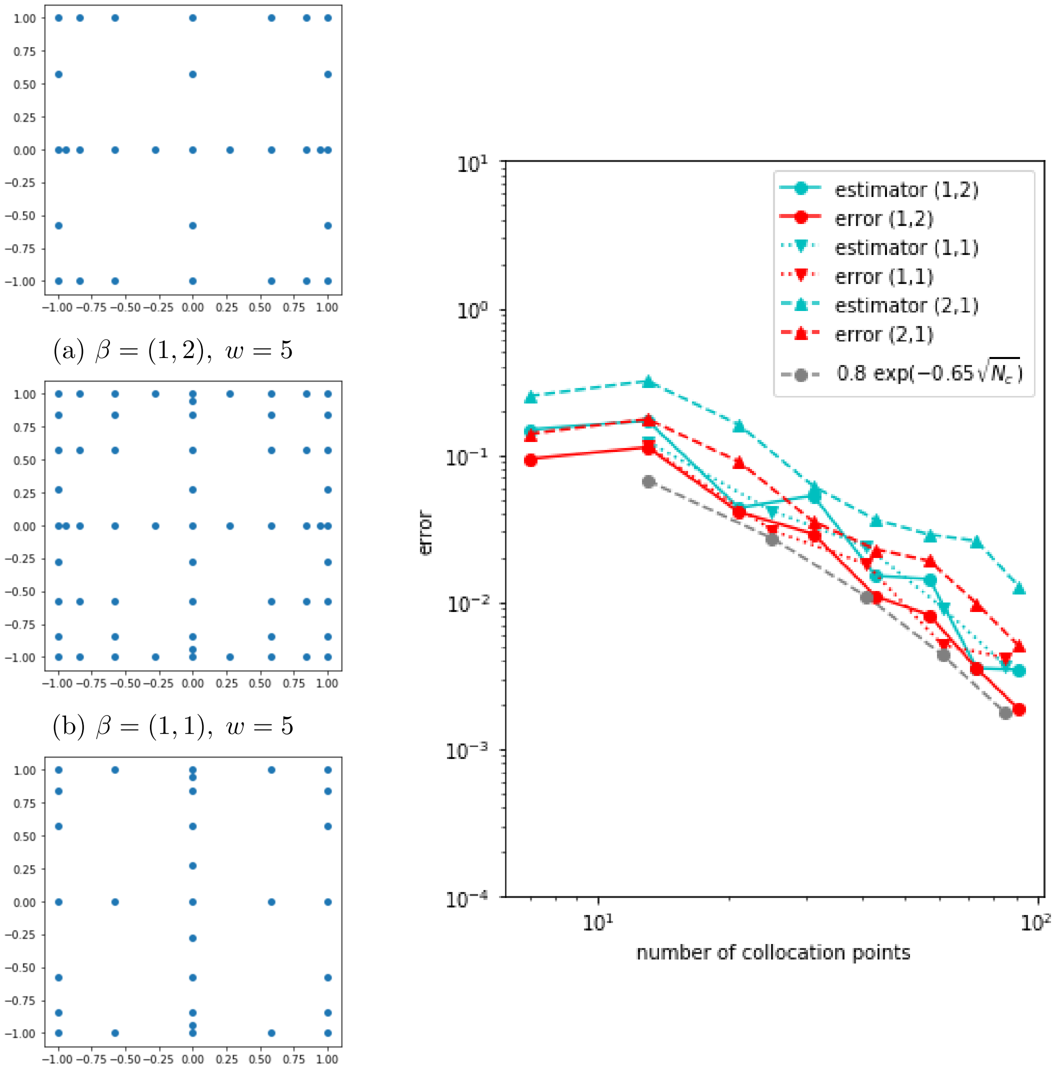


Figure 5.1 – Time error and estimator with respect to the number of time steps (right) for solutions computed on refinements of 3 time grids (left).

with $w = 1, \ldots, 8$ and $\beta = (\beta_1, \ldots, \beta_M)$, $\beta_m \geq 1$. The weights β were fixed to $(1, 2), (1, 1), (2, 1)$. Examples of such sparse grids with $w = 5$ can be seen in Figure 5.2 (left). The reference solution was computed with $w = 15$. Figure 5.2 (right) shows the error and the estimator for the 3 considered choices of sparse grids. We observe again that in all three cases we obtain a good estimation of the “true" error. Moreover, in all cases we observe a subexponential convergence which is consistent with sparse grid approximation results [NTW08b; NTT16].

Concerning the spatial estimator, we fixed the number of collocation points to 113 built as an isotropic sparse grid, the number of uniform time steps to 200 and computed the numerical solution on non uniform triangulations specified in Figure 5.3 (left). The convergence in Figure 5.3 was achieved by uniformly refining every cell, i.e. halving the diameter of every cell at each iteration of the convergence study with the use of refinement by longest edge bisection [BR14]. The Figure 5.3 shows that the estimators provide a good control of the “true" error in all three cases. We also plot the function $C/\sqrt{N_{FE}}$ where N_{FE} is the number of finite element cells. We see that in all three cases the error dacays as $O(\frac{1}{\sqrt{N_{FE}}})$ which corresponds to the theoretical order of convergence 1 with respect to the mesh size when using P1 finite elements on quasi-uniform meshes.

We now focus on the total error and consider several combinations of uniform refinements in the different components (spatial, temporal, stochastic). We report in Table 5.1 the behaviour of the estimator in all cases. From these results we conclude that the three components of the estimator behave in a fairly independent way. The only dependency

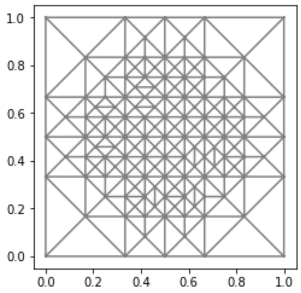


(a) $\beta = (1, 2),\ w = 5$

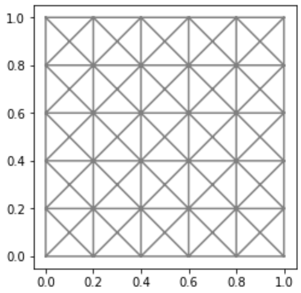
(b) $\beta = (1, 1),\ w = 5$

(c) $\beta = (2, 1),\ w = 5$

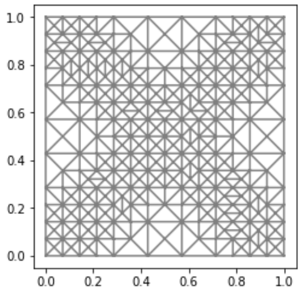
Figure 5.2 – Stochastic error and estimator with respect to the number of collocation points (right) for solutions computed on anisotropic sparse grids (left) of levels $w = 1, \ldots, 8$.



(a) Triangulation 1



(b) Triangulation 2



(c) Triangulation 3

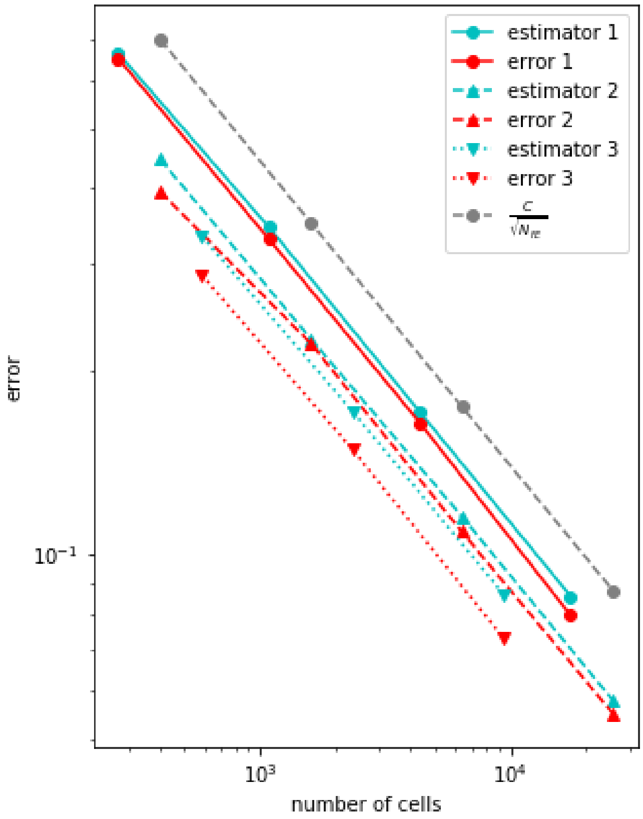


Figure 5.3 – Spatial error and estimator with respect to the number of cells (right) for solutions computed on refinements of 3 triangulations (left).

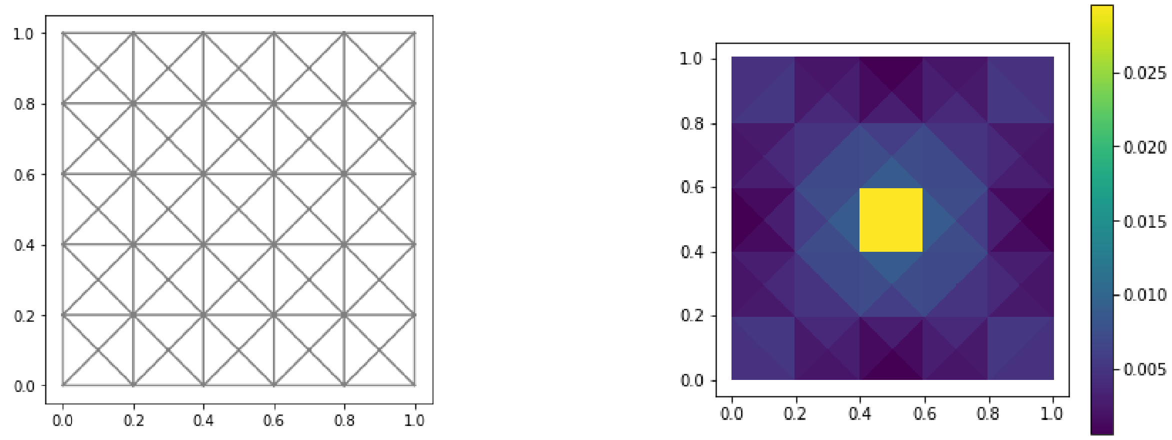
Table 5.1 – Error and error estimation when using different combinations of uniform triangulations, uniform time steps and isotropic sparse grids

h	τ	coll. pts	ϵ_{spa}	ϵ_{tem}	ϵ_{sto}	$\varepsilon_{\mathcal{T}_h,\tau,I}$	ε
0.2	0.025	13	0.73	0.13	0.017	0.74	0.66
0.1	0.025	13	0.38	0.13	0.1	0.42	0.39
0.04	0.025	13	0.16	0.13	0.18	0.27	0.2
0.02	0.025	13	0.08	0.13	0.2	0.25	0.16
0.01	0.025	13	0.04	0.13	0.2	0.24	0.157
0.02	0.05	13	0.08	0.26	0.2	0.33	0.22
0.02	0.025	13	0.08	0.13	0.2	0.25	0.16
0.02	0.0125	13	0.081	0.068	0.2	0.22	0.15
0.02	0.025	5	0.08	0.13	0.34	0.37	0.18
0.02	0.025	13	0.08	0.13	0.2	0.25	0.16
0.02	0.025	25	0.08	0.13	0.07	0.15	0.11

we can observe is the stochastic estimator being dependent on the spatial discretization. If the stochastic error is negligible compared to the spatial error, the stochastic estimator grows as the spatial estimator decreases while refining the spatial grid. When they reach a similar magnitude, decreasing the spatial error does not influence the stochastic estimator anymore. All the numerical solutions were computed on uniform triangulations, uniform time grids and isotropic sparse grids.

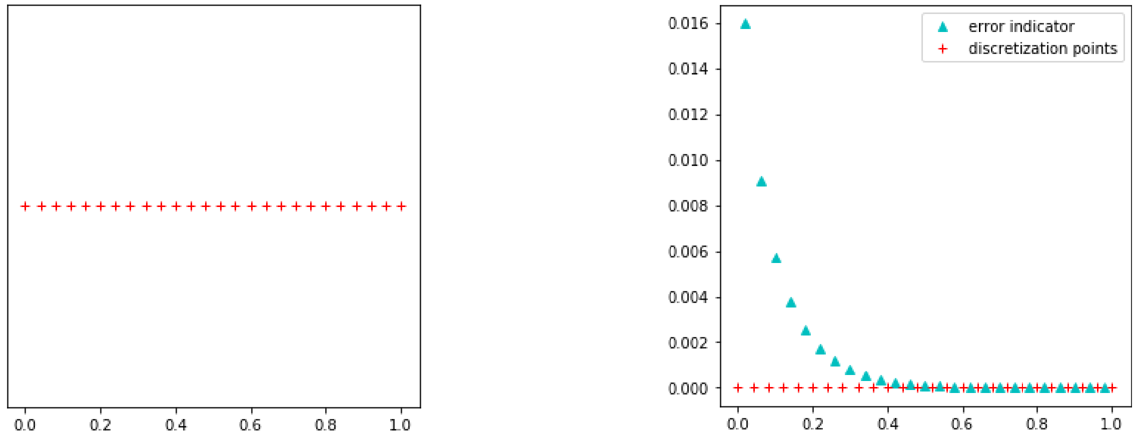
Numerical study of the performance of the adaptive algorithm

In this part we study the performance of the Algorithm 1 applied to problem (5.27). We set the tolerance to $TOL = 0.1$, $\alpha = 1.5$ and initialize the spatial grid as a uniform triangulation having 25 points and 100 triangles. The initial time discretization was set to have 25 equally spaced subintervals and the initial sparse grid was isotropic with 13 collocation points built over the index set $I = \{(1,1),(1,2),(1,3),(2,1),(2,2),(3,1)\}$. The initial discretizations are depicted in Figure 5.4 (left) with their corresponding error estimator (right). In Figure 5.5 (left) we show the final grids, the spatial triangulation having 7490 triangles, the time grid consisting of 155 steps and the stochastic sparse grid having 57 collocation points. As we can see, the algorithm was able to detect the location and discontinuity of the forcing term. The final time discretization is clearly consistent with the dissipative behaviour of this problem and the algorithm is also able to identify the dominant random variable Y_1. In Figure 5.5 (right) we can as well observe that the estimator provides a good control over the error throughout the whole process. As a last result we report in Table 5.2 the number of cells, time steps and collocation



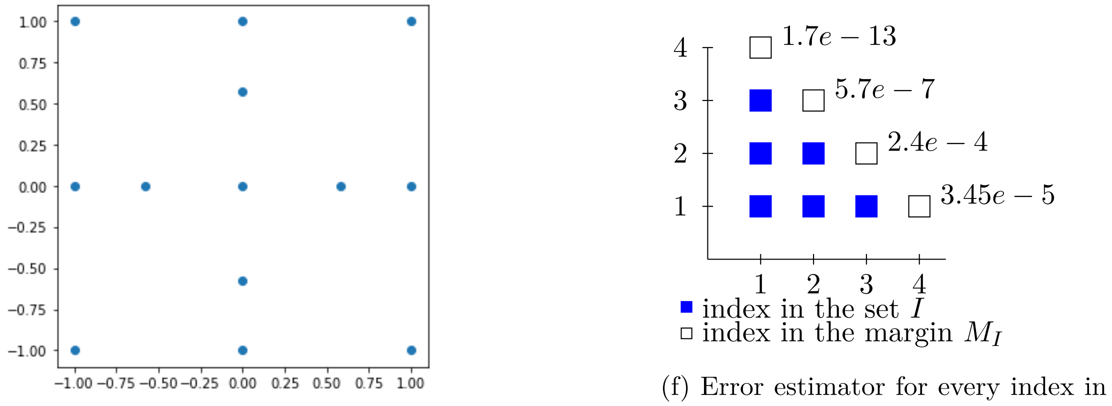
(a) Initial triangulation

(b) Error estimator for every triangle



(c) Initial time discretization

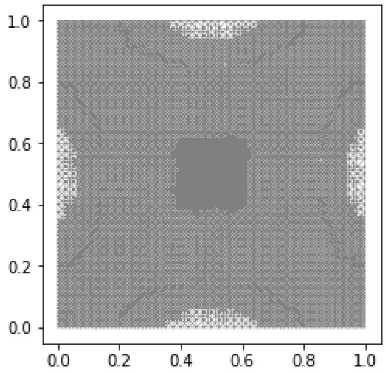
(d) Error estimator for every time step



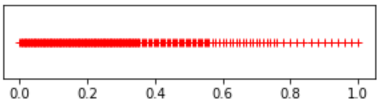
(e) Initial sparse grid

(f) Error estimator for every index in the margin

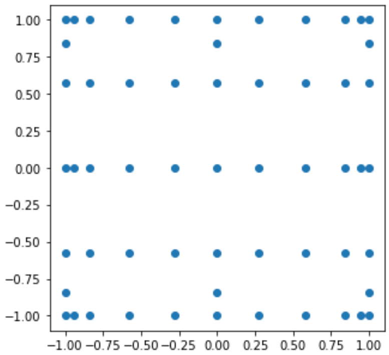
Figure 5.4 – Initial discretizations (left) when running the Algorithm 1 applied to problem (5.27) and corresponding error estimators (right) for elements, time intervals and multi-indices.



(a) Final triangulation



(b) Final time discretization



(c) Final sparse grid

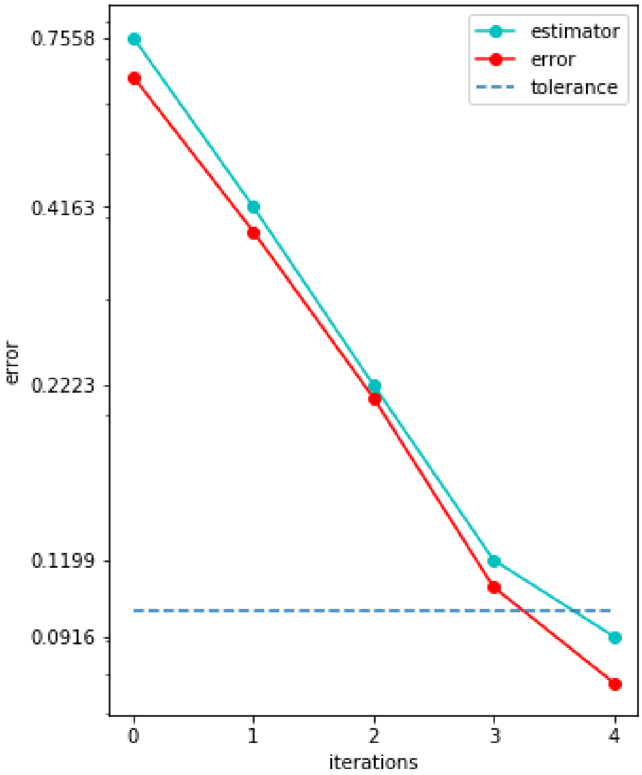


Figure 5.5 – Final discretizations (left) resulting from the Algorithm 1 applied to problem (5.27) and the evolution of the overall error and error estimation (right).

Table 5.2 – Number of cells, time steps and collocation points when using the Algorithm 1 with different tolerances.

TOL	no. of cells	no. of time steps	no. of coll. points	$\varepsilon_{\mathcal{T}_h,\tau,I}$
0.4	492	36	17	0.3676
0.2	1808	70	33	0.1927
0.1	7490	155	57	0.0947
0.05	29106	357	89	0.0487

points in the final discretizations for different tolerances. We can see that halving the tolerance results in approximately twice more time steps and four times more cells which agrees with the expected order of convergence.

5.2 A posteriori error estimation for a DLRA of a random parabolic equation

In this work we present a residual based a-posteriori error estimation for a DLR approximation of a random parabolic equation. The precise definition of the fully discrete DLR solution is available in Section 2.2.1. The derivation of the error applies the variational formulation (2.27), which requires the weights of the stochastic quadrature to be positive. Therefore, as opposed to the previous section, we consider here a stochastic discretization provided by tensor grids. The a-posteriori error estimate consists of four parts controlling the space discretization error, the time discretization error, the stochastic collocation error and the DLR error. These estimators can be used to drive an adaptive choice of spatial, stochastic, time discretization parameters and rank in the DLR approximation. This section starts with describing the governing problem and discretization techniques in Section 5.2.1. We follow by deriving a heuristic a-posteriori error estimation for a general random parabolic equation in Section 5.2.3. In Section 5.3 we derive a rigorous a-posteriori error estimation for the same random heat equation discussed in the previous section, with a diffusion coefficient affine w.r.t. the random variables. Finally, in Section 5.3.1 we propose an adaptive algorithm, where the derived estimators are used to drive an adaptive choice of time, spatial, stochastic discretization and rank R for the DLR approximation.

5.2.1 Discretization aspects

In this work we consider a DLR approximation of random parabolic equation, as described in Section 3.1. Furthermore, let $D \subset \mathbb{R}^d,\ 1 \leq d \leq 3$ be a polygonal domain with Lipschitz boundary. We assume that $V = H_0^1(D) =: H_0^1$, $H = L^2(D) =: L^2$, $V' = H^{-1}(D) =: H^{-1}$

and the scalar products $\langle v, w\rangle_{H,L^2_\rho}$, $\langle v, w\rangle_{V,L^2_\rho}$ are defined as

$$\begin{aligned}\langle v, w\rangle_{H,L^2_\rho} &= \int_\Omega \int_D v\, w \,\mathrm{d}x \,\mathrm{d}\rho\\ \langle v, w\rangle_{V,L^2_\rho} &= \int_\Omega \int_D \nabla v \cdot \nabla w \,\mathrm{d}x \,\mathrm{d}\rho.\end{aligned}$$

The discretization scheme is specified in Chapter 2. In particular, we apply the finite element method for the spatial discretization. We consider a triangulation $\mathcal{T}_h$ of D which satisfies $\bigcup_{K\in\mathcal{T}_h} K = D$ and a corresponding conforming finite element space V_h, which consists of continuous functions that are piecewise polynomials of degree $\leq r,\ r \geq 1$. For a rigorous estimation, we require the affine equivalence, admissibility and shape regularity conditions from Section 5.1.2 to be satisfied. Note that for the sake of simplicity, we assume that the spatial and stochastic discretizations, as well as the rank of the DLR approximation, do not change in time.

The stochastic discretization is performed by the stochastic collocation method ([XH05a; BNT10]), in particular the tensor grid method with Gaussian quadrature points. Let us assume that $\Omega \subset \mathbb{R}^M$ is a product of intervals Ω_m and that ρ factorizes as $\rho(\omega) = \Pi_{m=1}^M \rho_m(\omega_m),\ \forall\omega \in \Omega$. For each dimension $m = 1, \ldots, M$, let $\omega_{m,k_m},\ k_m = 1, \ldots, p_m + 1$ be the $p_m + 1$ roots of the polynomial q_{p_m+1} of degree p_m that is orthogonal w.r.t. the weight ρ_m, i.e.

$$\int_{\Omega_m} q_{p_m+1}(\omega)v(\omega)\rho_m(\omega)\,\mathrm{d}\omega = 0 \quad \forall v \in \mathbb{P}_{p_m}(\Omega_m),$$

where $\mathbb{P}_{p_m}(\Omega_m) = \mathrm{span}(\omega_m^j,\ j = 1, \ldots, p_m)$ consists of univariate polynomials of order at most p_m. By $\mathbb{P}_p(\Omega)$ we denote the span of tensor product polynomials with degree at most $p = (p_1, \ldots, p_m)$, i.e.

$$\mathbb{P}_p(\Omega) = \bigotimes_{m=1}^M \mathbb{P}_{p_m}(\Omega_m).$$

To any vector of indices $[k_1, \ldots, k_M]$, we associate the global index $k = k_1 + p_1(k_2 - 1) + p_1p_2(k_3 - 1) + \ldots$ and we denote by ω_k the point $\omega_k = [\omega_{1,k_1}, \omega_{2,k_2}, \ldots, \omega_{M,k_M}] \in \Omega$. We also introduce, for each $m = 1, 2, \ldots, M$, the Lagrangian basis $\{l_{m,j}\}_{j=1}^{p_m+1}$ of the space $\mathbb{P}_{p_m}$,

$$l_{m,j} \in \mathbb{P}_{p_m(\Omega_m)}, \quad l_{m,j}(\omega_{m,k}) = \delta_{j,k}, \quad j, k = 1, \ldots, p_m + 1$$

and we set $l_k(\omega) = \Pi_{m=1}^M l_{m,k_m}(\omega_m)$. The weights are obtained as

$$\lambda_{k_m} = \int_{\Omega_m} l_{k_m}^2(\omega)\rho_m(\omega)\,\mathrm{d}\omega, \quad \lambda_k = \Pi_{m=1}^M \lambda_{k_m} > 0.$$

The expectation $\mathbb{E}_\rho[g]$ of a function $g : \Omega \to V$ will be approximated by the Gauss quadrature formula

$$\mathbb{E}_{\hat{\rho}}[g] = \sum_{k=1}^{\hat{N}} \lambda_k g(\omega_k). \tag{5.28}$$

Note that $\hat{N} = \Pi_{m=1}^{M}(p_m + 1)$. We introduce the Lagrange interpolant operator $\mathcal{I}_p : C^0(\Omega; V) \to \mathbb{P}_p(\Omega)$ defined as

$$\mathcal{I}_p g(\omega) = \sum_{k=1}^{\hat{N}} g(\omega_k) l_k(\omega) \qquad \forall g \in C^0(\Omega; V).$$

5.2.2 Fully discrete problem

As described in Chapter 2, the fully discrete DLR solution $u^n_{h,\hat{\rho}} = ((u^n_{h,\hat{\rho}})_{(1)}, \dots, (u^n_{h,\hat{\rho}})_{(\hat{N})})$ satisfies the following equation, weakly in V_h

$$\frac{(u^{n+1}_{h,\hat{\rho}})_{(k)} - (u^n_{h,\hat{\rho}})_{(k)}}{\triangle t^n} + \left(\Pi^{h,\hat{\rho}}_{\tilde{U}^{n+1} Y^{n\intercal}}[\mathcal{L}(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}})]\right)(\omega_k) = \left(\Pi^{h,\hat{\rho}}_{\tilde{U}^{n+1} Y^{n\intercal}}[f^{n,n+1}]\right)(\omega_k), \tag{5.29}$$

where

$$\Pi^{h,\hat{\rho}}_{\tilde{U}^{n+1} Y^{n\intercal}}[g] = \mathbb{E}_{\hat{\rho}}[gY^n]Y^n + \Big((g, \tilde{U}^{n+1})_{V'V} - \mathbb{E}_{\hat{\rho}}\Big[(g, \tilde{U}^{n+1})_{V'V} Y^n\Big]Y^n\Big)(\tilde{M}^{n+1})^{-1}\tilde{U}^{n+1}, \quad g \in L^2_\rho(\Omega; V').$$

Note that in this work we allow for different time steps at different times, noting $\triangle t^n = t^{n+1} - t^n$. We proceed by redefining the DLR solution $u^n_{h,\hat{\rho}}$ and consider its interpolated version defined as

$$u^n_{h,\hat{\rho}}(\omega) = \sum_{k=1}^{\hat{N}} (u^n_{h,\hat{\rho}})_{(k)} l_k(\omega), \qquad \forall \omega \in \Omega.$$

Note that $u^n_{h,\hat{\rho}} \in V_h \otimes \mathbb{P}_p(\Omega)$, $\forall n = 0, \dots, N$. The projection on the tangent space $\Pi^{h,\hat{\rho}}_{\tilde{U}^{n+1} Y^{n\intercal}}$ involves computing the mean value $\mathbb{E}_\rho[\mathcal{L}(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}})Y^n]$, which is obtained as

$$\begin{aligned}\mathbb{E}_\rho[\mathcal{L}(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}})Y^n] \approx \mathbb{E}_{\hat{\rho}}[\mathcal{L}(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}})Y^n] &= \sum_{k=1}^{\hat{N}} \lambda_k \mathcal{L}(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}})(\omega_k) Y^n(\omega_k) \\ &= \mathbb{E}_{\hat{\rho}}[\mathcal{I}_p[\mathcal{L}(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}})]Y^n] = \mathbb{E}_\rho[\mathcal{I}_p[\mathcal{L}(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}})]Y^n],\end{aligned}$$

where in the last inequality we used the fact that the Gauss quadrature is exact for polynomials up to degree $2p_m + 1$. The integrand $\mathcal{I}_p[\mathcal{L}(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}})]Y^n$ is, in each variable, a univariate polynomial of degree $2p_m$, $m = 1, \dots, M$. The equation (5.29) can be therefore rewritten as

$$\frac{u_{h,\hat\rho}^{n+1}(\omega_k) - u_{h,\hat\rho}^{n}(\omega_k)}{\triangle t} + \Pi^h_{\tilde U^{n+1}Y^{n\intercal}}\Big[\mathcal{I}_p[\mathcal{L}(u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho})]\Big](\omega_k) = \Pi^h_{\tilde U^{n+1}Y^{n\intercal}}\Big[\mathcal{I}_p[f^{n,n+1}]\Big](\omega_k), \quad \forall k = 1, \dots, \hat N, \quad (5.30)$$

with the projection $\Pi^h_{\tilde U^{n+1}Y^{n\intercal}}$ involving computing the expectations exactly, i.e.

$$\Pi^h_{\tilde U^{n+1}Y^{n\intercal}}[g] = \mathbb{E}_\rho[gY^n]Y^n + \Big((g, \tilde U^{n+1})_{V'V} - \mathbb{E}_\rho[(g, \tilde U^{n+1})_{V'V}Y^n]Y^n\Big)(\tilde M^{n+1})^{-1}\tilde U^{n+1}, \quad g \in L^2_\rho(\Omega; V').$$

Multiplying (5.30) by $l_k(\omega)$ and summing over $k = 1, \dots, \hat N$, we see that the fully discrete DLR solution satisfies the following variational formulation, for any $v_h \in L^2_\rho(\Omega; V_h)$

$$\Big\langle \frac{u^{n+1}_{h,\hat\rho} - u^n_{h,\hat\rho}}{\triangle t^n}, v_h \Big\rangle_{H, L^2_\rho} + \Big(\Pi^h_{\tilde U^{n+1}Y^{n\intercal}}\mathcal{I}_p\mathcal{L}(u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho}), v_h\Big)_{V'V, L^2_\rho} = \Big\langle \Pi^h_{\tilde U^{n+1}Y^{n\intercal}}\mathcal{I}_p f^{n,n+1}, v_h\Big\rangle_{H, L^2_\rho}, \quad \forall v_h \in L^2_\rho(\Omega; V_h). \quad (5.31)$$

Based on the sequence of solutions $\{u^n_{h,\hat\rho}\}_{n=0}^N \subset L^2_\rho(\Omega; V_h)$, we build a piecewise affine function $\tilde u \in L^2(0, T; L^2_\rho(\Omega; V_h))$ on $[0, T]$, which equals $u^n_{h,\hat\rho}$ at times t^n, $n = 0, \dots, N$, i.e.

$$\tilde u(t) = \frac{t^{n+1} - t}{\triangle t^n} u^n_{h,\hat\rho} + \frac{t - t^n}{\triangle t^n} u^{n+1}_{h,\hat\rho}, \quad t \in [t^n, t^{n+1}]. \quad (5.32)$$

Note that

$$\dot{\tilde u} = \frac{u^{n+1}_{h,\hat\rho} - u^n_{h,\hat\rho}}{\triangle t^n} \text{ a.e. on } (t^n, t^{n+1}].$$

5.2.3 Residual based a-posteriori error estimation for a general random parabolic equation

The objective of our work is to derive a residual based a-posteriori error estimation for a DLR approximation of a random parabolic equation. First, we will start by deriving an error estimation for a general elliptic operator $\mathcal{L}$ and a discretization $\mathcal{L}(u^n_{h,\hat\rho}, u^{n+1}_{h,\hat\rho})$, which covers all explicit, implicit and semi-implicit scheme. Later, we will follow by specifying all error estimates for all three schemes, for the case of a random heat equation in Theorem 5.3.1.

In what follows, all equations hold a.e. in (t^n, t^{n+1}). For any $v \in L^2_\rho(\Omega; V)$ we have

$$\Big\langle \dot u_{\text{true}} - \dot{\tilde u}, v\Big\rangle_{H, L^2_\rho} + \Big(\mathcal{L}(u_{\text{true}} - \tilde u), v\Big)_{V'V, L^2_\rho} = \Big\langle f, v\Big\rangle_{H, L^2_\rho} - \Big\langle \dot{\tilde u}, v\Big\rangle_{H, L^2_\rho} - \Big(\mathcal{L}(\tilde u), v\Big)_{V'V, L^2_\rho}$$

$$
\begin{align*}
&= \underbrace{\Big\langle f - \mathcal{I}_p f, v \Big\rangle_{H,L^2_\rho} - \Big(\mathcal{L}(\tilde{u}) - \mathcal{I}_p\mathcal{L}(\tilde{u}), v\Big)_{V'V,L^2_\rho}}_{=:A_{sto}} \\
&\qquad + \Big\langle \mathcal{I}_p f, v \Big\rangle_{H,L^2_\rho} - \Big\langle \dot{\tilde{u}}, v \Big\rangle_{H,L^2_\rho} - \Big(\mathcal{I}_p\mathcal{L}(\tilde{u}), v\Big)_{V'V,L^2_\rho} \\
&= A_{sto} + \underbrace{\Big\langle \mathcal{I}_p f, v - v_h \Big\rangle_{H,L^2_\rho} - \Big\langle \dot{\tilde{u}}, v - v_h \Big\rangle_{H,L^2_\rho} - \Big(\mathcal{I}_p\mathcal{L}(\tilde{u}), v - v_h\Big)_{V'V,L^2_{\hat{\rho}}}}_{=:A_{spa}} \\
&\qquad + \Big\langle \mathcal{I}_p f, v_h \Big\rangle_{H,L^2_{\hat{\rho}}} - \Big\langle \dot{\tilde{u}}, v_h \Big\rangle_{H,L^2_\rho} - \Big(\mathcal{I}_p\mathcal{L}(\tilde{u}), v_h\Big)_{V'V,L^2_\rho} \\
&= A_{sto} + A_{spa} + \Big\langle \mathcal{I}_p f, v_h \Big\rangle_{H,L^2_\rho} - \Big\langle \dot{\tilde{u}}, v_h \Big\rangle_{H,L^2_\rho} - \Big(\mathcal{I}_p\mathcal{L}(\tilde{u}), v_h\Big)_{V'V,L^2_\rho} \\
&= A_{sto} + A_{spa} + \underbrace{\Big\langle \mathcal{I}_p f - \mathcal{I}_p f^{n,n+1}, v_h \Big\rangle_{H,L^2_\rho} - \Big(\mathcal{I}_p\mathcal{L}(\tilde{u}) - \mathcal{I}_p\mathcal{L}(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}}), v_h\Big)_{V'V,L^2_\rho}}_{=:A_{tem,1}} \\
&\qquad + \Big\langle \mathcal{I}_p f^{n,n+1}, v_h \Big\rangle_{H,L^2_\rho} - \Big\langle \frac{u^{n+1}_{h,\hat{\rho}} - u^n_{h,\hat{\rho}}}{\triangle t^n}, v_h \Big\rangle_{H,L^2_\rho} - \Big(\mathcal{I}_p\mathcal{L}(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}}), v_h\Big)_{V'V,L^2_\rho} \\
&= \Big\langle \Pi^h_{\tilde{U}^{n+1}Y^{n\intercal}}[\mathcal{I}_p f^{n,n+1}], v_h \Big\rangle_{H,L^2_\rho} - \Big\langle \frac{u^{n+1}_{h,\hat{\rho}} - u^n_{h,\hat{\rho}}}{\triangle t^n}, v_h \Big\rangle_{H,L^2_\rho} \tag{5.33} \\
&\qquad - \Big(\Pi^h_{\tilde{U}^{n+1}Y^{n\intercal}}[\mathcal{I}_p\mathcal{L}(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}})], v_h\Big)_{V'V,L^2_\rho} \\
&\qquad + A_{sto} + A_{spa} + A_{tem,1} + \Big({\Pi^h_{\tilde{U}^{n+1}Y^{n\intercal}}}^{\perp}[\mathcal{I}_p f^{n,n+1^*} - \mathcal{I}_p\mathcal{L}^*(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}})], v_h\Big)_{V'V,L^2_\rho} \\
&= A_{sto} + A_{spa} + A_{tem,1} + \underbrace{\Big({\Pi^h_{\tilde{U}^{n+1}Y^{n\intercal}}}^{\perp}[\mathcal{I}_p f^{n,n+1^*} - \mathcal{I}_p\mathcal{L}^*(\tilde{U}^{n+1}Y^{n^\intercal}), v_h\Big)_{V'V,L^2_\rho}}_{=:A_{rank}} \\
&\qquad + \underbrace{\Big({\Pi^h_{\tilde{U}^{n+1}Y^{n\intercal}}}^{\perp}[\mathcal{I}_p\mathcal{L}^*(\tilde{U}^{n+1}Y^{n^\intercal}) - \mathcal{I}_p\mathcal{L}^*(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}})], v_h\Big)_{V'V,L^2_\rho}}_{=:A_{tem,2}}.
\end{align*}
$$

In the last step, we applied equation (5.31) and split the remaining term into 2 terms, of which A_{rank} contributes to the rank-truncation error and $A_{tem,2}$ contributes to the time discretization error.

Concerning the term $A_{sto}(t)$, we can bound it as

$$
\begin{aligned}
|A_{sto}(t)| &\leq \|f - \mathcal{I}_p f\|_{H,L^2_\rho}\|v\|_{H,L^2_\rho} + \|\mathcal{L}(\tilde{u}) - \mathcal{I}_p\mathcal{L}(\tilde{u})\|_{V',L^2_\rho}\|v\|_{V,L^2_\rho} \\
&\leq C_{\mathrm{P}}\|f - \mathcal{I}_p f\|_{H,L^2_\rho}\|v\|_{V,L^2_\rho} + \|\mathcal{L}(\tilde{u}) - \mathcal{I}_p\mathcal{L}(\tilde{u})\|_{V',L^2_\rho}\|v\|_{V,L^2_\rho} \\
&\leq \underbrace{\frac{4C_{\mathrm{P}}^2}{C_{\mathcal{L}}}\|f - \mathcal{I}_p f\|^2_{H,L^2_\rho} + \frac{4}{C_{\mathcal{L}}}\|\mathcal{L}(\tilde{u}) - \mathcal{I}_p\mathcal{L}(\tilde{u})\|^2_{V',L^2_\rho}}_{\mathscr{E}_{sto}(t)} + \frac{C_{\mathcal{L}}}{8}\|v\|^2_{V,L^2_\rho}.
\end{aligned}
$$

As for the temporal terms $A_{tem,1}(t), A_{tem,2}(t)$, we can bound them as

$$
\begin{aligned}
|A_{tem,1}(t)| &\leq C_{\mathrm{P}}\|\mathcal{I}_p f - \mathcal{I}_p f^{n,n+1}\|_{H,L^2_\rho}\|v_h\|_{V,L^2_\rho} + \|\mathcal{I}_p\mathcal{L}(\tilde{u}) - \mathcal{I}_p\mathcal{L}(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}})\|_{V',L^2_\rho}\|v_h\|_{V,L^2_\rho} \\
&\leq \underbrace{\frac{4C_{\mathrm{P}}^2C_1^2}{C_{\mathcal{L}}}\|\mathcal{I}_p f - \mathcal{I}_p f^{n,n+1}\|^2_{H,L^2_\rho} + \frac{4C_1^2}{C_{\mathcal{L}}}\|\mathcal{I}_p\mathcal{L}(\tilde{u}) - \mathcal{I}_p\mathcal{L}(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}})\|^2_{V',L^2_\rho}}_{\mathscr{E}_{tem,1}(t)} + \frac{C_{\mathcal{L}}}{8}\|v\|^2_{V,L^2_\rho}.
\end{aligned}
$$

$$
\begin{aligned}
|A_{tem,2}(t)| &\leq \|{\Pi^{h,\hat{\rho}}_{\tilde{U}^{n+1}Y^{n\mathsf{T}}}}^{\perp}[\mathcal{I}_p\mathcal{L}(\tilde{U}^{n+1}Y^{n^{\mathsf{T}}}) - \mathcal{I}_p\mathcal{L}(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}})]\|_{V',L^2_\rho}\|v_h\|_{V,L^2_\rho} \\
&\leq \underbrace{\frac{2C_1^2}{C_{\mathcal{L}}}\|{\Pi^{h,\hat{\rho}}_{\tilde{U}^{n+1}Y^{n\mathsf{T}}}}^{\perp}[\mathcal{I}_p\mathcal{L}(\tilde{U}^{n+1}Y^{n^{\mathsf{T}}}) - \mathcal{I}_p\mathcal{L}(u^n_{h,\hat{\rho}}, u^{n+1}_{h,\hat{\rho}})]\|^2_{V',L^2_\rho}}_{\mathscr{E}_{tem,2}(t)} + \frac{C_{\mathcal{L}}}{8}\|v\|^2_{V,L^2_\rho}
\end{aligned}
$$

The spatial contribution will be further specified for the random heat equation example. For now, let us just assume that we can bound A_{spa} as

$$
|A_{spa}(t)| = \left|\left\langle \mathcal{I}_p f - \dot{\tilde{u}} - \mathcal{I}_p\mathcal{L}(\tilde{u}), v - v_h \right)_{V'V,L^2_\rho}\right| \leq \mathscr{E}_{spa}(t) + \frac{C_{\mathcal{L}}}{8}\|v\|^2_{V,L^2_\rho}.
$$

And lastly, the rank-truncation contribution will be bounded by

$$
\begin{aligned}
|A_{rank}(t)| &= \left|\left({\Pi^{h,\hat{\rho}}_{\tilde{U}^{n+1}Y^{n\mathsf{T}}}}^{\perp}[\mathcal{I}_p f^{n,n+1^*} - \mathcal{I}_p\mathcal{L}^*(\tilde{U}^{n+1}Y^{n^{\mathsf{T}}}), v_h\right)_{V'V,L^2_\rho}\right| \\
&\leq \left\|{\Pi^{h,\hat{\rho}}_{\tilde{U}^{n+1}Y^{n\mathsf{T}}}}^{\perp}[\mathcal{I}_p f^{n,n+1^*} - \mathcal{I}_p\mathcal{L}^*(\tilde{U}^{n+1}Y^{n^{\mathsf{T}}})\right\|_{V',L^2_\rho} C_1\|v\|_{V,L^2_\rho} \\
&\leq \underbrace{\frac{2C_1^2}{C_{\mathcal{L}}}\left\|{\Pi^{h,\hat{\rho}}_{\tilde{U}^{n+1}Y^{n\mathsf{T}}}}^{\perp}[\mathcal{I}_p f^{n,n+1^*} - \mathcal{I}_p\mathcal{L}^*(\tilde{U}^{n+1}Y^{n^{\mathsf{T}}})\right\|^2_{V',L^2_\rho}}_{\mathscr{E}_{rank}(t)} + \frac{C_{\mathcal{L}}}{8}\|v\|^2_{V,L^2_\rho},
\end{aligned}
$$

where

$$
\mathscr{E}_{rank}(t) = \frac{2C_1^2}{C_{\mathcal{L}}}\left\|\mathcal{P}^{\perp}_{\mathcal{Y}^n}\left[\mathcal{P}^{\perp}_{\tilde{\mathcal{U}}^{n+1}}\left[\mathcal{I}_p f^{n,n+1^*} - \mathcal{I}_p\mathcal{L}^*(\tilde{U}^{n+1}Y^{n^{\mathsf{T}}})\right]\right]\right\|^2_{V',L^2_\rho}.
$$

Based on our previous computations, we obtain

$$
\begin{aligned}
&\left\langle \dot{u}_{\mathrm{true}} - \dot{\tilde{u}}, v\right\rangle_{H,L^2_\rho} + \left(\mathcal{L}(u_{\mathrm{true}} - \tilde{u}), v\right)_{V'V,L^2_\rho} \\
&\qquad\qquad \leq \mathscr{E}_{sto} + \mathscr{E}_{spa} + \mathscr{E}_{tem,1} + \mathscr{E}_{tem,2} + \mathscr{E}_{rank} + \frac{5C_{\mathcal{L}}}{8}\|v\|^2_{V,L^2_\rho}.
\end{aligned}
$$

Now, taking $v = u_{\mathrm{true}} - \tilde{u} \in L^2_\rho(\Omega; V)$ we derive

$$
\begin{aligned}
&\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u_{\mathrm{true}} - \tilde{u}\|^2_{H,L^2_\rho}(t) + C_{\mathcal{L}}\|u_{\mathrm{true}} - \tilde{u}\|^2_{V,L^2_\rho}(t) \\
&\qquad \leq \mathscr{E}_{sto}(t) + \mathscr{E}_{spa}(t) + \mathscr{E}_{tem,1}(t) + \mathscr{E}_{tem,2}(t) + \mathscr{E}_{rank}(t) + \frac{5C_{\mathcal{L}}}{8}\|u_{\mathrm{true}} - \tilde{u}\|^2_{V,L^2_\rho}(t)
\end{aligned}
$$

which results in

$$\frac{\mathrm{d}}{\mathrm{d}t}\|u_{\text{true}} - \tilde{u}\|^2_{H,L^2_\rho}(t) + \frac{3C_{\mathcal{L}}}{4}\|u_{\text{true}} - \tilde{u}\|^2_{V,L^2_\rho}(t) \\ \leq 2(\mathscr{E}_{sto}(t) + \mathscr{E}_{spa}(t) + \mathscr{E}_{tem,1}(t) + \mathscr{E}_{tem,2}(t) + \mathscr{E}_{rank}(t)).$$

The last step is to integrate this inequality w.r.t. t over $(0,T)$. In conclusion, we obtain

$$\|u_{\text{true}}(T) - \tilde{u}(T)\|^2_{H,L^2_\rho} + \frac{3C_{\mathcal{L}}}{4}\|u_{\text{true}} - \tilde{u}\|^2_{L^2(0,T;L^2_\rho(\Omega;V))} \leq \|u_{\text{true}}(0) - \tilde{u}(0)\|^2_{H,L^2_\rho} \\ + 2\int_0^T \Big(\mathscr{E}_{sto}(t) + \mathscr{E}_{spa}(t) + \mathscr{E}_{tem,1}(t) + \mathscr{E}_{tem,2}(t) + \mathscr{E}_{rank}(t)\Big)\,\mathrm{d}t. \quad (5.34)$$

5.3 A posteriori error estimation for a DLRA of a random heat equation

In this section, we detail all error estimators for the case of a random heat equation, described in Section 3.5. In addition, let us assume that the diffusion coefficient a is affine w.r.t. $\omega = (\omega^1, \dots, \omega^M) \in \Omega$, i.e.

$$a(x,\omega) = \bar{a}(x) + \sum_{m=1}^{M} a_m(x)\omega^m, \quad \forall x \in D, \omega \in \Omega, \quad (5.35)$$

and that there exist $a_{\min}$, $a_{\max} \in \mathbb{R}$ s.t.

$$\rho(\omega \in \Omega:\ 0 < a_{\min} \leq a(x,\omega) \leq a_{\max} < \infty \quad \forall x \in D) = 1. \quad (5.36)$$

By $a_{stoch}(x,\omega)$ we will denote the stochastic part of the diffusion coefficient, i.e. $a_{stoch}(x,\omega) := \sum_{m=1}^{M} a_m(x)\omega^m$.

For any element, face or edge S, h_S denotes its diameter. With every edge $(d = 2)$ or face $(d = 3)$ E, we identify a unit vector η_E orthogonal to it and denote the jump across E in direction η_E by $[\cdot]_E$.

Theorem 5.3.1. *Consider a random heat equation with a diffusion coefficient affine w.r.t. random variables satisfying* (5.36)*. Let u_{true} be a solution of* (5.4) *and $\tilde{u}$ be defined as in* (5.32)*. Then there exists a constant $C > 0$, independent of the time step, mesh size, tensor grid choice and DLR rank, such that*

$$\|u_{\text{true}}(T) - \tilde{u}(T)\|^2_{H,L^2_\rho} + \frac{3}{4}a_{\min}\|u_{\text{true}} - \tilde{u}\|^2_{L^2(0,T;L^2_\rho(\hat{\Omega};V))} \leq \|u_{\text{true}}(0) - \tilde{u}(0)\|^2_{H,L^2_\rho} \\ + \varepsilon_{sto} + \varepsilon_{spa} + \varepsilon_{tem} + \varepsilon_{rank}, \quad (5.37)$$

where

$$
\begin{aligned}
\varepsilon_{sto} &= C\|f - \mathcal{I}_p f\|^2_{L^2(0,T;L^2_\rho(\Omega;H))} \\
&\quad + C\sum_{n=0}^{N-1} \triangle t^n \left(\left\| a\nabla u^{n+1}_{h,\hat\rho} - \mathcal{I}_p[a\nabla u^{n+1}_{h,\hat\rho}] \right\|^2_{H,L^2_\rho} + \left\| a\nabla u^{n}_{h,\hat\rho} - \mathcal{I}_p[a\nabla u^{n}_{h,\hat\rho}] \right\|^2_{H,L^2_\rho} \right), \qquad (5.38)
\end{aligned}
$$

$$
\begin{aligned}
\varepsilon_{spa} = C\sum_{k=1}^{\hat N} \lambda_k \Big(\sum_{K\in\mathscr{T}_h} h_K^2 \left\| f(\omega_k) - \dot{\tilde u}(\omega_k) + \nabla\cdot(a(\omega_k)\nabla\tilde u(\omega_k)) \right\|^2_{L^2(0,T;L^2(K))} \\
+ \sum_{E\in\mathscr{E}_h} h_E \left\| [a(\omega_k)\nabla\tilde u(\omega_k)\cdot\eta_E]_E \right\|^2_{L^2(0,T;L^2(E))} \Big), \qquad (5.39)
\end{aligned}
$$

$$
\begin{aligned}
\varepsilon_{rank} &= C\sum_{n=0}^{N-1} \triangle t^n \sum_{k=1}^{\hat N} \lambda_k \left\| {\Pi^{h,\hat\rho}_{\tilde U^{n+1}Y^{n\intercal}}}^{\perp} [f^{n+1^*}](\omega_k) \right\|^2_H \\
&\quad + C\sum_{n=0}^{N-1} \triangle t^n \sum_{k=1}^{\hat N} \lambda_k \Big(\sum_{K\in\mathcal{T}_h} \left\| {\Pi^{h,\hat\rho}_{\tilde U^{n+1}Y^{n\intercal}}}^{\perp} [(-\nabla\cdot(a\nabla\tilde U^{n+1})Y^{n\intercal})^*](\omega_k) \right\|^2_{L^2(K)} \\
&\qquad + \sum_{E\in\mathscr{E}_h} \left\| \mathcal{P}^{\hat\rho\perp}_{\mathcal{Y}^n} \Big[\Big([a\nabla\tilde U^{n+1}\cdot\eta_E]_E Y^{n\intercal} \Big)^* \Big] (\omega_k) \right\|^2_{L^2(E)} \\
&\quad + \|\tilde U^{n+1}\tilde M^{n+1^{-1}}\|^2_H \left\| \tilde U^{n+1\intercal} \mathcal{P}^{\hat\rho\perp}_{\mathcal{Y}^n} \Big[\Big([a\nabla\tilde U^{n+1}\cdot\eta_E]_E Y^{n\intercal}(\omega_k) \Big)^* \Big] \right\|^2_{L^2(E)}, \qquad (5.40)
\end{aligned}
$$

$$
\varepsilon_{tem} = \begin{cases} \varepsilon^{\mathrm{im}}_{tem} & \textit{for implicit scheme} \\ \varepsilon^{\mathrm{ex}}_{tem} & \textit{for explicit scheme} \\ \varepsilon^{\mathrm{semi}}_{tem} & \textit{for semi-implicit scheme,} \end{cases} \qquad (5.41)
$$

where

$$
\begin{aligned}
\varepsilon^{\mathrm{im}}_{tem} &= C\sum_{n=0}^{N-1}\sum_{k=1}^{\hat N} \lambda_k \|f(\omega_k) - f^{n+1}(\omega_k)\|^2_{L^2(t^n,t^{n+1};H)} \\
&\qquad + C\sum_{n=0}^{N-1}\sum_{k=1}^{\hat N} \lambda_k \frac{\triangle t^n}{3} \|a(\omega_k)\nabla\Big(u^{n+1}_{h,\hat\rho}(\omega_k) - u^n_{h,\hat\rho}(\omega_k)\Big)\|^2_H \\
&\qquad + C\sum_{n=0}^{N-1} \triangle t^n \sum_{k=1}^{\hat N} \lambda_k \Big(\|a(\omega_k)\nabla\tilde U^{n+1}(\tilde Y^{n+1} - Y^n)^\intercal(\omega_k)\|^2_H \\
&\qquad + \|a(\omega_k)\tilde M^{n+1^{-1}}\nabla\tilde U^{n+1\intercal}\nabla\tilde U^{n+1}(\tilde Y^{n+1} - Y^n)^\intercal(\omega_k)\|^2_H \|\tilde U^{n+1}\|^2_H \Big) \qquad (5.42)
\end{aligned}
$$

$$
\begin{aligned}
\varepsilon_{tem}^{\mathrm{ex}} = C \sum_{n=0}^{N-1} \sum_{k=1}^{\hat{N}} \lambda_k \| f(\omega_k) - f^n(\omega_k) \|^2_{L^2(t^n,t^{n+1};H)} \\
+ C \sum_{n=0}^{N-1} \sum_{k=1}^{\hat{N}} \lambda_k \frac{\triangle t^n}{3} \| a(\omega_k) \nabla \Big(u_{h,\hat{\rho}}^{n+1}(\omega_k) - u_{h,\hat{\rho}}^{n}(\omega_k) \Big) \|^2_H \\
+ C \sum_{n=0}^{N-1} \triangle t^n \sum_{k=1}^{\hat{N}} \lambda_k \Big(\| a(\omega_k) \nabla (\tilde{U}^{n+1} - U^n) Y^{n^\intercal}(\omega_k) \|^2_H \\
+ \| a(\omega_k) \tilde{M}^{n+1^{-1}} \nabla (\tilde{U}^{n+1} - U^n)^\intercal \nabla \tilde{U}^{n+1} Y^{n^\intercal}(\omega_k) \|^2_H \| \tilde{U}^{n+1} \|^2_H \Big) \qquad (5.43)
\end{aligned}
$$

$$
\begin{aligned}
\varepsilon_{tem}^{\mathrm{semi}} = C \sum_{n=0}^{N-1} \sum_{k=1}^{\hat{N}} \lambda_k \| f(\omega_k) - f^{n+1}(\omega_k) \|^2_{L^2(t^n,t^{n+1};H)} \\
+ C \sum_{n=0}^{N-1} \sum_{k=1}^{\hat{N}} \lambda_k \frac{\triangle t^n}{3} \Big(\| \bar{a} \nabla \Big(u_{h,\hat{\rho}}^{n+1}(\omega_k) - u_{h,\hat{\rho}}^{n}(\omega_k) \Big) \|^2_H \\
+ \| a_{stoch}(\omega_k) \nabla \Big(u_{h,\hat{\rho}}^{n+1}(\omega_k) - u_{h,\hat{\rho}}^{n}(\omega_k) \Big) \|^2_H \Big) \\
+ C \sum_{n=0}^{N-1} \triangle t^n \sum_{k=1}^{\hat{N}} \lambda_k \Big(\| \bar{a} \nabla \tilde{U}^{n+1} (\tilde{Y}^{n+1} - Y^n)^\intercal (\omega_k) \|^2_H \\
+ \| \bar{a} \tilde{M}^{n+1^{-1}} \nabla \tilde{U}^{n+1^\intercal} \nabla \tilde{U}^{n+1} (\tilde{Y}^{n+1} - Y^n)^\intercal (\omega_k) \|^2_H \| \tilde{U}^{n+1} \|^2_H \\
+ \| a_{stoch}(\omega_k) \nabla (\tilde{U}^{n+1} - U^n) Y^{n^\intercal}(\omega_k) \|^2_H \\
+ \| a_{stoch}(\omega_k) \tilde{M}^{n+1^{-1}} \nabla (\tilde{U}^{n+1} - U^n)^\intercal \nabla \tilde{U}^{n+1} Y^{n^\intercal}(\omega_k) \|^2_H \| \tilde{U}^{n+1} \|^2_H \Big), \qquad (5.44)
\end{aligned}
$$

where λ_k denotes the weight corresponding to the collocation point ω_k, $k = 1, \dots, \hat{N}$.

Proof. The proof comprises detailing the estimators provided in the previous section. As for the time contribution, we have

$$
\begin{aligned}
\int_0^T \mathscr{E}_{tem,1}(t)\, \mathrm{d}t = \frac{4C_{\mathrm{P}}^2 C_1^2}{C_{\mathcal{L}}} \sum_{n=0}^{N-1} \int_{t^n}^{t^{n+1}} \| \mathcal{I}_p f - \mathcal{I}_p f^{n,n+1} \|^2_{H,L^2_\rho} \mathrm{d}t \\
+ \frac{4C_1^2}{C_{\mathcal{L}}} \sum_{n=0}^{N-1} \int_{t^n}^{t^{n+1}} \| \mathcal{I}_p \mathcal{L}(\tilde{u}) - \mathcal{I}_p \mathcal{L}(u_{h,\hat{\rho}}^n, u_{h,\hat{\rho}}^{n+1}) \|^2_{V',L^2_\rho} \mathrm{d}t.
\end{aligned}
$$

We follow by further bounding this term for implicit, explicit and semi-implicit scheme.

For the implicit scheme, we derive

$$\begin{aligned}\int_0^T \mathscr{E}^{\mathrm{im}}_{tem,1}(t)\,\mathrm{d}t &= \frac{4C_{\mathrm{P}}^2C_1^2}{C_{\mathcal{L}}}\sum_{n=0}^{N-1}\sum_{k=1}^{\hat N}\lambda_k\|f(\omega_k)-f^{n+1}(\omega_k)\|^2_{L^2(t^n,t^{n+1};H)}\\ &\quad+\frac{4C_1^2}{C_{\mathcal{L}}}\sum_{n=0}^{N-1}\sum_{k=1}^{\hat N}\lambda_k\int_{t^n}^{t^{n+1}}\|a(\omega_k)\nabla\Big(u^{n+1}_{h,\hat\rho}(\omega_k)-\tilde u(t,\omega_k)\Big)\|^2_H\mathrm{d}t\\ &=\frac{4C_{\mathrm{P}}^2C_1^2}{C_{\mathcal{L}}}\sum_{n=0}^{N-1}\sum_{k=1}^{\hat N}\lambda_k\|f(\omega_k)-f^{n+1}(\omega_k)\|^2_{L^2(t^n,t^{n+1};H)}\\ &\quad+\frac{4C_1^2}{C_{\mathcal{L}}}\sum_{n=0}^{N-1}\sum_{k=1}^{\hat N}\lambda_k\int_{t^n}^{t^{n+1}}\Big(1-\frac{t-t^n}{\triangle t^n}\Big)^2\|a(\omega_k)\nabla\Big(u^{n+1}_{h,\hat\rho}(\omega_k)-u^n_{h,\hat\rho}(\omega_k)\Big)\|^2_H\mathrm{d}t\\ &=\frac{4C_{\mathrm{P}}^2C_1^2}{C_{\mathcal{L}}}\sum_{n=0}^{N-1}\sum_{k=1}^{\hat N}\lambda_k\|f(\omega_k)-f^{n+1}(\omega_k)\|^2_{L^2(t^n,t^{n+1};H)}\\ &\quad+\frac{4C_1^2}{C_{\mathcal{L}}}\sum_{n=0}^{N-1}\sum_{k=1}^{\hat N}\lambda_k\frac{\triangle t^n}{3}\|a(\omega_k)\nabla\Big(u^{n+1}_{h,\hat\rho}(\omega_k)-u^n_{h,\hat\rho}(\omega_k)\Big)\|^2_H,\end{aligned}$$

where we used the fact that $u^{n+1}_{h,\hat\rho}-\tilde u=\Big(1-\frac{t-t^n}{\triangle t^n}\Big)(u^{n+1}_{h,\hat\rho}-u^n_{h,\hat\rho}),\ t\in(t^n,t^{n+1})$. Analogously, for the explicit scheme we have

$$\begin{aligned}\int_0^T \mathscr{E}^{\mathrm{ex}}_{tem,1}(t)\,\mathrm{d}t &= \frac{4C_{\mathrm{P}}^2C_1^2}{C_{\mathcal{L}}}\sum_{n=0}^{N-1}\sum_{k=1}^{\hat N}\lambda_k\|f(\omega_k)-f^{n}(\omega_k)\|^2_{L^2(t^n,t^{n+1};H)}\\ &\quad+\frac{4C_1^2}{C_{\mathcal{L}}}\sum_{n=0}^{N-1}\sum_{k=1}^{\hat N}\lambda_k\int_{t^n}^{t^{n+1}}\|a(\omega_k)\nabla\Big(u^{n}_{h,\hat\rho}(\omega_k)-\tilde u(t,\omega_k)\Big)\|^2_H\mathrm{d}t\\ &=\frac{4C_{\mathrm{P}}^2C_1^2}{C_{\mathcal{L}}}\sum_{n=0}^{N-1}\sum_{k=1}^{\hat N}\lambda_k\|f(\omega_k)-f^{n}(\omega_k)\|^2_{L^2(t^n,t^{n+1};H)}\\ &\quad+\frac{4C_1^2}{C_{\mathcal{L}}}\sum_{n=0}^{N-1}\sum_{k=1}^{\hat N}\lambda_k\int_{t^n}^{t^{n+1}}\Big(\frac{t^n-t}{\triangle t^n}\Big)^2\|a(\omega_k)\nabla\Big(u^{n+1}_{h,\hat\rho}(\omega_k)-u^n_{h,\hat\rho}(\omega_k)\Big)\|^2_H\mathrm{d}t\\ &=\frac{4C_{\mathrm{P}}^2C_1^2}{C_{\mathcal{L}}}\sum_{n=0}^{N-1}\sum_{k=1}^{\hat N}\lambda_k\|f(\omega_k)-f^{n}(\omega_k)\|^2_{L^2(t^n,t^{n+1};H)}\\ &\quad+\frac{4C_1^2}{C_{\mathcal{L}}}\sum_{n=0}^{N-1}\sum_{k=1}^{\hat N}\lambda_k\frac{\triangle t^n}{3}\|a(\omega_k)\nabla\Big(u^{n+1}_{h,\hat\rho}(\omega_k)-u^n_{h,\hat\rho}(\omega_k)\Big)\|^2_H,\end{aligned}$$

where we used the fact that $u^n_{h,\hat\rho}-\tilde u=\Big(\frac{t^n-t}{\triangle t^n}\Big)(u^{n+1}_{h,\hat\rho}-u^n_{h,\hat\rho})$. Lastly we deal with the semi-implicit scheme, for which we obtain

$$\int_0^T \mathscr{E}^{\mathrm{semi}}_{tem,1}(t)\,\mathrm{d}t \le \frac{4C_{\mathrm{P}}^2C_1^2}{C_{\mathcal{L}}}\sum_{n=0}^{N-1}\sum_{k=1}^{\hat N}\lambda_k\|f(\omega_k)-f^{n+1}(\omega_k)\|^2_{L^2(t^n,t^{n+1};H)}$$

$$
\begin{aligned}
&+ \frac{4C_1^2}{C_{\mathcal{L}}} \sum_{n=0}^{N-1} \sum_{k=1}^{\hat{N}} \lambda_k \int_{t^n}^{t^{n+1}} \|\bar{a}\nabla\Big(u_{h,\hat{\rho}}^{n+1}(\omega_k) - \tilde{u}(t,\omega_k)\Big)\|_H^2 \mathrm{d}t \\
&+ \frac{4C_1^2}{C_{\mathcal{L}}} \sum_{n=0}^{N-1} \sum_{k=1}^{\hat{N}} \lambda_k \int_{t^n}^{t^{n+1}} \|a_{stoch}(\omega_k)\nabla\Big(u_{h,\hat{\rho}}^{n}(\omega_k) - \tilde{u}(t,\omega_k)\Big)\|_H^2 \mathrm{d}t \\
= &\frac{4C_{\mathrm{P}}^2C_1^2}{C_{\mathcal{L}}} \sum_{n=0}^{N-1} \sum_{k=1}^{\hat{N}} \lambda_k \|f(\omega_k) - f^{n+1}(\omega_k)\|_{L^2(t^n,t^{n+1};H)}^2 \\
&+ \frac{4C_1^2}{C_{\mathcal{L}}} \sum_{n=0}^{N-1} \sum_{k=1}^{\hat{N}} \lambda_k \int_{t^n}^{t^{n+1}} \Big(1 - \frac{t-t^n}{\triangle t^n}\Big)^2 \|\bar{a}\nabla\Big(u_{h,\hat{\rho}}^{n+1}(\omega_k) - u_{h,\hat{\rho}}^{n}(\omega_k)\Big)\|_H^2 \mathrm{d}t \\
&+ \frac{4C_1^2}{C_{\mathcal{L}}} \sum_{n=0}^{N-1} \sum_{k=1}^{\hat{N}} \lambda_k \int_{t^n}^{t^{n+1}} \Big(\frac{t^n-t}{\triangle t^n}\Big)^2 \|a_{stoch}(\omega_k)\nabla\Big(u_{h,\hat{\rho}}^{n+1}(\omega_k) - u_{h,\hat{\rho}}^{n}(\omega_k)\Big)\|_H^2 \mathrm{d}t \\
= &\frac{4C_{\mathrm{P}}^2C_1^2}{C_{\mathcal{L}}} \sum_{n=0}^{N-1} \sum_{k=1}^{\hat{N}} \lambda_k \|f(\omega_k) - f^{n+1}(\omega_k)\|_{L^2(t^n,t^{n+1};H)}^2 \\
&+ \frac{4C_1^2}{C_{\mathcal{L}}} \sum_{n=0}^{N-1} \sum_{k=1}^{\hat{N}} \lambda_k \frac{\triangle t^n}{3} \Big(\|\bar{a}\nabla\Big(u_{h,\hat{\rho}}^{n+1}(\omega_k) - u_{h,\hat{\rho}}^{n}(\omega_k)\Big)\|_H^2 \\
&\qquad + \|a_{stoch}(\omega_k)\nabla\Big(u_{h,\hat{\rho}}^{n+1}(\omega_k) - u_{h,\hat{\rho}}^{n}(\omega_k)\Big)\|_H^2\Big)
\end{aligned}
$$

The second part of the time contribution will be again distinguished for the implicit, explicit and semi-implicit scheme. For the implicit case, we have

$$
\begin{aligned}
\left|\int_0^T \mathscr{E}_{tem,2}^{\mathrm{im}}(t)\ \mathrm{d}t\right| &= \frac{2C_1^2}{C_{\mathcal{L}}} \sum_{n=0}^{N-1} \triangle t^n \left\|\mathcal{P}_{\mathcal{Y}^n}^{\perp}\Big[\mathcal{P}_{\tilde{U}^{n+1}}^{\perp}[\mathcal{I}_p\mathcal{L}\Big(\tilde{U}^{n+1}(\tilde{Y}^{n+1} - Y^n)^{\mathsf{T}}\Big)]\Big]\right\|_{V',L_\rho^2}^2 \\
&\leq \frac{2C_1^2}{C_{\mathcal{L}}} \sum_{n=0}^{N-1} \triangle t^n \left\|\mathcal{P}_{\tilde{U}^{n+1}}^{\perp}[\mathcal{I}_p\mathcal{L}\Big(\tilde{U}^{n+1}(\tilde{Y}^{n+1} - Y^n)^{\mathsf{T}}\Big)]\right\|_{V',L_\rho^2}^2 \\
&\leq \frac{2C_1^2}{C_{\mathcal{L}}} \sum_{n=0}^{N-1} \triangle t^n \sum_{k=1}^{\hat{N}} \lambda_k \Bigg(\|a(\omega_k)\nabla\tilde{U}^{n+1}(\tilde{Y}^{n+1} - Y^n)^{\mathsf{T}}(\omega_k)\|_H^2 \\
&\qquad + \|a(\omega_k)\tilde{M}^{{n+1}^{-1}}\nabla\tilde{U}^{{n+1}^{\mathsf{T}}}\nabla\tilde{U}^{n+1}(\tilde{Y}^{n+1} - Y^n)^{\mathsf{T}}(\omega_k)\|_H^2\|\tilde{U}^{n+1}\|_H^2\Bigg).
\end{aligned}
$$

For the explicit case, we analogously derive

$$
\begin{aligned}
\left|\int_0^T \mathscr{E}_{tem,2}^{ex}(t)\ \mathrm{d}t\right| &= \frac{2C_1^2}{C_{\mathcal{L}}} \sum_{n=0}^{N-1} \triangle t^n \left\|\mathcal{P}_{\mathcal{Y}^n}^{\perp}\Big[\mathcal{P}_{\tilde{U}^{n+1}}^{\perp}[\mathcal{I}_p\mathcal{L}\Big((\tilde{U}^{n+1} - U^n)Y^{n^{\mathsf{T}}}\Big)]\Big]\right\|_{V',L_\rho^2}^2 \\
&\leq \frac{2C_1^2}{C_{\mathcal{L}}} \sum_{n=0}^{N-1} \triangle t^n \sum_{k=1}^{\hat{N}} \lambda_k \Bigg(\|a(\omega_k)\nabla(\tilde{U}^{n+1} - U^n)Y^{n^{\mathsf{T}}}(\omega_k)\|_H^2 \\
&\qquad + \|a(\omega_k)\tilde{M}^{{n+1}^{-1}}\nabla(\tilde{U}^{n+1} - U^n)^{\mathsf{T}}\nabla\tilde{U}^{n+1}Y^{n^{\mathsf{T}}}(\omega_k)\|_H^2\|\tilde{U}^{n+1}\|_H^2\Bigg).
\end{aligned}
$$

The last scheme to analyze, is the semi-implicit scheme, which satisfies

$$\begin{aligned}
\left|\int_0^T \mathscr{E}^{semi}_{tem,2}(t)\,\mathrm{d}t\right| &= \frac{2C_1^2}{C_{\mathcal{L}}}\sum_{n=0}^{N-1}\triangle t^n\left\|\mathcal{P}^{\perp}_{\mathcal{Y}^n}\left[\mathcal{P}^{\perp}_{\tilde{U}^{n+1}}\left[\mathcal{I}_p[\mathcal{L}(\tilde{U}^{n+1}Y^{n^\intercal})-\mathcal{L}(u^n_{h,\hat{\rho}},u^{n+1}_{h,\hat{\rho}})]\right]\right]\right\|^2_{V',L^2_\rho}\\
&\leq \frac{2C_1^2}{C_{\mathcal{L}}}\sum_{n=0}^{N-1}\triangle t^n\sum_{k=1}^{\hat{N}}\lambda_k\Big(\|\bar{a}\nabla\tilde{U}^{n+1}(\tilde{Y}^{n+1}-Y^n)^\intercal(\omega_k)\|^2_H\\
&\qquad\qquad + \|\bar{a}\tilde{M}^{n+1^{-1}}\nabla\tilde{U}^{n+1^\intercal}\nabla\tilde{U}^{n+1}(\tilde{Y}^{n+1}-Y^n)^\intercal(\omega_k)\|^2_H\|\tilde{U}^{n+1}\|^2_H\\
&\quad + \|a_{stoch}(\omega_k)\nabla(\tilde{U}^{n+1}-U^n)Y^{n^\intercal}(\omega_k)\|^2_H\\
&\qquad\quad + \|a_{stoch}(\omega_k)\tilde{M}^{n+1^{-1}}\nabla(\tilde{U}^{n+1}-U^n)^\intercal\nabla\tilde{U}^{n+1}Y^{n^\intercal}(\omega_k)\|^2_H\|\tilde{U}^{n+1}\|^2_H\Big).
\end{aligned}$$

For the spatial part, we will follow the estimation provided in [Ver03]. We denote by J_h any of the quasi interpolation operators of [Ver99b] defined on V and with values in the space of continuous, piecewise linear finite element functions corresponding to $\mathcal{T}_h$. Then, combining the interpolation error estimates of [Ver99b], a standard trace theorem [Ver99b, Lemma 3.2], the following estimates hold for every $v \in V$ and for any element $K \in \mathcal{T}_h$ and interior edge/face $E \in \mathcal{E}_h$

$$\begin{aligned}
\|\nabla(v-J_hv)\|_{L^2(K)} &\leq c_0\|\nabla v\|_{L^2(\tilde{\gamma}_K)},\\
\|v-J_hv\|_{L^2(K)} &\leq \tilde{c}_1h_K\|\nabla v\|_{L^2(\tilde{\gamma}_K)},\\
\|v-J_hv\|_{L^2(E)} &\leq c_2\Big\{h_E^{-1/2}\|v-J_hv\|_{L^2(K)}+h_E^{1/2}\|\nabla(v-J_hv)\|_{L^2(K)}\Big\}\\
&\leq \tilde{c}_2h_E^{1/2}\|\nabla v\|_{L^2(\tilde{\gamma}_K)},
\end{aligned}\tag{5.45}$$

where $\tilde{\gamma}_K$ denotes the subset that consists of all elements of $\mathcal{T}_h$ sharing at least a vertex with K. The constants $c_0, \tilde{c}_1, c_2, \tilde{c}_2$ only depend on the maximal ratio of the diameter of any element to the diameter of its largest inscribed ball. Note that these estimates are equivalent to the estimates in (5.20).

With η_K denoting a unit outward pointing normal we further derive

$$\begin{aligned}
|A_{spa}| &= \left|\Big\langle\mathcal{I}_pf-\dot{\tilde{u}}-\mathcal{I}_p\mathcal{L}(\tilde{u}),v-v_h\Big\rangle_{V'V,L^2_\rho}\right|\\
&= \int_\Omega\Big(\int_D\mathcal{I}_pf(\omega)\Big(v-v_h\Big)(\omega)-\int_D\dot{\tilde{u}}(\omega)\Big(v-v_h\Big)(\omega)\\
&\qquad\qquad -\int_D\mathcal{I}_p[a(\omega)\nabla\tilde{u}(\omega)]\nabla\Big(v-v_h\Big)(\omega)\Big)\,\mathrm{d}\rho\\
&= \int_\Omega\Big(\sum_{K\in\mathscr{T}_h}\int_K\Big(\mathcal{I}_pf(\omega)-\dot{\tilde{u}}(\omega)+\mathcal{I}_p[\nabla\cdot(a(\omega)\nabla\tilde{u}(\omega))]\Big)\Big(v-v_h\Big)(\omega)
\end{aligned}$$

$$- \sum_{E \in \mathscr{E}_h} \int_E \mathcal{I}_p \Big[[a(\omega) \nabla \tilde{u}(\omega) \cdot \eta_E]_E \Big] \Big(v - v_h \Big)(\omega) \Big) \, \mathrm{d}\rho$$

Considering $v_h(\omega) = J_h v(\omega)$ and applying (5.45) leads to

$$\begin{aligned}
|A_{spa}| \leq \int_\Omega \Big[& \sum_{K \in \mathscr{T}_h} \tilde{c}_1 h_K \Big\| \mathcal{I}_p f(\omega) - \dot{\tilde{u}}(\omega) + \mathcal{I}_p [\nabla \cdot (a(\omega) \nabla \tilde{u}(\omega))] \Big\|_{L^2(K)} \Big\| \nabla v(\omega) \Big\|_{L^2(\tilde{\gamma}_K)} \\
& + \sum_{E \in \mathscr{E}_h} \tilde{c}_2 h_E^{1/2} \Big\| \mathcal{I}_p \Big[[a(\omega) \nabla \tilde{u}(\omega) \cdot \eta_E]_E \Big] \Big\|_{L^2(E)} \Big\| \nabla v(\omega) \Big\|_{L^2(\tilde{\gamma}_K)} \Big] \mathrm{d}\rho \\
\leq C_2 \Bigg\| \Big(& \sum_{K \in \mathscr{T}_h} h_K^2 \Big\| \mathcal{I}_p f - \dot{\tilde{u}} + \mathcal{I}_p [\nabla \cdot (a \nabla \tilde{u})] \Big\|^2_{L^2(K)} \Big)^{1/2} \Bigg\|_{L^2_\rho} \Big\| \nabla v \Big\|_{H, L^2_\rho} \\
& + \Bigg\| \Big(\sum_{E \in \mathscr{E}_h} h_E \Big\| \mathcal{I}_p \Big[[a \nabla \tilde{u} \cdot \eta_E]_E \Big] \Big\|^2_{L^2(E)} \Big)^{1/2} \Bigg\|_{L^2_\rho} \Big\| \nabla v \Big\|_{H, L^2_{\hat{\rho}}} \\
= C_2 \Big(& \sum_{k=1}^{\hat{N}} \lambda_k \Big(\sum_{K \in \mathscr{T}_h} h_K^2 \Big\| f(\omega_k) - \dot{\tilde{u}}(\omega_k) + \nabla \cdot (a(\omega_k) \nabla \tilde{u}(\omega_k)) \Big\|^2_{L^2(K)} \\
& + \sum_{E \in \mathscr{E}_h} h_E \Big\| [a(\omega_k) \nabla \tilde{u}(\omega_k) \cdot \eta_E]_E \Big\|^2_{L^2(E)} \Big) \Big)^{1/2} \|v\|_{V, L^2_\rho} \\
\leq \frac{2C_2^2}{C_{\mathcal{L}}} \Big(& \sum_{k=1}^{\hat{N}} \lambda_k \Big(\sum_{K \in \mathscr{T}_h} h_K^2 \Big\| f(\omega_k) - \dot{\tilde{u}}(\omega_k) + \nabla \cdot (a(\omega_k) \nabla \tilde{u}(\omega_k)) \Big\|^2_{L^2(K)} \\
& + \sum_{E \in \mathscr{E}_h} h_E \Big\| [a(\omega_k) \nabla \tilde{u}(\omega_k) \cdot \eta_E]_E \Big\|^2_{L^2(E)} \Big) \Big) + \frac{C_{\mathcal{L}}}{8} \|v\|^2_{V, L^2_\rho}
\end{aligned}$$

i.e.

$$\begin{aligned}
\mathscr{E}^{spa} = \frac{2C_2^2}{C_{\mathcal{L}}} \Big(\sum_{k=1}^{\hat{N}} \lambda_k \Big(& \sum_{K \in \mathscr{T}_h} h_K^2 \Big\| f(\omega_k) - \dot{\tilde{u}}(\omega_k) + \nabla \cdot (a(\omega_k) \nabla \tilde{u}(\omega_k)) \Big\|^2_{L^2(K)} \\
& + \sum_{E \in \mathscr{E}_h} h_E \Big\| [a(\omega_k) \nabla \tilde{u}(\omega_k) \cdot \eta_E]_E \Big\|^2_{L^2(E)} \Big) \Big).
\end{aligned}$$

Concerning the rank error estimation, we proceed as follows

$$\begin{aligned}
\mathscr{E}^{rank} &= \frac{2C_1^2}{C_{\mathcal{L}}} \Big\| \mathcal{P}^\perp_{\mathcal{Y}^n} \Big[\mathcal{P}^\perp_{\tilde{\mathcal{U}}^{n+1}} \Big[\mathcal{I}_p f^{n+1^*} - \mathcal{I}_p \mathcal{L}^* (\tilde{U}^{n+1} Y^{n^\intercal}) \Big] \Big] \Big\|^2_{V', L^2_\rho} \\
&\leq \frac{4C_1^2}{C_{\mathcal{L}}} \Big(\Big\| \mathcal{P}^\perp_{\mathcal{Y}^n} \Big[\mathcal{P}^\perp_{\tilde{\mathcal{U}}^{n+1}} \Big[\mathcal{I}_p f^{n+1^*} \Big] \Big] \Big\|^2_{H, L^2_\rho} + \Big\| \mathcal{P}^\perp_{\mathcal{Y}^n} \Big[\mathcal{P}^\perp_{\tilde{\mathcal{U}}^{n+1}} \Big[\mathcal{I}_p \mathcal{L}^* (\tilde{U}^{n+1} Y^{n^\intercal}) \Big] \Big] \Big\|^2_{V', L^2_\rho} \Big).
\end{aligned}$$

For the second term, we perform the following computation. For $v \in L^2_\rho(\Omega; V)$, it holds

$$
\begin{aligned}
&\Big(\mathcal{P}^\perp_{\tilde{U}^{n+1}}\Big[\mathcal{P}^\perp_{\mathcal{Y}^n}\Big[\mathcal{I}_p\mathcal{L}^*(\tilde{U}^{n+1}Y^{n^\mathsf{T}})\Big]\Big], v\Big)_{V'V,L^2_\rho} = \int_\Omega\int_D \mathcal{P}^\perp_{\mathcal{Y}^n}\Big[\mathcal{I}_p[-\nabla\cdot(a\nabla\tilde{U}^{n+1})Y^{n^\mathsf{T}}]^*\Big] v\,\mathrm{d}x\,\mathrm{d}\rho \\
&\qquad - \int_\Omega\int_D \tilde{U}^{n+1}\tilde{M}^{n+1^{-1}}\int_D \tilde{U}^{n+1^\mathsf{T}}\mathcal{P}^\perp_{\mathcal{Y}^n}\Big[\mathcal{I}_p[-\nabla\cdot(a\nabla\tilde{U}^{n+1})Y^{n^\mathsf{T}}]^*\Big]\,\mathrm{d}\hat{x}\; v\,\mathrm{d}x\,\mathrm{d}\rho \\
&\quad = \int_\Omega\sum_K\int_K \mathcal{P}^\perp_{\mathcal{Y}^n}\Big[\mathcal{I}_p[-\nabla\cdot(a\nabla\tilde{U}^{n+1})Y^{n^\mathsf{T}}]^*\Big] v\,\mathrm{d}x \\
&\qquad + \sum_E\int_E \mathcal{P}^\perp_{\mathcal{Y}^n}\Big[\mathcal{I}_p[[a\nabla\tilde{U}^{n+1}\cdot\eta_E]_E Y^{n^\mathsf{T}}]^*\Big]\, v\,\mathrm{d}x\,\mathrm{d}\rho \\
&\qquad - \int_\Omega\int_D \tilde{U}^{n+1}\tilde{M}^{n+1^{-1}}\Big(\sum_K\int_K \tilde{U}^{n+1^\mathsf{T}}\mathcal{P}^\perp_{\mathcal{Y}^n}\Big[\mathcal{I}_p[-\nabla\cdot(a\nabla\tilde{U}^{n+1})Y^{n^\mathsf{T}}]^*\Big]\mathrm{d}\hat{x} \\
&\qquad\qquad + \sum_E\int_E \tilde{U}^{n+1^\mathsf{T}}\mathcal{P}^\perp_{\mathcal{Y}^n}\Big[\mathcal{I}_p[[a\nabla\tilde{U}^{n+1}\cdot\eta_E]_E Y^{n^\mathsf{T}}]^*\Big]\mathrm{d}\hat{x}\Big)\, v\,\mathrm{d}x\,\mathrm{d}\rho \\
&\quad = \int_\Omega\sum_K\int_K \mathcal{P}^\perp_{\mathcal{Y}^n}\Big[\mathcal{I}_p[-\nabla\cdot(a\nabla\tilde{U}^{n+1})Y^{n^\mathsf{T}}]^*\Big] v\,\mathrm{d}x \\
&\qquad + \sum_E\int_E \mathcal{P}^\perp_{\mathcal{Y}^n}\Big[\mathcal{I}_p[[a\nabla\tilde{U}^{n+1}\cdot\eta_E]_E Y^{n^\mathsf{T}}]^*\Big]\, v\,\mathrm{d}x\,\mathrm{d}\rho \\
&\qquad - \int_\Omega\sum_K\int_K \tilde{U}^{n+1}\tilde{M}^{n+1^{-1}}\Big(\sum_K\int_K \tilde{U}^{n+1^\mathsf{T}}\mathcal{P}^\perp_{\mathcal{Y}^n}\Big[\mathcal{I}_p[-\nabla\cdot(a\nabla\tilde{U}^{n+1})Y^{n^\mathsf{T}}]^*\Big]\mathrm{d}\hat{x}\Big)\, v\,\mathrm{d}x\,\mathrm{d}\rho \\
&\qquad - \int_\Omega\int_D \tilde{U}^{n+1}\tilde{M}^{n+1^{-1}}\Big(\sum_E\int_E \tilde{U}^{n+1^\mathsf{T}}\mathcal{P}^\perp_{\mathcal{Y}^n}\Big[\mathcal{I}_p[[a\nabla\tilde{U}^{n+1}\cdot\eta_E]_E Y^{n^\mathsf{T}}]^*\Big]\mathrm{d}\hat{x}\Big)\, v\,\mathrm{d}x\,\mathrm{d}\rho \\
&\quad = \int_\Omega\sum_K\int_K \Big(\mathcal{P}^\perp_{\mathcal{Y}^n}\Big[\mathcal{I}_p[-\nabla\cdot(a\nabla\tilde{U}^{n+1})Y^{n^\mathsf{T}}]^*\Big] \\
&\qquad\qquad - \tilde{U}^{n+1}\tilde{M}^{n+1^{-1}}\Big(\sum_K\int_K \tilde{U}^{n+1^\mathsf{T}}\mathcal{P}^\perp_{\mathcal{Y}^n}\Big[\mathcal{I}_p[-\nabla\cdot(a\nabla\tilde{U}^{n+1})Y^{n^\mathsf{T}}]^*\Big]\mathrm{d}\hat{x}\Big)\, v\,\mathrm{d}x\,\mathrm{d}\rho \\
&\qquad + \int_\Omega\sum_E\int_E \mathcal{P}^\perp_{\mathcal{Y}^n}\Big[\mathcal{I}_p[[a\nabla\tilde{U}^{n+1}\cdot\eta_E]_E Y^{n^\mathsf{T}}]^*\Big]\, v\,\mathrm{d}x\,\mathrm{d}\rho \\
&\qquad - \int_\Omega\int_D \tilde{U}^{n+1}\tilde{M}^{n+1^{-1}}\Big(\sum_E\int_E \tilde{U}^{n+1^\mathsf{T}}\mathcal{P}^\perp_{\mathcal{Y}^n}\Big[\mathcal{I}_p[[a\nabla\tilde{U}^{n+1}\cdot\eta_E]_E Y^{n^\mathsf{T}}]^*\Big]\mathrm{d}\hat{x}\Big)\, v\,\mathrm{d}x\,\mathrm{d}\rho \\
&\quad = \int_\Omega\sum_K\int_K {\Pi^h_{\tilde{U}^{n+1}Y^{n^\mathsf{T}}}}^{\perp}\Big[\mathcal{I}_p\Big[\Big(-\nabla\cdot(a\nabla\tilde{U}^{n+1})Y^{n^\mathsf{T}}\Big)^*\Big]\Big] v\,\mathrm{d}x\,\mathrm{d}\rho \\
&\qquad + \int_\Omega\sum_E\int_E \mathcal{P}^\perp_{\mathcal{Y}^n}\Big[\mathcal{I}_p[[a\nabla\tilde{U}^{n+1}\cdot\eta_E]_E Y^{n^\mathsf{T}}]^*\Big]\, v\,\mathrm{d}x\,\mathrm{d}\rho \\
&\qquad - \int_\Omega\int_D \tilde{U}^{n+1}\tilde{M}^{n+1^{-1}}\Big(\sum_E\int_E \tilde{U}^{n+1^\mathsf{T}}\mathcal{P}^\perp_{\mathcal{Y}^n}\Big[\mathcal{I}_p[[a\nabla\tilde{U}^{n+1}\cdot\eta_E]_E Y^{n^\mathsf{T}}]^*\Big]\mathrm{d}\hat{x}\Big)\, v\,\mathrm{d}x\,\mathrm{d}\rho
\end{aligned}
$$

where the expression ${\Pi^h_{\tilde{U}^{n+1}Y^{n^\mathsf{T}}}}^{\perp}\Big[\mathcal{I}_p\Big[\Big(-\nabla\cdot(a\nabla\tilde{U}^{n+1})Y^{n^\mathsf{T}}\Big)^*\Big]\Big]$ is considered as a function defined only inside of the elements K. We can further proceed as

$$
\begin{aligned}
&\left|\left(\mathcal{P}^{\perp}_{\mathcal{Y}^n}\left[\mathcal{P}^{\perp}_{\tilde{\mathcal{U}}^{n+1}}\left[\mathcal{I}_p\mathcal{L}(\tilde{U}^{n+1}Y^{n^\intercal})\right]\right], v\right)_{V'V,L^2_\rho}\right| \\
&\quad\le \sum_K \left\|{\Pi^{h,\hat{\rho}}_{\tilde{U}^{n+1}Y^{n^\intercal}}}^{\perp}[(-\nabla\cdot(a\nabla\tilde{U}^{n+1})Y^{n^\intercal})^*]\right\|_{L^2(K),L^2_\rho}\|v\|_{L^2(K),L^2_\rho} \\
&\qquad+\sum_E \left\|\mathcal{P}^{\hat{\rho}\perp}_{\mathcal{Y}^n}\left[\left([a\nabla\tilde{U}^{n+1}\cdot\eta_E]_E Y^{n^\intercal}\right)^*\right]\right\|_{L^2(E),L^2_\rho}\|v\|_{L^2(E),L^2_\rho} \\
&\qquad+\sum_E \|\tilde{U}^{n+1}\tilde{M}^{n+1^{-1}}\|_{H,L^2_\rho}\|\tilde{U}^{n+1^\intercal}\mathcal{P}^{\perp}_{\mathcal{Y}^n}\left[\mathcal{I}_p[[a\nabla\tilde{U}^{n+1}\cdot\eta_E]_E Y^{n^\intercal}]^*\right]\|_{L^2(E),L^2_\rho}\|v\|_{H,L^2_\rho} \\
&\quad\le C_3\Bigg(\sum_K \left\|{\Pi^{h,\hat{\rho}}_{\tilde{U}^{n+1}Y^{n^\intercal}}}^{\perp}[(-\nabla\cdot(a\nabla\tilde{U}^{n+1})Y^{n^\intercal})^*]\right\|_{L^2(K),L^2_\rho} \\
&\qquad+\sum_E \left\|\mathcal{P}^{\hat{\rho}\perp}_{\mathcal{Y}^n}\left[\left([a\nabla\tilde{U}^{n+1}\cdot\eta_E]_E Y^{n^\intercal}\right)^*\right]\right\|_{L^2(E),L^2_\rho} \\
&\qquad+\sum_E \|\tilde{U}^{n+1}\tilde{M}^{n+1^{-1}}\|_{H,L^2_\rho}\|\tilde{U}^{n+1^\intercal}\mathcal{P}^{\perp}_{\mathcal{Y}^n}\left[\mathcal{I}_p[[a\nabla\tilde{U}^{n+1}\cdot\eta_E]_E Y^{n^\intercal}]^*\right]\|_{L^2(E),L^2_\rho}\Bigg)\|v\|_{V,L^2_\rho} \\
&\quad= C_3\Bigg(\sum_{k=1}^{\hat{N}}\lambda_k\Bigg(\sum_K \left\|{\Pi^{h,\hat{\rho}}_{\tilde{U}^{n+1}Y^{n^\intercal}}}^{\perp}[(-\nabla\cdot(a\nabla\tilde{U}^{n+1})Y^{n^\intercal})^*](\omega_k)\right\|^2_{L^2(K)} \\
&\qquad+\sum_E \left\|\mathcal{P}^{\hat{\rho}\perp}_{\mathcal{Y}^n}\left[\left([a\nabla\tilde{U}^{n+1}\cdot\eta_E]_E Y^{n^\intercal}\right)^*\right](\omega_k)\right\|^2_{L^2(E)} \\
&\qquad+\sum_E \|\tilde{U}^{n+1}\tilde{M}^{n+1^{-1}}\|^2_{H,L^2_\rho}\left\|\tilde{U}^{n+1^\intercal}\mathcal{P}^{\perp}_{\mathcal{Y}^n}\left[\mathcal{I}_p[[a\nabla\tilde{U}^{n+1}\cdot\eta_E]_E Y^{n^\intercal}(\omega_k)]^*\right]\right\|^2_{L^2(E),L^2_\rho}\Bigg)\Bigg)^{1/2}\|v\|_{V,L^2_\rho}.
\end{aligned}
$$

Plugging this in the definition of $\mathscr{E}^{rank}$, we derive that

$$
\begin{aligned}
\mathscr{E}^{rank} = \frac{4C_1^2}{C_{\mathcal{L}}}\Bigg(&\left\|\mathcal{P}^{\perp}_{\mathcal{Y}^n}\left[\mathcal{P}^{\perp}_{\tilde{\mathcal{U}}^{n+1}}\left[\mathcal{I}_p[f^{n+1^*}]\right]\right]\right\|^2_{H,L^2_\rho} \\
&+ C_3\sum_{k=1}^{\hat{N}}\lambda_k\Bigg(\sum_K \left\|{\Pi^{h,\hat{\rho}}_{\tilde{U}^{n+1}Y^{n^\intercal}}}^{\perp}[(-\nabla\cdot(a\nabla\tilde{U}^{n+1})Y^{n^\intercal})^*](\omega_k)\right\|^2_{L^2(K)} \\
&\qquad+\sum_E\Bigg(\left\|\mathcal{P}^{\hat{\rho}\perp}_{\mathcal{Y}^n}\left[\left([a\nabla\tilde{U}^{n+1}\cdot\eta_E]_E Y^{n^\intercal}\right)^*\right](\omega_k)\right\|^2_{L^2(E)} \\
&\qquad+\|\tilde{U}^{n+1}\tilde{M}^{n+1^{-1}}\|_H\left\|\tilde{U}^{n+1^\intercal}\mathcal{P}^{\perp}_{\mathcal{Y}^n}\left[\mathcal{I}_p[[a\nabla\tilde{U}^{n+1}\cdot\eta_E]_E Y^{n^\intercal}(\omega_k)]^*\right]\right\|^2_{L^2(E)}\Bigg)\Bigg).
\end{aligned}
$$

The last error contribution to analyze is the stochastic error contribution. We recall that

$$
\int_0^T \mathscr{E}_{sto}(t)\mathrm{d}t = \int_0^T \frac{4C_{\mathrm{P}}^2}{C_{\mathcal{L}}}\|f-\mathcal{I}_p f\|^2_{H,L^2_\rho} + \frac{4}{C_{\mathcal{L}}}\|\mathcal{L}(\tilde{u})-\mathcal{I}_p\mathcal{L}(\tilde{u})\|^2_{V',L^2_\rho}\mathrm{d}t.
$$

For the second term, we proceed by

$$\begin{aligned}
\int_0^T \|\mathcal{L}(\tilde{u}) - \mathcal{I}_p\mathcal{L}(\tilde{u})\|^2_{V',L^2_\rho}\mathrm{d}t &\le \int_0^T \Big\| a\nabla\tilde{u} - \mathcal{I}_p[a\nabla\tilde{u}]\Big\|^2_{H,L^2_\rho}\mathrm{d}t \\
&= \int_\Omega\int_D \sum_{n=0}^{N-1}\int_{t^n}^{t^{n+1}} \Big(\Big(\frac{t-t^n}{\triangle t^n}\Big)\Big(a\nabla u^{n+1}_{h,\hat\rho} - \mathcal{I}_p a\nabla u^{n+1}_{h,\hat\rho}\Big) \\
&\quad + \Big(\frac{t^{n+1}-t}{\triangle t^n}\Big)\Big(a\nabla u^{n}_{h,\hat\rho} - \mathcal{I}_p a\nabla u^{n}_{h,\hat\rho}\Big)\Big)^2 \mathrm{d}t\,\mathrm{d}x\,\mathrm{d}\rho \\
&\le 2\int_\Omega\int_D \sum_{n=0}^{N-1}\int_{t^n}^{t^{n+1}} \Big(\Big(\frac{t-t^n}{\triangle t^n}\Big)\Big(a\nabla u^{n+1}_{h,\hat\rho} - \mathcal{I}_p a\nabla u^{n+1}_{h,\hat\rho}\Big)\Big)^2 \\
&\quad + \Big(\Big(\frac{t^{n+1}-t}{\triangle t^n}\Big)\Big(a\nabla u^{n}_{h,\hat\rho} - \mathcal{I}_p a\nabla u^{n}_{h,\hat\rho}\Big)\Big)^2 \mathrm{d}t\,\mathrm{d}x\,\mathrm{d}\rho \\
&= \frac{2}{3}\triangle t^n \sum_{n=0}^{N-1}\Big(\Big\|a\nabla u^{n+1}_{h,\hat\rho} - \mathcal{I}_p a\nabla u^{n+1}_{h,\hat\rho}\Big\|^2_{H,L^2_\rho} + \Big\|a\nabla u^{n}_{h,\hat\rho} - \mathcal{I}_p a\nabla u^{n}_{h,\hat\rho}\Big\|^2_{H,L^2_\rho}\Big).
\end{aligned}$$

□

5.3.1 Adaptive algorithm

The estimators from the preceding section provide us with a fully computable upper bound of the error caused by the spatial, time, stochastic discretization as well as the rank truncation. In this section, we will see how these estimators can be naturally localized in all variables – time, space, stochastics and rank. Up to this point, we only considered DLR approximation with a rank R fixed in time. However, in what follows we will allow different ranks R^n for different time intervals $[t^n, t^{n+1}]$, $n = 0, \dots, N-1$. At time $t = t^{n+1}$, the new obtained solution $u^{n+1}_{h,\hat\rho}$ is of rank R^n. When $R^n \neq R^{n+1}$, we need to update the solution $u^{n+1}_{h,\hat\rho}$ to be of rank R^{n+1} so that the method can proceed with the new time step. This is performed in the following way. If $R^n > R^{n+1}$, we lose $(R^n - R^{n+1})$ terms in $(u^{n+1}_{h,\hat\rho})^* = \sum_{r=1}^{R^n} U^{n+1}_r Y^{n+1}_r$ corresponding to the smallest singular values of $(u^{n+1}_{h,\hat\rho})^*$. This results in $(u^{n+1}_{h,\hat\rho})^* = \sum_{r=1}^{R^{n+1}} U^{n+1}_r Y^{n+1}_r$ of rank R^{n+1}. If, on the other hand, $R^{n+1} > R^n$, we define a new solution $(u^{n+1}_{h,\hat\rho})^* = \sum_{r=1}^{R^{n+1}} U^{n+1}_r Y^{n+1}_r$, where $U^{n+1}_r = 0$, for $r = R^n + 1, \dots, R^{n+1}$ and Y^{n+1}_r, $r = R^n + 1, \dots, R^{n+1}$ has to be chosen in a way that $\{Y^{n+1}_r\}_{r=1}^{R^{n+1}}$ forms an orthonormal basis. Different possible choices of $\{Y^{n+1}_r\}_{r=1}^{R^{n+1}}$ will be discussed here after.

We here propose an adaptive algorithm for the FE meshes, time discretization, tensor grids and rank, with the goal to obtain the overall error

$$\|u_{\mathrm{true}}(T) - \tilde{u}(T)\|^2_{H,L^2_\rho} + \frac{3}{4}a_{\min}\|u_{\mathrm{true}} - \tilde{u}\|^2_{L^2(0,T;L^2_\rho(\hat\Omega;V))}$$

under a prescribed tolerance TOL. For a deterministic right hand side, the corresponding error estimators are summarized in the following theorem.

Theorem 5.3.2. *Let us assume that the term f is deterministic. Then the stochastic error estimator ε_{sto} from* (5.38)*, the spatial error estimator ε_{spa} from* (5.39)*, the time error estimator ε_{tem} from* (5.41) *and the rank truncation error estimator ε_{rank} from* (5.40) *can be alternatively expressed as*

$$\varepsilon_{spa} = \sum_{K \in \mathscr{T}_h} \varepsilon_{spa,K}$$

$$\begin{aligned} \varepsilon_{spa,K} = C h_K^2 \Big(\sum_{k=1}^{\hat{N}} \lambda_k \Big(\Big\| f(\omega_k) - \dot{\tilde{u}}(\omega_k) + \nabla \cdot (a(\omega_k) \nabla \tilde{u}(\omega_k)) \Big\|^2_{L^2(0,T;L^2(K))} \Big) \\ + \sum_{E \subset \partial K} h_E \Big(\sum_{k=1}^{\hat{N}} \lambda_k \Big(\Big\| [a(\omega_k) \nabla \tilde{u}(\omega_k) \cdot \eta_E]_E \Big\|^2_{L^2(0,T;L^2(E))} \Big), \end{aligned} \tag{5.46}$$

and

$$\varepsilon_{stoch} = \sum_{m=1}^{M} \varepsilon_{stoch,m}$$

$$\begin{aligned} \varepsilon_{stoch,m} = C \sum_{n=0}^{N-1} \triangle t^n M \Big(\Big\| a_m \omega^m \nabla u_{h,\hat{\rho}}^{n+1} - \mathcal{I}_p [a_m \omega^m \nabla u_{h,\hat{\rho}}^{n+1}] \Big\|^2_{H, L^2_\rho} \\ + \Big\| a_m \omega^m \nabla u_{h,\hat{\rho}}^{n} - \mathcal{I}_p [a_m \omega^m \nabla u_{h,\hat{\rho}}^{n}] \Big\|^2_{H, L^2_\rho} \Big) \end{aligned} \tag{5.47}$$

and

$$\varepsilon_{tem}^{\mathrm{im}} = \sum_{n=0}^{N-1} \varepsilon_{tem,n}^{\mathrm{im}}$$

$$\begin{aligned} \varepsilon_{tem,n}^{\mathrm{im}} = C \sum_{k=1}^{\hat{N}} \lambda_k \| f(\omega_k) - f^{n+1}(\omega_k) \|^2_{L^2(t^n, t^{n+1}; H)} \\ + \sum_{n=0}^{N-1} \sum_{k=1}^{\hat{N}} \lambda_k \frac{\triangle t^n}{3} \| a(\omega_k) \nabla \Big(u_{h,\hat{\rho}}^{n+1}(\omega_k) - u_{h,\hat{\rho}}^{n}(\omega_k) \Big) \|^2_H \\ + \sum_{n=0}^{N-1} \triangle t^n \sum_{k=1}^{\hat{N}} \lambda_k \Big(\| a(\omega_k) \nabla \tilde{U}^{n+1} (\tilde{Y}^{n+1} - Y^n)^\intercal (\omega_k) \|^2_H \\ + \| a(\omega_k) \tilde{M}^{{n+1}^{-1}} \nabla \tilde{U}^{{n+1}^\intercal} \nabla \tilde{U}^{n+1} (\tilde{Y}^{n+1} - Y^n)^\intercal (\omega_k) \|^2_H \| \tilde{U}^{n+1} \|^2_H \Big) \end{aligned} \tag{5.48}$$

and analogously for $\varepsilon_{tem}^{\text{ex}}$*,* $\varepsilon_{tem}^{\text{semi}}$*. The rank estimator can be localized in time*

$$\varepsilon_{rank} = \sum_{n=0}^{N-1} \varepsilon_{rank,n},$$

where

$$\begin{aligned}
\varepsilon_{rank,n} = C\triangle t^n \sum_{k=1}^{\hat{N}} \lambda_k \Bigg(\sum_{K\in\mathcal{T}_h} \Big\| {\Pi_{\tilde{U}^{n+1}Y^{n\intercal}}^{h,\hat{\rho}}}^{\perp} \Big[&- \nabla\cdot(a_{stoch}\nabla\tilde{U}^{n+1})Y^{n^\intercal} \\
&+ \mathbb{E}_{\hat{N}}[\nabla\cdot(a_{stoch}\nabla\tilde{U}^{n+1})Y^{n^\intercal}]\Big](\omega_k)\Big\|_{L^2(K)}^2 \\
+ \sum_{E\in\mathscr{E}_h} \Big\| \mathcal{P}_{\mathcal{Y}^n}^{\hat{\rho}\perp}\Big[[a_{stoch}\nabla\tilde{U}^{n+1}\cdot\eta_E]_E Y^{n^\intercal} &- \mathbb{E}_{\hat{N}}[[a_{stoch}\nabla\tilde{U}^{n+1}\cdot\eta_E]_E Y^{n^\intercal}]\Big](\omega_k)\Big\|_{L^2(E)}^2 \\
+ \|\tilde{U}^{n+1}{\tilde{M}^{n+1}}{}^{-1}\|_H^2 \Big\| \tilde{U}^{n+1\intercal}\mathcal{P}_{\mathcal{Y}^n}^{\hat{\rho}\perp}\Big[&[a_{stoch}\nabla\tilde{U}^{n+1}\cdot\eta_E]_E Y^{n^\intercal} \\
&- \mathbb{E}_{\hat{N}}[[a_{stoch}\nabla\tilde{U}^{n+1}\cdot\eta_E]_E Y^{n^\intercal}]\Big](\omega_k)\Big\|_{L^2(E)}^2 \Bigg).
\end{aligned}$$

In addition, $\varepsilon_{rank,n}$ *can be further localized in new random directions for every* $n = 0,\dots,N-1$ *as*

$$\varepsilon_{rank,n} = \sum_{m=1}^{M}\sum_{r=1}^{R^n} \varepsilon_{rank,n,m,r}, \tag{5.49}$$

where

$$\begin{aligned}
\varepsilon_{rank,n,m,r} = C\triangle t^n \sum_{k=1}^{\hat{N}} \lambda_k \Bigg(\sum_{K\in\mathcal{T}_h} \Big\| {\Pi_{\tilde{U}^{n+1}Y^{n\intercal}}^{h,\hat{\rho}}}^{\perp} \Big[&- \nabla\cdot(a_m\nabla\tilde{U}_r^{n+1}) \\
&(\omega^m Y_r^n - \mathbb{E}_{\hat{N}}[\omega^m Y_r^n])\Big](\omega_k)\Big\|_{L^2(K)}^2 \\
+ \sum_{E\in\mathscr{E}_h} \Big\| \mathcal{P}_{\mathcal{Y}^n}^{\hat{\rho}\perp}\Big[[a_m\nabla\tilde{U}_r^{n+1}\cdot\eta_E]_E&(\omega^m Y_r^n - \mathbb{E}_{\hat{N}}[\omega^m Y_r^n])\Big](\omega_k)\Big\|_{L^2(E)}^2 \\
+ \|\tilde{U}^{n+1}{\tilde{M}^{n+1}}{}^{-1}\|_H^2 \Big\| \tilde{U}^{n+1\intercal}\mathcal{P}_{\mathcal{Y}^n}^{\hat{\rho}\perp}\Big[[a_m\nabla\tilde{U}_r^{n+1}\cdot\eta_E]_E&(\omega^m Y_r^n - \mathbb{E}_{\hat{N}}[\omega^m Y_r^n])\Big](\omega_k)\Big\|_{L^2(E)}^2 \Bigg).
\end{aligned} \tag{5.50}$$

Proof. The localization of the spatial error estimator into every element and time estimator into every time step stems from a natural rearrangement of terms in the spatial and time error estimates (5.39), (5.42). For a deterministic forcing term f, the stochastic error estimate can be bounded as

$$\varepsilon_{sto} = \sum_{n=0}^{N-1} \triangle t^n \Big(\Big\| a\nabla u_{h,\hat{\rho}}^{n+1} - \mathcal{I}_p[a\nabla u_{h,\hat{\rho}}^{n+1}]\Big\|_{H,L^2_\rho}^2 + \Big\| a\nabla u_{h,\hat{\rho}}^{n} - \mathcal{I}_p[a\nabla u_{h,\hat{\rho}}^{n}]\Big\|_{H,L^2_\rho}^2 \Big)$$

$$
\begin{aligned}
&= \sum_{n=0}^{N-1} \triangle t^n \Big(\Big\| a_{sto} \nabla u_{h,\hat{\rho}}^{n+1} - \mathcal{I}_p[a_{sto} \nabla u_{h,\hat{\rho}}^{n+1}] \Big\|^2_{H,L^2_\rho} + \Big\| a_{sto} \nabla u_{h,\hat{\rho}}^{n} - \mathcal{I}_p[a_{sto} \nabla u_{h,\hat{\rho}}^{n}] \Big\|^2_{H,L^2_\rho} \Big) \\
&\leq \sum_{n=0}^{N-1} \triangle t^n \Big(\Big\| \sum_{m=1}^{M} a_m \omega^m \nabla u_{h,\hat{\rho}}^{n+1} - \mathcal{I}_p[a_m \omega^m \nabla u_{h,\hat{\rho}}^{n+1}] \Big\|^2_{H,L^2_\rho} \\
&\qquad\qquad + \Big\| \sum_{m=1}^{M} a_m \omega^m \nabla u_{h,\hat{\rho}}^{n} - \mathcal{I}_p[a_m \omega^m \nabla u_{h,\hat{\rho}}^{n}] \Big\|^2_{H,L^2_\rho} \Big) \\
&\leq \sum_{n=0}^{N-1} \triangle t^n \sum_{m=1}^{M} M \Big(\Big\| a_m \omega^m \nabla u_{h,\hat{\rho}}^{n+1} - \mathcal{I}_p[a_m \omega^m \nabla u_{h,\hat{\rho}}^{n+1}] \Big\|^2_{H,L^2_\rho} \\
&\qquad\qquad + \Big\| a_m \omega^m \nabla u_{h,\hat{\rho}}^{n} - \mathcal{I}_p[a_m \omega^m \nabla u_{h,\hat{\rho}}^{n}] \Big\|^2_{H,L^2_\rho} \Big) \\
&= \sum_{m=1}^{M} M \sum_{n=0}^{N-1} \triangle t^n \Big(\Big\| a_m \omega^m \nabla u_{h,\hat{\rho}}^{n+1} - \mathcal{I}_p[a_m \omega^m \nabla u_{h,\hat{\rho}}^{n+1}] \Big\|^2_{H,L^2_\rho} \\
&\qquad\qquad + \Big\| a_m \omega^m \nabla u_{h,\hat{\rho}}^{n} - \mathcal{I}_p[a_m \omega^m \nabla u_{h,\hat{\rho}}^{n}] \Big\|^2_{H,L^2_\rho} \Big).
\end{aligned}
$$

As for the rank truncation error estimator, we proceed by

$$
\begin{aligned}
\varepsilon_{rank} &= \sum_{n=0}^{N-1} \triangle t^n \sum_{k=1}^{\hat{N}} \lambda_k \Big(\sum_{K \in \mathcal{T}_h} \Big\| {\Pi_{\tilde{U}^{n+1} Y^{n^\intercal}}^{h,\hat{\rho}}}^{\perp} \Big[-\nabla \cdot (a_{stoch} \nabla \tilde{U}^{n+1}) Y^{n^\intercal} \\
&\qquad\qquad + \mathbb{E}_{\hat{N}}[\nabla \cdot (a_{stoch} \nabla \tilde{U}^{n+1}) Y^{n^\intercal}] \Big] (\omega_k) \Big\|^2_{L^2(K)} \\
&+ \sum_{E \in \mathscr{E}_h} \Big\| \mathcal{P}_{\mathcal{Y}^n}^{\hat{\rho}\perp} \Big[[a_{stoch} \nabla \tilde{U}^{n+1} \cdot \eta_E]_E Y^{n^\intercal} - \mathbb{E}_{\hat{N}}[[a_{stoch} \nabla \tilde{U}^{n+1} \cdot \eta_E]_E Y^{n^\intercal}] \Big] (\omega_k) \Big\|^2_{L^2(E)} \\
&\qquad\qquad + \| \tilde{U}^{n+1} \tilde{M}^{{n+1}^{-1}} \|^2_H \Big\| \tilde{U}^{{n+1}^\intercal} \mathcal{P}_{\mathcal{Y}^n}^{\hat{\rho}\perp} \Big[[a_{stoch} \nabla \tilde{U}^{n+1} \cdot \eta_E]_E Y^{n^\intercal} \\
&\qquad\qquad\qquad - \mathbb{E}_{\hat{N}}[[a_{stoch} \nabla \tilde{U}^{n+1} \cdot \eta_E]_E Y^{n^\intercal}] \Big] (\omega_k) \Big\|^2_{L^2(E)} \Big) \\
&= \sum_{n=0}^{N-1} \triangle t^n \sum_{k=1}^{\hat{N}} \lambda_k \Big(\sum_{K \in \mathcal{T}_h} \Big\| {\Pi_{\tilde{U}^{n+1} Y^{n^\intercal}}^{h,\hat{\rho}}}^{\perp} \Big[\sum_{m=1}^{M} \Big(-\nabla \cdot (a_m \nabla \tilde{U}^{n+1}) \\
&\qquad\qquad (\omega^m Y^{n^\intercal} - \mathbb{E}_{\hat{N}}[\omega^m Y^{n^\intercal}]) \Big) \Big] (\omega_k) \Big\|^2_{L^2(K)} \\
&\qquad + \sum_{E \in \mathscr{E}_h} \Big\| \mathcal{P}_{\mathcal{Y}^n}^{\hat{\rho}\perp} \Big[\sum_{m=1}^{M} [a_m \nabla \tilde{U}^{n+1} \cdot \eta_E]_E \, (\omega^m Y^{n^\intercal} - \mathbb{E}_{\hat{N}}[\omega^m Y^{n^\intercal}]) \Big] (\omega_k) \Big\|^2_{L^2(E)} \\
&\quad + \| \tilde{U}^{n+1} \tilde{M}^{{n+1}^{-1}} \|^2_H \Big\| \tilde{U}^{{n+1}^\intercal} \mathcal{P}_{\mathcal{Y}^n}^{\hat{\rho}\perp} \Big[[\sum_{m=1}^{M} a_m \nabla \tilde{U}^{n+1} \cdot \eta_E]_E (\omega^m Y^{n^\intercal} - \mathbb{E}_{\hat{N}}[\omega^m Y^{n^\intercal}]) \Big] (\omega_k) \Big\|^2_{L^2(E)} \Big) \\
&= \sum_{n=0}^{N-1} \triangle t^n \sum_{k=1}^{\hat{N}} \lambda_k \Big(\sum_{K \in \mathcal{T}_h} \Big\| {\Pi_{\tilde{U}^{n+1} Y^{n^\intercal}}^{h,\hat{\rho}}}^{\perp} \Big[\sum_{m=1}^{M} \sum_{r=1}^{R^n} -\nabla \cdot (a_m \nabla \tilde{U}_r^{n+1}) \\
&\qquad\qquad\qquad (\omega^m Y_r^n - \mathbb{E}_{\hat{N}}[\omega^m Y_r^n]) \Big] (\omega_k) \Big\|^2_{L^2(K)}
\end{aligned}
$$

$$+ \sum_{E\in\mathscr{E}_h} \Big\|\mathcal{P}^{\hat{\rho}^\perp}_{\mathcal{Y}^n}\Big[\sum_{m=1}^{M}\sum_{r=1}^{R^n}[a_m\nabla\tilde{U}_r^{n+1}\cdot\eta_E]_E\,(\omega^m Y_r^n - \mathbb{E}_{\hat{N}}[\omega^m Y_r^n])\Big](\omega_k)\Big\|^2_{L^2(E)}$$

$$+ \|\tilde{U}^{n+1}\tilde{M}^{{n+1}^{-1}}\|^2_H\Big\|\tilde{U}^{{n+1}^\intercal}\mathcal{P}^{\hat{\rho}^\perp}_{\mathcal{Y}^n}\Big[[\sum_{m=1}^{M}\sum_{r=1}^{R^n}a_m\nabla\tilde{U}_r^{n+1}\cdot\eta_E]_E$$

$$(\omega^m Y_r^n - \mathbb{E}_{\hat{N}}[\omega^m Y_r^n])\Big](\omega_k)\Big\|^2_{L^2(E)}\Bigg)$$

□

The overall error ε can be bounded by

$$\varepsilon^2 \leq \sum_{K\in\mathscr{T}_h}\varepsilon_{spa,K} + \sum_{m=1}^{M}\varepsilon_{stoch,m} + \sum_{n=1}^{N}\varepsilon_{tem,n} + \sum_{n=1}^{N}\varepsilon_{rank,n}. \tag{5.51}$$

Let $\tau = \{t^n\}_{n=1}^N$ denote the time discretization, $\mathscr{P} = \{p_m\}_{m=1}^M$ the set of stochastic polynomial degrees determining the tensor grid, and $\mathscr{R} = \{R^n\}_{n=0}^{N-1}$ the sequence of DLR ranks for every time interval. The algorithm will start with fairly coarse grids $\mathcal{T}_h, \tau$, low rank $R = R^n, \forall n$ and low polynomial orders $p = (p_1, \ldots, p_M)$ determining the tensor grid. We compute the numerical solution $\tilde{u}$ and compute the estimators (5.46), (5.48), (5.47), and (5.49) for every cell, time subinterval, dimension of the stochastic space and time interval. Let $\mathscr{N} = |\mathcal{T}_h| + 2N + M$ denote the total number of elements in the error estimate (5.51), i.e. number of cells + number of subintervals (N) for time discretization + number of dimensions in the stochastic space (M) + number of time intervals (N) for the rank adaptivity. Then we will refine a cell K whenever $\varepsilon_{spa,K} \geq \alpha TOL/\mathscr{N}$, divide a time interval $[t^n, t^{n+1}]$ into 2 equal subintervals whenever $\varepsilon_{tem,n} \geq \alpha TOL/\mathscr{N}$, increase a polynomial order p_m by 1 whenever $\varepsilon_{stoch,m} \geq \alpha TOL/\mathscr{N}$ and increase the DLR rank R^n by 1 whenever $\varepsilon_{rank,n} \geq \alpha TOL/\mathscr{N}$, where $\alpha > 1$. With the new refined mesh, time grid, tensor grid and DLR ranks we compute a new solution $\tilde{u}$ and continue until the stopping criterion

$$\varepsilon_{\mathcal{T}_h,\tau,\mathscr{P},\mathscr{R}} := \sum_{K\in\mathscr{T}_h}\varepsilon_{spa,K} + \sum_{m=1}^{M}\varepsilon_{stoch,m} + \sum_{n=0}^{N-1}\varepsilon_{tem,n} + \sum_{n=0}^{N-1}\varepsilon_{rank,n} < TOL$$

is satisfied. This procedure is described in Algorithm 2. We note that there is no proof of convergence for this algorithm.

As mentioned before, when increasing the rank of $u_{h,\hat{\rho}}^{n+1}$, to proceed with the computation of $u_{h,\hat{\rho}}^{n+2}$ of rank $R^{n+1} > R^n$, we need to choose the basis $\{Y_r^{n+1}\}_{r=R^n+1}^{R^{n+1}}$, so that $\{Y_r^{n+1}\}_{r=1}^{R^{n+1}}$ forms an orthonormal basis in $L^2_{\hat{\rho}}$ and that it leads to an improved approximate solution. There are several options for this choice. We propose to perform a Karhunen-Loève expansion of ${\Pi^{h,\hat{\rho}}_{\tilde{U}^{n+1}Y^{n^\intercal}}}^{\perp}\Big[\mathcal{I}_p[\mathcal{L}^*(\tilde{U}^{n+1}Y^{n^\intercal})]\Big]$, which, as derived in the

Algorithm 2: Adaptive algorithm for DLRA

```
Data: $TOL > 0$
Result: $\mathcal{T}_h, \tau, \mathscr{P}, \mathscr{R}$ and $\tilde{u}$ s.t. $\varepsilon_{\mathcal{T}_h,\tau,\mathscr{P},\mathscr{R}} < TOL$
Initialize $\mathcal{T}_h, \tau, \mathscr{P}, \mathscr{R}$;
compute $\tilde{u}$ on $\mathcal{T}_h, \tau, \mathscr{P}, \mathscr{R}$;
compute $\varepsilon_{spa,K}$, $\varepsilon_{tem,n}$, $\varepsilon_{stoch,m}$, $\varepsilon_{rank,n}$;
while $\varepsilon_{\mathcal{T}_h,\tau,\mathscr{P},\mathscr{R}} \geq TOL$ do
    set $\mathscr{N} = |\mathcal{T}_h| + 2N + M$;
    for $K \in \mathcal{T}_h$ do
        if $\varepsilon_{spa,K} > \alpha \frac{TOL}{\mathscr{N}}$ then
            refine $K$
    for $n \in \{0, \dots, N-1\}$ do
        if $\epsilon_{tem,n} > \alpha \frac{TOL}{\mathscr{N}}$ then
            refine $[t_n, t_{n+1}]$
    for $m \in \{1, \dots, M\}$ do
        if $\epsilon_{sto,m} > \alpha \frac{TOL}{\mathscr{N}}$ then
            $p_m = p_m + 1$
    for $n \in \{0, \dots, N-1\}$ do
        if $\epsilon_{rank,n} > \alpha \frac{TOL}{\mathscr{N}}$ then
            $R^n = R^n + 1$
    update $\mathcal{T}_h, \tau, \mathscr{P}, \mathscr{R}$;
    compute $\tilde{u}$ on new $\mathcal{T}_h, \tau, \mathscr{P}, \mathscr{R}$;
    compute $\varepsilon_{spa,K}$, $\varepsilon_{tem,n}$, $\varepsilon_{stoch,m}$, $\varepsilon_{rank,n}$;
```

proof of Theorem 5.3.1, for a random heat equation with an affine diffusion coefficient and a deterministic forcing term takes the form

$$\begin{aligned}\Big(\Pi^{h,\hat{\rho}}_{\tilde{U}^{n+1}Y^{n^\intercal}}{}^{\perp}\Big[\mathcal{I}_p[\mathcal{L}^*(\tilde{U}^{n+1}Y^{n^\intercal})]\Big], v\Big)_{V'V,L^2_\rho} \\ = \int_\Omega \sum_K \int_K \Pi^{h}_{\tilde{U}^{n+1}Y^{n^\intercal}}{}^{\perp}\Big[\mathcal{I}_p\Big[\Big(-\nabla\cdot(a\nabla\tilde{U}^{n+1})Y^{n^\intercal}\Big)^*\Big]\Big] v\,\mathrm{d}x \\ + \sum_E \int_E \mathcal{P}^{\perp}_{\mathcal{Y}^n}\Big[\mathcal{I}_p\Big[\Big([a\nabla\tilde{U}^{n+1}\cdot\eta_E]_E Y^{n^\intercal}\Big)^*\Big]\Big] v\,\mathrm{d}x \\ - \int_D \tilde{U}^{n+1}\tilde{M}^{n+1^{-1}}\Big(\sum_E \int_E \tilde{U}^{n+1^\intercal}\mathcal{P}^{\perp}_{\mathcal{Y}^n}\Big[\mathcal{I}_p[[a\nabla\tilde{U}^{n+1}\cdot\eta_E]_E Y^{n^\intercal}]^*\Big]\mathrm{d}\hat{x}\Big)\, v\,\mathrm{d}x\,\mathrm{d}\rho.\end{aligned}$$

We then set $\{Y_r^{n+1}\}_{r=R^n+1}^{R^{n+1}}$ to be the $R^{n+1}-R^n$ random eigenvectors corresponding to the $R^{n+1}-R^n$ most dominant singular values of $\Pi^{h,\hat{\rho}}_{\tilde{U}^{n+1}Y^{n^\intercal}}{}^{\perp}\Big[\mathcal{I}_p[\mathcal{L}^*(\tilde{U}^{n+1}Y^{n^\intercal})]\Big]$. Note that $\Pi^{h,\hat{\rho}}_{\tilde{U}^{n+1}Y^{n^\intercal}}{}^{\perp}\Big[\mathcal{I}_p[\mathcal{L}^*(\tilde{U}^{n+1}Y^{n^\intercal})]\Big]$ is at most of rank $R^n\cdot M$. A different approach is to compute $\varepsilon_{rank,n,m,r}$ from (5.50) for all the $R^n\cdot M$ new directions, choose directions with the highest error contribution $\varepsilon_{rank,n,m,r}$ and orthonormalize them. Implementation of this algorithm together with possible further improvements is a part of an ongoing project.

Part II

Dynamical low rank approximation for data assimilation

6 Discrete filtering problem: overview

This chapter provides a mathematical formulation of the filtering problem as well as a brief overview of the commonly used approaches to tackle it. We start by introducing the problem in Section 6.1. In Section 6.2, we continue by describing some standard algorithms used to deal with both linear and nonlinear filtering problems. The last section provides a numerical comparison of the presented methods applied to a 40-dimensional Lorenz-96 chaotic system of equations. We shall highlight that, apart from the numerical experiments, none of the results stated in this chapter are new and we follow to a large extent the book [LSZ15].

6.1 Problem statement

The problem introduced in Section 1.1 was set in an abstract Hilbert space H. Discretizing the system in the physical variable leads to a finite-dimensional system, on which we focus in this part of the thesis.

Let us consider a sequence of N_h-dimensional states $u = \{u^n\}_{n\in\mathbb{N}} \subset \mathbb{R}^{N_h}$, also called a *signal*, defined by the random recursion

$$\begin{aligned} u^{n+1} &= \Psi(u^n) + \xi^n, \quad n \in \mathbb{N}^0 \\ u^0 &\sim N(m^0, C^0), \end{aligned} \tag{6.1}$$

where $\xi = \{\xi^n\}_{n\in\mathbb{N}^0}$ is an i.i.d. sequence with $\xi^0 \sim N(0, \Sigma),\ \Sigma > 0,\ \Sigma \in \mathbb{R}^{N_h\times N_h}$ and accounts for the model error. The operator $\Psi \in C(\mathbb{R}^{N_h}, \mathbb{R}^{N_h})$ is assumed to be deterministic. Further, we assume that u^0 and ξ are independent. In what follows, we denote the joint probability density function of a multivariate Gaussian variable with mean μ and covariance matrix Σ by $\mathbb{R}^{N_h} \ni \xi \to N(\xi; \mu, \Sigma) \in \mathbb{R}^+$. We will use the notation $N(\mu, \Sigma)$ if there is no need to highlight the variable ξ.

At certain time instants $\{t^n\}_{n\in\mathbb{N}}$, we are provided with (complete or incomplete) observations $\{z^n\}_{n\in\mathbb{N}}$ of the signal

$$z^{n+1} = Hu^{n+1} + \eta^{n+1}, \quad n \in \mathbb{N}^0 \tag{6.2}$$

where $H \in \mathbb{R}^{l\times N_h}$, $l \leq N_h$ is a linear operator called an observation operator and $\eta = \{\eta^n\}_{n\in\mathbb{N}} \subset \mathbb{R}^l$ is an i.i.d. sequence, independent of (u^0, ξ), with $\eta^1 \sim N(0,\Gamma)$, $\Gamma > 0$ and accounts for the observation error.

In this work, we assume the initial condition, model error and observation error to be normally distributed. This assumption is not necessarily satisfied in all real-world problems, however many of the techniques considered in this work do rely on this assumption.

In our setting, the function Ψ is the solution operator for a dynamical system of the form

$$\begin{aligned} \dot{u} &= \mathcal{F}(u), \quad t \in (t^n, t^{n+1}) \\ u(t^n) &= u^n, \end{aligned} \tag{6.3}$$

meaning that $\Psi(u^n)$ in (6.1) gives the solution of (6.3) at time t^{n+1}, assuming that the solution exists uniquely.

Let $Z^n = \{z^k\}_{k=1}^n \subset \mathbb{R}^l$ denote the accumulated data up to time n. The discrete filtering problem refers to a sequential update of the probability distribution of the signal, given the data (observations). The objective is to determine $\mathbb{P}(u^n|Z^n)$, the probability density function w.r.t. the Lebesgue measure , associated with the probability measure of the random variable $u^n|Z^n$ (see e.g. [Bau11; LSZ15] for definition of conditional probability distributions). This is done by applying two steps: *forecast (or prediction)* and *analysis*. The forecast step takes in the filtering distribution at time t^n, $\mathbb{P}(u^n|Z^n)$, and through the forward model (6.1) results in the so called forecasted distribution $\mathbb{P}(u^{n+1}|Z^n)$. The analysis step takes in the forecasted distribution $\mathbb{P}(u^{n+1}|Z^n)$ and incorporates the newly observed data z^{n+1} via the Bayes' formula resulting in the posterior filtering distribution $\mathbb{P}(u^{n+1}|Z^{n+1})$. The following formulas form a basis for numerous algorithms used to approximate the sequence of filtering distributions $\mathbb{P}(u^n|Z^n)$, $n \in \mathbb{N}$.

Concerning the forecast step, it holds

$$\mathbb{P}(u^{n+1}|Z^n) = \int_{\mathbb{R}^{N_h}} \mathbb{P}(u^{n+1}|u^n)\mathbb{P}(u^n|Z^n)\,\mathrm{d}u^n. \tag{6.4}$$

Since the probability distribution $\mathbb{P}(u^{n+1}|u^n)$ is determined by the forward model (6.1), the forecast step provides a mapping $\mathbb{P}(u^n|Z^n) \mapsto \mathbb{P}(u^{n+1}|Z^n)$. As for the analysis step, thanks to the Bayes' formula we have that

$$\mathbb{P}(u^{n+1}|Z^{n+1}) = \mathbb{P}(u^{n+1}|Z^n, z^{n+1}) = \frac{\mathbb{P}(z^{n+1}|u^{n+1})\mathbb{P}(u^{n+1}|Z^n)}{\mathbb{P}(z^{n+1}|Z^n)}. \tag{6.5}$$

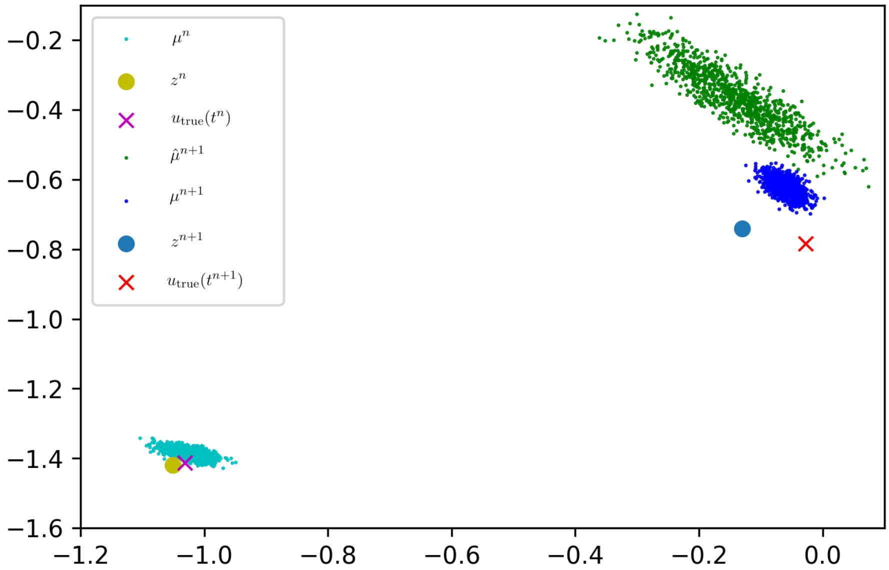


Figure 6.1 – Evolution of an empirical measure of the first two components of a 40-dimensional Lorenz-96 system, computed at the beginning of a forecast step (μ^n), at the end of the forecast step ($\hat{\mu}^{n+1}$) and at the end of the analysis step (μ^{n+1}).

The probability distribution $\mathbb{P}(z^{n+1}|u^{n+1})$ is determined by the observation model (6.2). Therefore, the analysis step provides a mapping $\mathbb{P}(u^{n+1}|Z^n) \mapsto \mathbb{P}(u^{n+1}|Z^{n+1})$. By μ^n we denote the probability measure on $\mathbb{R}^{N_h}$ corresponding to the distribution $\mathbb{P}(u^n|Z^n)$, and $\hat{\mu}^{n+1}$, the probability measure on $\mathbb{R}^{N_h}$ corresponding to $\mathbb{P}(u^{n+1}|Z^n)$. The Figure 6.1 gives an illustration of the measure updates, for a 40-dimensional Lorenz-96 system of equations with full observations (see section 6.3 for more details on Lorenz-96). The continuous measures are here approximated by empirical measures. The illustration depicts the evolution of the first two components of the 40-dimensional state. The filtering measure at time t^n, μ^n (light blue), evolves into a forecasted measure at time t^{n+1}, $\hat{\mu}^{n+1}$ (green), which is then updated by the observations z^{n+1} into a filtering distribution at time t^{n+1}, μ^{n+1} (dark blue). We see that in this particular case, the analysis step reduces the variance by assimilating the observations and moves the measure closer to them.

6.2 Standard algorithms

This section is dedicated to describing various methods used to deal with the discrete filtering problem. We start in subsection 6.2.1 with a linear problem with additive Gaussian noise, for which the Kalman filter provides an exact algorithm. In subsection 6.2.2 we describe how the Kalman filter technique can be extended to efficiently treat nonlinear problems. Lastly, we introduce the particle filter in subsection 6.2.3.

6.2.1 The Kalman filter

Let us consider the scenario where Ψ is an affine operator, i.e. there exists a matrix $M \in \mathbb{R}^{N_h \times N_h}$ and a vector $a \in \mathbb{R}^{N_h}$ which satisfy

$$\Psi(v) = Mv + a \qquad \forall v \in \mathbb{R}^{N_h}.$$

In this case, both the forecasted distribution $\mathbb{P}(u^{n+1}|Z^n)$ and the filtering distribution $\mathbb{P}(u^n|Z^n)$, $\forall n \in \mathbb{N}$ are Gaussian (see e.g. [LSZ15, Lemma 1.5, 1.6] for a proof) and therefore can be entirely characterized through its mean and covariance.

Let (m^n, C^n) denote the mean and the covariance matrix of $u^n|Z^n$ and $(\hat{m}^{n+1}, \hat{C}^{n+1})$ denote the mean and the covariance of $u^{n+1}|Z^n$. The forecasted mean and covariance matrix satisfy

$$\hat{m}^{n+1} = Mm^n + a, \qquad \hat{C}^{n+1} = MC^nM^\intercal + \Sigma. \tag{6.6}$$

The Kalman filter then calculates the exact posterior filtering distribution $\mathbb{P}(u^{n+1}|Z^{n+1})$, characterized by m^{n+1}, C^{n+1}, which are obtained via the formulas from the following lemma.

Lemma 6.2.1. *Assume that* $C^0, \Sigma, \Gamma > 0$. *Then* $C^n > 0$, $\forall n \in \mathbb{N}$, *and*

$$\begin{aligned} m^{n+1} &= \hat{m}^{n+1} + K^{n+1}d^{n+1}, \\ C^{n+1} &= (I - K^{n+1}H)\hat{C}^{n+1}, \end{aligned} \tag{6.7}$$

where

$$\begin{aligned} d^{n+1} &= z^{n+1} - H\hat{m}^{n+1}, \\ S^{n+1} &= H\hat{C}^{n+1}H^\intercal + \Gamma, \\ K^{n+1} &= \hat{C}^{n+1}H^\intercal(S^{n+1})^{-1}. \end{aligned} \tag{6.8}$$

For the proof we refer the reader to [LSZ15, p. 80-81]. The quantity d^{n+1} measures the difference between the observations of the predicted mean and the data and is referred to as the *innovation* at time step $n+1$. The matrix K^{n+1} is called the *Kalman gain*. Note that the matrix inversion in (6.8) takes place in the data space, whose dimension l is, in many applications, much smaller than the state space dimension N_h.

Remark 7. In the linear setting, the Kalman filter update formulas (6.7)–(6.8) can be reinterpreted as solutions to a minimization problem. The update equation for the mean m^{n+1} (6.7) can be rewritten as

$$m^{n+1} = \arg\min_v \mathrm{J}(v) \tag{6.9}$$

$$\mathrm{J}(v) := \frac{1}{2}\Big(\|z^{n+1} - Hv\|_\Gamma^2 + \|v - \hat{m}^{n+1}\|_{\hat{C}^{n+1}}\Big) \tag{6.10}$$

with $\hat{m}^{n+1}, \hat{C}^{n+1}$ given by (6.6) and $\|v\|_K^2$ defined as $\|v\|_K^2 = \langle v, K^{-1}v\rangle$ for K being a symmetric and positive-definite matrix.

6.2.2 Approximate Gaussian filters

Dealing with nonlinear operators Ψ is more challenging. The Kalman filter can be extended to treat nonlinear problems by invoking a Gaussian ansatz in the analysis step of the filter. These approaches are proved to not approximate the filtering distribution correctly in a general setting. We will mention three algorithms, namely *3DVAR*, *extended Kalman filter* and *ensemble Kalman filter*. All of these three algorithms approximate the generally non-Gaussian forecasted distribution by a Gaussian distribution $\mathbb{P}(u^{n+1}|Z^n) \approx N(\hat{m}^{n+1}, \hat{C}^{n+1})$ and the subsequent analysis step applies the standard Kalman formulas from Lemma 6.2.1.

3DVAR

The 3DVAR algorithm simply fixes the model covariance matrix for all time steps $\hat{C}^n = \hat{C}$ using some prior knowledge and evolves the mean as $\hat{m}^{n+1} = \Psi(m^{n+1})$. Note that the role of constant $\hat{C}$ can be interpreted as fixing the regularization term in (6.9)–(6.10) for all n.

We remark that the 3DVAR is a rather computationally inexpensive method, as it only requires to evolve the mean value through the forward model.

Extended Kalman filter

The extended Kalman filter (ExKF) computes the model mean $\hat{m}^{n+1}$ in the same way as 3DVAR, but the model covariance $\hat{C}^{n+1}$ is obtained through a linearization of (6.1). The resulting formulas for the prediction step are

$$\hat{m}^{n+1} = \Psi(m^{n+1}), \qquad \hat{C}^{n+1} = \left(\frac{\partial \Psi}{\partial u}\bigg|_{m^n}\right) C^n \left(\frac{\partial \Psi}{\partial u}\bigg|_{m^n}\right)^{\intercal} + \Sigma. \tag{6.11}$$

The analysis step then applies standard Kalman formulas.

The ExKF relies on a linearization of the forward operator Ψ so that the resulting distribution is at all times Gaussian. For a highly nonlinear Ψ this strategy introduces a large error in the forecast step.

Ensemble Kalman filter

The idea of the ensemble Kalman filter (EnKF) is to propagate a set of particles (called an ensemble of particles), $\{u^n_{(j)}\}_{j=1}^{\hat{N}}$, through the forecast model (6.1). The model mean and covariance $(\hat{m}^{n+1}, \hat{C}^{n+1})$ are then approximated by sample mean and sample covariance estimators using this ensemble

$$\begin{aligned}
\hat{u}^{n+1}_{(j)} &= \Psi(u^n_{(j)}) + \xi^n_{(j)}, \quad j = 1, \dots, \hat{N}, \\
\hat{m}^{n+1} &= \frac{1}{\hat{N}} \sum_{j=1}^{\hat{N}} \hat{u}^{n+1}_{(j)}, \\
\hat{C}^{n+1} &= \frac{1}{\hat{N}-1} \sum_{j=1}^{\hat{N}} (\hat{u}^{n+1}_{(j)} - \hat{m}^{n+1})(\hat{u}^{n+1}_{(j)} - \hat{m}^{n+1})^{\intercal},
\end{aligned}$$

where $\xi^n_{(j)} \overset{\text{iid}}{\sim} \xi^n$ and at time $t = 0$, $u^0_{(j)} \overset{\text{iid}}{\sim} u^0$. The analysis step then applies standard Kalman formulas (6.7)–(6.8) to update the mean and covariance (see [Eve09] for an overview of the methodology, written by one of its founders, and [LE96a] for an early example of the power of the method). To restart the process, one needs to generate a new ensemble of particles. There are many ways of achieving this (see e.g. [BEM01; And01; WH02]). One of the most commonly used is the perturbed observation EnKF, which applies

$$\begin{aligned}
u^{n+1}_{(j)} &= (I - K^{n+1}H)\hat{u}^{n+1}_{(j)} + K^{n+1} z^{n+1}_{(j)}, \quad j = 1, \dots, \hat{N}, \\
z^{n+1}_{(j)} &= z^{n+1} + \eta^{n+1}_{(j)}, \quad j = 1, \dots, \hat{N},
\end{aligned} \tag{6.12}$$

for the analysis step. Here, $\eta^{n+1}_{(j)}$ denote i.i.d. samples from $N(0, \Gamma)$. The matrices S^{n+1} and K^{n+1} are obtained as in (6.8). The algorithm provides update rules of the form

$$\{u^n_{(j)}\}_{j=1}^{\hat{N}} \quad \mapsto \quad \{\hat{u}^{n+1}_{(j)}\}_{j=1}^{\hat{N}} \quad \mapsto \quad \{u^{n+1}_{(j)}\}_{j=1}^{\hat{N}},$$

resulting in approximations to the forecasted and filtering measures by empirical measures

$$\hat{\mu}^{n+1} \approx \frac{1}{\hat{N}} \sum_{j=1}^{\hat{N}} \delta_{\hat{u}^{n+1}_{(j)}}, \qquad \mu^{n+1} \approx \frac{1}{\hat{N}} \sum_{j=1}^{\hat{N}} \delta_{u^{n+1}_{(j)}}.$$

Note, that, except for linear problems, the approximations do not converge to the true distribution μ^n for $\hat{N} \to \infty$. This limitation is overcome by the particle filter introduced in the following subsection. Despite this limitation, the EnKF is still widely used in practice as it often performs well in high-dimensional nonlinear problems.

6.2.3 The particle filter

Similarly to EnKF, the particle filter provides an approximation of the sought (filtering and forecasted) measures by a convex combination of Dirac probability measures (also called empirical measure):

$$\mu^n \approx \sum_{j=1}^{\hat{N}} \lambda^n_{(j)} \delta_{u^n_{(j)}}, \qquad \hat{\mu}^{n+1} \approx \sum_{j=1}^{\hat{N}} \hat{\lambda}^{n+1}_{(j)} \delta_{\hat{u}^{n+1}_{(j)}}, \qquad \sum_{j=1}^{\hat{N}} \hat{\lambda}^{n+1}_{(j)} = \sum_{j=1}^{\hat{N}} \lambda^n_{(j)} = 1,$$

which require the knowledge of the locations and weights $\{u^n_{(j)}, \lambda^n_{(j)}\}_{j=1}^{\hat{N}}$, $\{\hat{u}^{n+1}_{(j)}, \hat{\lambda}^{n+1}_{(j)}\}_{j=1}^{\hat{N}}$ of the associated particles. Note that in this case the weights are not assumed to be uniform. Thus the objective of particle filter algorithms is to propose the update rules

$$\{u^n_{(j)}, \lambda^n_{(j)}\}_{j=1}^{\hat{N}} \quad \mapsto \quad \{\hat{u}^{n+1}_{(j)}, \hat{\lambda}^{n+1}_{(j)}\}_{j=1}^{\hat{N}} \quad \mapsto \quad \{u^{n+1}_{(j)}, \lambda^{n+1}_{(j)}\}_{j=1}^{\hat{N}},$$

for the forecast and the analysis step, respectively.

Unlike the approximate Gaussian filters, the particle filter provides an approximation which converges to the true posterior filtering distribution as $\hat{N} \to \infty$ ([LSZ15, Th. 4.5]). However, particle filters do not perform well in real-world applications and further improvements are necessary.

We will describe the algorithm in its basic form: the *bootstrap filter*.

Forecast
In the forecast step, we keep the weights unchanged $\hat{\lambda}^{n+1}_{(j)} = \lambda^n_{(j)}$ and we let the particle locations evolve through the system (6.1), i.e. $\hat{u}^{n+1}_{(j)} = \Psi(u^n_{(j)}) + \xi^n_{(j)}$, $j = 1, \dots, \hat{N}$, where $\{\xi^n_{(j)}\}_{j=1}^{\hat{N}}$ is a set of i.i.d. samples from the distribution of ξ^n. We thus obtain a particle approximation of the forecasted measure

$$\hat{\mu}^{n+1} \approx \sum_{j=1}^{\hat{N}} \lambda^n_{(j)} \delta_{\hat{u}^{n+1}_{(j)}}.$$

Analysis
The analysis step is performed via the Bayes' formula (6.5) resulting in

$$\mu^{n+1} \approx \sum_{j=1}^{\hat{N}} \lambda^{n+1}_{(j)} \delta_{\hat{u}^{n+1}_{(j)}}, \tag{6.13}$$

where the particle locations stay unchanged whereas the weights are updated as

$$\lambda^{n+1}_{(j)} = \frac{\hat{\lambda}^{n+1}_{(j)}}{\left(\sum_{j=1}^{\hat{N}} \hat{\lambda}^{n+1}_{(j)}\right)}, \qquad \hat{\lambda}^{n+1}_{(j)} = g^n(\hat{u}^{n+1}_{(j)})\, \lambda^n_{(j)}. \tag{6.14}$$

Here, $g^n(u)$ is given by

$$g^n(u^{n+1}) \propto \mathbb{P}(z^{n+1} | u^{n+1}), \tag{6.15}$$

where the constant of proportionality is irrelevant as the weights $\{\hat{\lambda}_{(j)}^{n+1}\}_{j=1}^{\hat{N}}$ are renormalized afterwards. The weights update (6.14) could lead to troublesome scenarios when one of the particle weights approaches 1 and consequently all others approach 0. This phenomenon is often referred to as the degeneracy of the particle filter. It can be partially overcome by resampling, i.e. drawing $\hat{N}$ samples from the measure (6.13) and assigning the weight $\frac{1}{\hat{N}}$ to each of them.

6.2.4 Optimal proposal particle filter

The optimal proposal particle filter (OP-PF) addresses the degeneracy issue in particle filters with the goal of ensuring that all posterior particles have similar weights. The OP-PF does not strictly follow the prediction and analysis paradigm of the 'standard' particle filter. The update of particle locations in the forecast step of the particle filter can be reinterpreted as drawing samples $\hat{u}_{(j)}^{n+1} \sim p(u_{(j)}^n, \cdot)$ from a Markov kernel $p(u^n, u^{n+1}) = \mathbb{P}(u^{n+1}|u^n)$ (see [LSZ15, Sec. 4.3.2.] for more details and e.g. [LSZ15, Sec. 1.4.1.] for more details on Markov kernels and Markov chains) and updating the weights in the analysis step by incorporating the data via the Bayes's law. The optimal proposal aims to improve the proposal distirbution w.r.t. which the new particle locations are drawn by including the data.

The so-called optimal proposal particle filter is found by choosing the Markov kernel

$$p(u^n, u^{n+1}) = \mathbb{P}(u^{n+1}|u^n, z^{n+1})$$

thus, the final update position of the j-th particle is drawn from the posteriori distribution $u_{(j)}^{n+1} \sim \mathbb{P}(\cdot|u_{(j)}^n, z^{n+1})$. Applying Bayes' law twice (see e.g. [Sny12] for more details), one can show that the weight update for the j-th particle drawn from $\mathbb{P}(u_{(j)}^{n+1}|u_{(j)}^n, z^{n+1})$ satisfies

$$\lambda_{(j)}^{n+1} \propto \mathbb{P}(z^{n+1}|u_{(j)}^n)\lambda_{(j)}^n.$$

For general stochastic forward models, obtaining samples from $\mathbb{P}(u_{(j)}^{n+1}|u_{(j)}^n, z^{n+1})$ is not always possible. However, for the case of the forward model considered in this work (6.1), i.e. a deterministic forward operator Ψ with additive Gaussian noise ξ, the distribution and the weights are given in a closed form. The optimal proposal update of each particle is Gaussian with $\mathbb{P}(u_{(j)}^{n+1}|u_{(j)}^n, z^{n+1}) = N(m_{(j)}^{n+1}, Q)$, where

$$\begin{aligned} Q^{-1} &= \Sigma^{-1} + H^\intercal \Gamma^{-1} H, \\ m_{(j)}^{n+1} &= \Psi(u_{(j)}^n) + Q H^\intercal \Gamma^{-1}\Big(z^{n+1} - H\Psi(u_{(j)}^n)\Big). \end{aligned} \tag{6.16}$$

The update of the weights satisfies

$$\hat{\lambda}^{n+1}_{(j)} = \mathbb{P}(z^{n+1}|u^n_{(j)})\lambda^n_{(j)}, \qquad \lambda^{n+1}_{(j)} = \frac{\hat{\lambda}^{n+1}_{(j)}}{\sum_{j=1}^{\hat{N}} \hat{\lambda}^{n+1}_{(j)}}$$

with

$$\mathbb{P}(z^{n+1}|u^n_{(j)}) \propto \exp\Big(-\frac{1}{2}\Big(z^{n+1} - H\Psi(u^n_{(j)})\Big)^\intercal \Big(\Sigma + H\Gamma H^\intercal\Big)^{-1} \Big(z^{n+1} - H\Psi(u^n_{(j)})\Big)\Big).$$

Applying the Woodbury matrix identity [LSZ15, Lemma 4.4], the formulas (6.16) can be rewritten as Kalman formulas

$$Q = (I - KH)\Sigma, \qquad K = \Sigma H^\intercal \Big(H\Sigma H^\intercal + \Gamma\Big)^{-1} \tag{6.17}$$

$$m^{n+1}_{(j)} = \Psi(u^n_{(j)}) + Kd_{(j)}, \qquad d_{(j)} = \Big(z^{n+1} - H\Psi(u^n_{(j)})\Big). \tag{6.18}$$

6.2.5 The Gaussian mixture filter

As mentioned above, generally, the OP-PF does not follow the prediction and analysis steps typical of most filtering algorithms. However, when applied to systems of the form (6.1), it can be reinterpreted as a 2-step algorithm, solving for the forecast and analysis steps. To show this, we first introduce Gaussian mixture models.

Gaussian mixture models (GMMs) provide an attractive framework to approximate unknown distributions based on a set of ensemble realizations. We say that a random vector $u \in \mathbb{R}^{N_h}$ is distributed according to a GMM if

$$\mathbb{P}(u) = \sum_{m=1}^{M} \pi_m \times N(u; m_m, C_m), \qquad \sum_{m=1}^{M} \pi_m = 1. \tag{6.19}$$

Here $M \in \mathbb{N}$ is referred to as the mixture complexity; $\pi_m \in [0,1]$ are called the mixture weights; $m_m \in \mathbb{R}^{N_h}$ are the mixture mean values and $C_m \in \mathbb{R}^{N_h \times N_h}$ are the mixture covariance matrices. An important property of a GMM is that their Bayesian update remains a GMM, if the observation operator is linear and the observation error is Gaussian. This is in fact the scenario of the analysis step in our work, since observations are obtained as (6.2).

Lemma 6.2.2. *Let the forecasted disribution $\mathbb{P}(u^{n+1}|Z^n)$ be a GMM, i.e.*

$$\mathbb{P}(u^{n+1}|Z^n) = \sum_{m=1}^{M} \hat{\pi}^{n+1}_m \times N(u^{n+1}; \hat{m}^{n+1}_m, \hat{C}^{n+1}_m).$$

Then the filtering distribution $\mathbb{P}(u^{n+1}|Z^{n+1})$ *is a GMM*

$$\mathbb{P}(u^{n+1}|Z^{n+1}) = \sum_{m=1}^{M} \pi_m^{n+1} \times N(u^{n+1}; m_m^{n+1}, C_m^{n+1}),$$

where the updated weights, mean values and covariances are given by

$$\begin{aligned} m_m^{n+1} &= \hat{m}_m^{n+1} + K_m(z^{n+1} - H\hat{m}_m^{n+1}), \\ C_m^{n+1} &= (I - K_m H)\hat{C}_m^{n+1}, \\ \pi_m^{n+1} &= \frac{\hat{\pi}_m^{n+1} \cdot g_m(z^{n+1})}{\sum_{i=1}^{M} \hat{\pi}_i^{n+1} \cdot g_i(z^{n+1})}, \end{aligned} \tag{6.20}$$

with K_m *and* $g_m(z)$ *defined as*

$$\begin{aligned} K_m &= \hat{C}_m^{n+1} H^\intercal (H\hat{C}_m^{n+1} H^\intercal + \Gamma)^{-1}, \\ g_m(z) &\propto \exp\Big(-\frac{1}{2}(z - H\hat{m}_m^{n+1})^\intercal (H\hat{C}_m^{n+1} H^\intercal + \Gamma)^{-1} (z - H\hat{m}_m^{n+1})\Big). \end{aligned}$$

Proof. For the details of the proof we refer the reader to [CB01]. □

Note that the matrix K_m in (6.20) is the Kalman gain matrix corresponding to the m-th mixture component. The individual mixture mean values and mixture covariances are updated in accordance with familiar Kalman update formulas (6.8). The coupling occurs only through the weights, which are updated as in the particle filter (6.14).

Now, let us describe the Gaussian mixture filter (GMF) algorithm. Let the filtering measure at time t^n be approximated by an empirical measure

$$\mu^n = \sum_{j=1}^{\hat{N}} \lambda_{(j)}^n \delta_{u_{(j)}^n}.$$

The GMF is realized through a forecast and an analysis step.

Forecast. In the forecast step we first keep the weights unchanged and we let the particle locations evolve through the deterministic forward operator Ψ from (6.1), i.e. $\hat{u}_{(j)}^{n+1} = \Psi(u_{(j)}^n)$, $j = 1, \ldots, \hat{N}$. Incorporating the model error results in a forecasted distribution which is a Gaussian mixture where the mixture covariance matrix is constant across all the mixture components

$$\mathbb{P}(u^{n+1}|Z^n) = \sum_{j=1}^{\hat{N}} \lambda_{(j)}^n \times N(u^{n+1}; \hat{u}_{(j)}^{n+1}, \Sigma). \tag{6.21}$$

Analysis. Applying Lemma 6.2.2 we see that the analysis step results as well in a Gaussian mixture with the mixture covariance matrices constant across all the mixture components:

$$\mathbb{P}(u^{n+1}|Z^{n+1}) = \sum_{j=1}^{\hat{N}} \lambda_{(j)}^{n+1} \times N(u^{n+1}; m_j^{n+1}, C^{n+1}), \tag{6.22}$$

where

$$\begin{aligned} C^{n+1} &= (I - KH)\Sigma, \quad K = \Sigma H^\intercal (H\Sigma H^\intercal + \Gamma)^{-1}, \\ m_j^{n+1} &= \hat{u}_{(j)}^{n+1} + K(z^{n+1} - H\hat{u}_{(j)}^{n+1}), \\ \lambda_{(j)}^{n+1} &= \frac{\lambda_{(j)}^n \cdot g_j(z^{n+1})}{\sum_{m=1}^{\hat{N}} \lambda_{(m)}^n \cdot g_m(z^{n+1})}, \end{aligned} \tag{6.23}$$

with $g_j(z) \propto \exp(-\frac{1}{2}(z - H\hat{u}_{(j)}^{n+1})^\intercal (H\Sigma H^\intercal + \Gamma)^{-1}(z - H\hat{u}_{(j)}^{n+1}))$. To restart the process, one needs to generate a new ensemble of particles. There are many ways of achieving this. Here we propose the following

$$u_{(j)}^{n+1} = m_j^{n+1} + \gamma_{(j)}^{n+1}, \tag{6.24}$$

where $\{\gamma_{(j)}^{n+1}\}_{j=1}^{\hat{N}}$ are i.i.d. samples from $N(0, C^{n+1})$. The resulting empirical filtering distribution $\{\lambda_{(j)}^{n+1}, u_{(j)}^{n+1}\}_{j=1}^{\hat{N}}$ is, in fact, equivalent to the distribution resulting from the optimal proposal particle filter (6.16), which means that the proposed GMF is the OP-PF. This is true specifically for the case of deterministic operator Ψ, normally distributed additive model and observation error and linear observation operator.

6.3 Numerical example: Lorenz-96

The preceding section lead us through some of the most standard methods applied to deal with the discrete filtering problem. In this section, we introduce a numerical example which will be used throughout the whole work as a test case for comparison of different filtering algorithms.

The performance of the aforementioned algorithms, as well as the performance of the new-proposed techniques provided in the next chapter, will be tested on a guiding numerical example, the 40-dimensional Lorenz-96 dynamical system ([Lor96]). This system of equations is a simplified model of the chaotic nature of atmospheric processes, and therefore a popular test case to assess the performance of filtering procedures in turbulent dynamics. The equations take the form

$$\begin{aligned} &\dot{u}_i = u_{i-1}(u_{i+1} - u_{i-2}) - u_i + F, \quad i = 1, \dots, 40 \\ &u_0 = u_{40},\ u_{41} = u_1,\ u_{-1} = u_{39}. \end{aligned} \tag{6.25}$$

This is a specific case of a more general quadratic system formulation

$$\dot{u} = (L + D)u + B(u, u) + F, \tag{6.26}$$

where L is a skew-symmetric linear operator, i.e. $L^{\intercal} = -L$, D is a negative definite symmetric operator $D^{\intercal} = D$. The quadratic term B conserves energy which means $\langle B(u, u), u\rangle = 0$ and F is a forcing term. Many turbulent dynamical systems in geosciences and engineering have such structure [Sal98; MX06].

To asses the performance of the various algorithms, we show the behavior of the RMSE (root mean square error) of the signal u^n over time, defined as

$$RMSE(t^n) = \sqrt{\frac{1}{N_h \hat{N}} \sum_{j=1}^{\hat{N}} \|u^n_{(j)} - u_{\text{true}}(t^n)\|^2}.$$

For the numerical experiment in this section, we set the following parameters: final time $T = 30$, time between observations $\triangle t = 0.05$, the model error covariance matrix $\Sigma = 10^{-4} \cdot \text{Id}$, the observation error covariance matrix $\Gamma = 10^{-2} \cdot \text{Id}$ and the observation operator $H = \text{Id}$. The data $\{z^n\}_{n\in\mathbb{N}}$ is obtained synthetically, i.e. there is a true signal u_{true} satisfying (6.1), for which $z^n = Hu_{\text{true}}(t^n) + \eta$, with η being a sample of η^n.

We focus on the 3DVAR, extended Kalman filter (ExKF) and the ensemble Kalman filter (EnKF) as examples of the approximate Gaussian filters from Section 6.2.2, and the bootstrap filter (PF) and the Gaussian mixture filter (GMF) as examples of the particle filters (in our setting GMF is equivalent to the optimal proposal particle filter and so we consider it to be an example of a particle filter). The behaviour of their RMSEs is depicted in Figure 6.2 and Figure 6.3. Since full observations are available, an algorithm performs reliably if the RMSE is below the observation noise (otherwise one can consider a trivial filtering with $u^{n+1} = z^{n+1}$ resulting in RMSE being equal to the observation noise).

We see from the figure that the performance of the 3DVAR depends on the choice of the constant covariance matrix $\hat{C}$. Choosing $\hat{C} = 10^{-4} \cdot \text{Id}$ expresses a rather high confidence on the capability of the mean value to represent well the predicted measure $\hat{\mu}$. For a highly non-linear system, like the Lorenz-96, this is not justified and the algorithm looses track of the signal very quickly. Choosing $\hat{C} = 1. \cdot \text{Id}$, on the contrary, expresses a high uncertainty in the predicted measure. The updated filtering distribution relies mostly on the information from the data and consequently the RMSE does not improve upon the observation noise. Finally, setting $\hat{C} = 10^{-2} \cdot \text{Id}$ provides satisfactory results. We conclude that in this test case, 3DVAR does manage to track the signal, but an a-priori insight into the choice of the constant covariance matrix is necessary. As for the extended Kalman filter, the method manages to track the signal at the very beginning of the experiment but then eventually loses track.

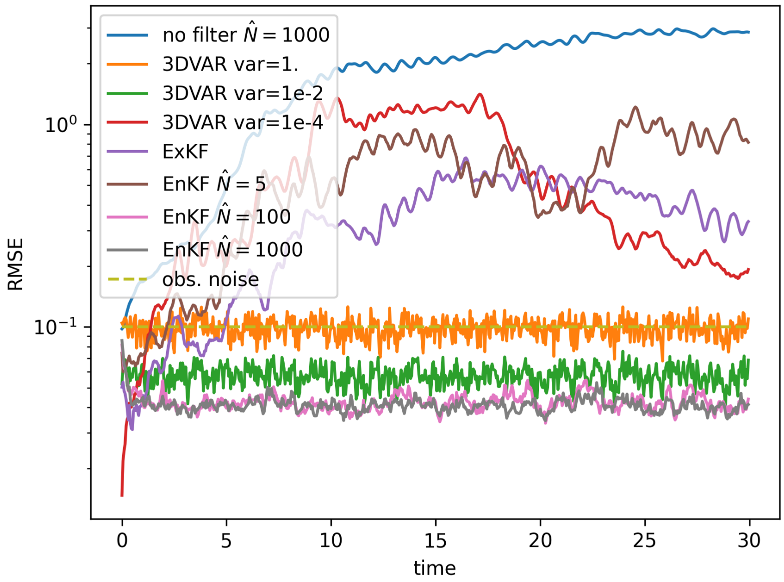


Figure 6.2 – Comparison of RMSEs when applying approximate Gaussian filters, namely 3DVAR with different covariance matrices, ExKF and EnKF with different number of particles.

It comes as no surprise that the ensemble Kalman filter deals with the problem much more efficiently, as it involves evolving a set of particles through a full non-linear system. Its performance depends on the number of considered particles. In Figure 6.2 we see a clear improvement when increasing $\hat{N}$, where in fact $\hat{N} = 100$ is sufficient to track the signal. Although the EnKF is not consistent with the true filtering distribution, as long as the signal is tracked, the filtering distribution remains well concentrated around the 'true value' and a Gaussian approximation turns out to perform well.

The particle filters are proven to provide an approximation which converges to the true filtering distribution in the large-particle limit. However, in Figure 6.3 we see that to avoid degeneracy, the most basic form, the bootstrap filter, requires evolving a high number of particles $\hat{N} = 1000$. This demand is weakened by the optimal-proposal particle filter (or the Gaussian mixture filter), for which $\hat{N} = 100$ is sufficient.

As a last test case, we provide a comparison of all algorithms: 3DVAR, ExKF, EnKF, PF, GMF to a problem with partial observations. We set the following parameters: final time $T = 30$, time between observations $\triangle t = 0.05$, the model error covariance matrix $\Sigma = 10^{-2} \cdot \mathrm{Id}$, the observation error covariance matrix $\Gamma = 10^{-2} \cdot \mathrm{Id}$. The observation operator observes every second value of the 40-dimensional signal, i.e. $l = 20$. We note that in the case of partial observations, we do not expect the RMSE of any of

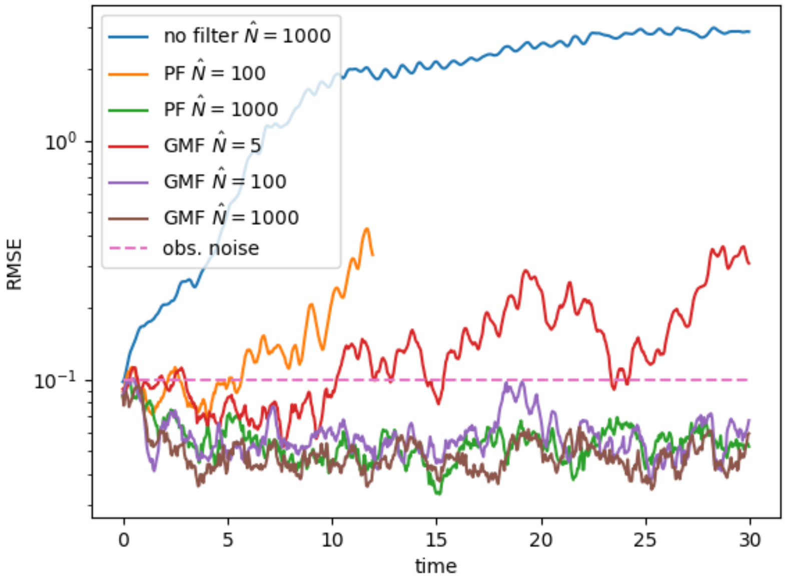


Figure 6.3 – Comparison of RMSEs when applying particle filters, namely bootstrap filter (PF) and optimal proposal particle filter (GMF), with different number of particles.

the algorithms to be necessarily below the observation noise. In Figure 6.4 we see that indeed, none of the filters managed to keep the error below the observation noise. The best results were achieved by the EnKF and GMF.

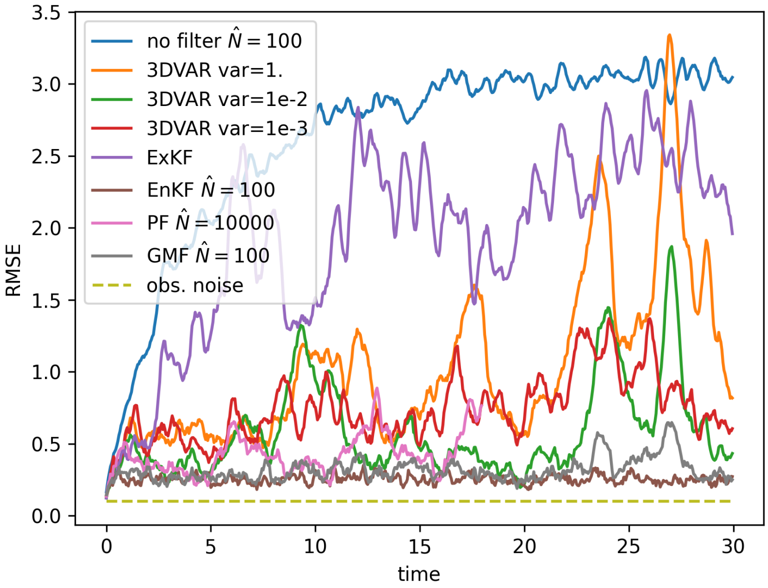


Figure 6.4 – Comparison of RMSEs when applying approximate Gaussian filters and particle filters to a problem with partial observations ($l = 20$).

7 Dynamical low rank approximation for filtering problems

A concern with nonlinear data assimilation schemes as those presented in the previous chapter is their difficulty in handling the dimensionality of the state space N_h, which can be very high in many atmospheric and geophysical applications. When running a full-order model is extremely expensive and only a very few runs (particles) can be performed in the forecast step, one needs to rely on reduced-order modeling techniques. These typically consist in looking for a solution in a low-dimensional subspace. However, especially in the context of data assimilation, the "optimal" subspace that approximates well the whole solution (or a large ensemble of particles) can significantly vary in time. In this respect, employing the dynamical low rank approximation (DLRA) in the forecast step seems very advantageous. The dominant subspace evolves in time, adjusting to the underlying dynamical system at every time instant as well as the incoming observations. In this chapter, we will first explore the idea of applying simple DLRA in the forecast step, combined with standard algorithms (introduced in Section 6.2) in the analysis step. In Section 7.1, we will see that keeping the signal in a simple DLR form and completely dismissing the omitted modes in the DLRA can result in unsatisfactory scenarios, as shown in the numerical experiments presented in Section 7.1.3. In Section 7.2 we will propose two new algorithms, which complement the signal in the DLR form by a Gaussian component. In Section 7.2.5, their performance will be tested on a Lorenz-96 system of equations (see Section 6.3 for the definition of the test case). We would like to point out that all the work presented in this chapter is original.

7.1 Simple dynamical low rank approximation in the forecast step

We will start our exploration of applying DLRA in filtering problems with the most straightforward idea: replacing the full-order model in the forecast step by a signal in the DLR form and performing the analysis step using some of the standard algorithms introduced in Section 6.2.

First, we recall the DLR formulation applied to our setting. Let us consider an abstract probability triple $(\Omega, \mathcal{S}, \mu)$, where Ω is the sample space characterizing all uncertainty in the model, i.e. model error, observation error and initial condition.

We will denote by $\omega \in \Omega$ an elementary event and the space of all square integrable random functions taking values in $\mathbb{R}^m$, $m \in \mathbb{N}$, will be denoted by $L^2(\Omega; \mathbb{R}^m)$

$$L^2(\Omega; \mathbb{R}^m) = \{f : \Omega \to \mathbb{R}^m,\ \mathcal{S}\text{-measurable} \mid \int_\Omega \|f(\omega)\|^2 \,\mathrm{d}\mu(\omega) < \infty\},$$

where $\|\cdot\| = \langle \cdot, \cdot \rangle$ is the standard Euclidean norm in $\mathbb{R}^m$. By $L^2_0(\Omega; \mathbb{R}^m) \subset L^2(\Omega; \mathbb{R}^m)$ we denote its subspace of all functions with zero mean.

7.1.1 Forecast step

We are looking for a signal $u^{\mathrm{DLR}} \in L^2(\Omega; \mathbb{R}^{N_h})$, approximating the solution of (6.3) for $t \in (t^n, t^{n+1})$ in the form

$$u_{\mathrm{true}}(t) \approx u^{\mathrm{DLR}}(t) = \bar{u}^{\mathrm{DLR}}(t) + \sum_{r=1}^{R} U_r(t) Y_r(t), \tag{7.1}$$

where $\bar{u}^{\mathrm{DLR}}$ is the mean value of u^{DLR}, $U = (U_1, \ldots, U_R)$ denotes a deterministic basis orthonormal w.r.t. the Euclidean inner product $\langle \cdot, \cdot \rangle$ and $Y = (Y_1, \ldots, Y_R)$ are zero mean random coefficients with bounded second moments, for which the covariance matrix $C_Y := \mathbb{E}[Y^\intercal Y]$ is full rank. The rank R is kept constant in time. Note that in this work we consider the DLR form where the deterministic basis U is kept orthonormal, as opposed to Chapter 2, in which the stochastic basis Y was kept orthonormal.

Initial condition

As the first step, we describe how to obtain the initial condition at time t^n in the form (7.1) starting from an arbitrary signal $u^n \in L^2(\Omega; \mathbb{R}^{N_h})$. Throughout the rest of the work, we use the notation $f^* = f - \mathbb{E}[f]$. To start with, the Karhunen–Loève expansion of u^{n*} results in

$$u^{n*} = \sum_{k=1}^{N_h} \xi_k e_k,$$

where $\{e_k\}_{k=1}^{N_h} \subset \mathbb{R}^{N_h}$ are pairwise orthonormal vectors in $\mathbb{R}^{N_h}$ and $\xi = \{\xi_1, \ldots, \xi_{N_h}\}$ are zero mean, pairwise uncorrelated random variables ordered by decreasing variance. We can then set

$$\begin{gathered} u^{\mathrm{DLR},n} = \bar{u}^{\mathrm{DLR},n} + \sum_{r=1}^{R} U_r^n Y_r^n \quad \text{with} \\ \bar{u}^{\mathrm{DLR},n} = \mathbb{E}[u^{\mathrm{DLR},n}], \quad U_r^n = e_r,\ Y_r^n = \xi_r, \quad \forall 1 \le r \le R. \end{gathered} \tag{7.2}$$

In the rest of the work we will apply to following vector notation $U = (U_1, \dots, U_R) \in \mathbb{R}^{N_h \times R}$, $Y = (Y_1, \dots, Y_R) \in L^2(\Omega; \mathbb{R}^R)$. By $\mathcal{P}_U[v]$ we denote the projection

$$\mathcal{P}_U[v] := \sum_{i=1}^{R} U_i \langle U_i, v \rangle, \quad v \in \mathbb{R}^{N_h}.$$

Definition 7.1.1. We define the DLR solution of the problem (6.3) as

$$u^{\mathrm{DLR}}(t,\omega) = \bar{u}^{\mathrm{DLR}}(t) + \sum_{r=1}^{R} U_r(t) Y_r(t,\omega) = \bar{u}^{\mathrm{DLR}}(t) + U(t) Y^{\intercal}(t,\omega),$$

where $\bar{u}^{\mathrm{DLR}}$ is the solution of

$$\dot{\bar{u}}^{\mathrm{DLR}} = \mathbb{E}[\mathcal{F}(u^{\mathrm{DLR}})], \qquad t \in (t^n, t^{n+1}), \tag{7.3}$$

and $\{U_r\}_{r=1}^R, \{Y_r\}_{r=1}^R$ are solutions of the following variational formulation for $t \in (t^n, t^{n+1})$

$$\begin{aligned}
\mathbb{E}\Big[\Big\langle \dot{U} Y^{\intercal} + U \dot{Y}^{\intercal},\, v \Big\rangle\Big] = \mathbb{E}\Big[\Big\langle \mathcal{F}^*\Big(u^{\mathrm{DLR}}\Big),\, v \Big\rangle\Big]& \\
\forall v \in \{w = \delta U Y^{\intercal} + U \delta Y^{\intercal} \text{ with } \delta U \in \mathbb{R}^{N_h \times R},\ \langle \delta U^{\intercal}, U \rangle = 0& \\
\delta Y \in L_0^2(\Omega; \mathbb{R}^R)\},&
\end{aligned} \tag{7.4}$$

with the initial conditions $\bar{u}^{\mathrm{DLR}}(t^n), \{U_r(t^n)\}_{r=1}^R$ and $\{Y_r(t^n)\}_{r=1}^R$ obtained as described in (7.2).

For completeness, we specify the detailed evolution equations for the DLR modes in the following theorem.

Theorem 7.1.1. *The variational formulation* (7.4) *results in the following system of equations for the mean value* $\dot{\bar{u}}^{\mathrm{DLR}}$ *and the DLR modes* $\{U_r, Y_r\}_{r=1}^R$*:*

$$\begin{aligned}
\dot{\bar{u}}^{\mathrm{DLR}} &= \mathbb{E}[\mathcal{F}(u^{\mathrm{DLR}})], \\
\dot{Y}_r &= \Big\langle U_r, \mathcal{F}^*(u^{\mathrm{DLR}}) \Big\rangle, \quad r = 1, \dots, R, \\
\sum_{l=1}^{R} (C_Y)_{lr} \dot{U}_l &= \Big(\mathbb{E}\Big[\mathcal{F}^*(u^{\mathrm{DLR}}) Y_r\Big] - U \langle U, \mathbb{E}\Big[\mathcal{F}^*(u^{\mathrm{DLR}}) Y_r\Big] \rangle \Big), \quad r = 1, \dots, R, \\
&\qquad\qquad \textit{with } C_Y = \mathbb{E}[Y^{\intercal} Y].
\end{aligned} \tag{7.5}$$

Proof. The proof can be found in [SL09]. □

The evolving measure is approximated by an empirical measure $\{u^{\mathrm{DLR}}_{(j)}, \lambda_{(j)}\}_{j=1}^{\hat{N}}$ with

$$u^{\mathrm{DLR}}_{(j)} = \bar{u}^{\mathrm{DLR}} + \sum_{r=1}^{R} U_r Y_{r,(j)}. \tag{7.6}$$

The weights do not evolve in time. In the computation of (7.5), we replace all expectations $\mathbb{E}[\cdot]$ by the sample averages $\mathbb{E}_{\hat{N}}[\cdot]$, that is

$$\mathbb{E}_{\hat{N}}[\mathcal{F}(u^{\mathrm{DLR}})] = \sum_{j=1}^{\hat{N}} \lambda_{(j)} \mathcal{F}(u^{\mathrm{DLR}}_{(j)}) \quad \text{and} \quad \mathbb{E}_{\hat{N}}[\mathcal{F}^*(u^{\mathrm{DLR}})Y_r] = \sum_{j=1}^{\hat{N}} \lambda_{(j)} \mathcal{F}^*(u^{\mathrm{DLR}}_{(j)}) Y_{r,(j)}.$$

The resulting system of equations discretized in the random variable is

$$\begin{aligned}
\dot{\bar{u}}^{\mathrm{DLR}} &= \mathbb{E}_{\hat{N}}[\mathcal{F}(u^{\mathrm{DLR}})], \\
\dot{Y}_{r,(j)} &= \left\langle U_r, \mathcal{F}^*(u^{\mathrm{DLR}}_{(j)}) \right\rangle, \quad r = 1, \dots, R, \\
\sum_{l=1}^{R} (C_Y)_{lr} \dot{U}_l &= \left(\mathbb{E}_{\hat{N}}\left[\mathcal{F}^*(u^{\mathrm{DLR}})Y_r\right] - U\langle U, \mathbb{E}_{\hat{N}}\left[\mathcal{F}^*(u^{\mathrm{DLR}})Y_r\right]\rangle \right), \quad r = 1, \dots, R, \\
&\qquad \text{with } C_Y = \mathbb{E}_{\hat{N}}[Y^{\mathsf{T}} Y].
\end{aligned} \tag{7.7}$$

By Ψ^{DLR} we will denote the DLR approximation of the forward operator Ψ from (6.1), i.e. an operator that takes in a set of $\hat{N}$ particles and weights characterizing the filtering distribution at time t^n: $u^{\mathrm{DLR},n} = \{u^{\mathrm{DLR},n}_{(j)}, \lambda^n_{(j)}\}_{j=1}^{\hat{N}}$, and applies the DLR method to approximate the evolution of the system (6.3) on (t^n, t^{n+1}), returning a set of forecasted particles and weights $\tilde{u}^{\mathrm{DLR},n+1} = \{u^{\mathrm{DLR}}_{(j)}(t^{n+1}), \lambda^{n+1}_{(j)}\}_{j=1}^{\hat{N}}$, $\tilde{u}^{\mathrm{DLR},n+1} = \Psi^{\mathrm{DLR}}(u^{\mathrm{DLR},n})$.

7.1.2 Model error and analysis step: ensemble Kalman filter and particle filter

By $\hat{\bar{u}}^{\mathrm{DLR},n+1}, \hat{U}^{n+1}$ let us denote the mean value and deterministic basis, respectively, resulting from (7.5) at time t^{n+1}. The last part of the forecast step involves incorporating the model error. The standard particle or ensemble Kalman filter adds $\hat{N}$ independent samples $\{\xi^n_{(j)}\}_{j=1}^{\hat{N}}$ of the model error to the particle locations obtained from evolving the system. Since generally $\{\xi^n_{(j)}\}_{j=1}^{\hat{N}} \subset \mathbb{R}^{N_h}$ are not restricted to any subspace, this process would result in losing the low-rank structure (7.6) of the signal. Instead, we add samples from the model error projected on the subspace spanned by $\hat{U}^{n+1}$. The resulting forecasted distribution is characterized by the set of particles $\hat{u}^{\mathrm{DLR},n+1} = \{\hat{u}^{\mathrm{DLR},n+1}_{(j)}\}_{j=1}^{\hat{N}}$, which satisfy

$$\hat{u}^{\mathrm{DLR},n+1}_{(j)} = \left(\Psi^{\mathrm{DLR}}(u^{\mathrm{DLR},n})\right)_{(j)} + \mathcal{P}_{\hat{U}^{n+1}}[\xi^n_{(j)}], \qquad j = 1, \dots, \hat{N},$$

where $\mathcal{P}_{\hat{U}^{n+1}}[\xi^n_{(j)}] = \hat{U}^{n+1}\langle \hat{U}^{n+1}, \xi^n_{(j)}\rangle$. Note that the projected model error samples can be alternatively obtained as $\xi^n_{(j)} = \hat{U}^{n+1}\chi^n_{(j)}$ with $\chi^n_{(j)} \sim N(0, \Sigma_{\hat{U}^{n+1}})$ with $\Sigma_{\hat{U}^{n+1}} = \hat{U}^{n+1\intercal}\Sigma\hat{U}^{n+1} \in \mathbb{R}^{R\times R}$.

It is worth keeping in mind that, as opposed to EnKF or particle filter, where the particles are evolved independently of each other and can be computed in parallel, the operator Ψ^{DLR} requires the knowledge of all particle locations and consequently creates correlations among them.

To incorporate the observation, we apply standard algorithms presented in Section 6.2 that deal with empirical measures, namely ensemble Kalman filter (EnKF) and particle filter. The EnKF computes the sample mean and sample covariance using the forecasted set of particles and updates the position of every particle via the Kalman formulas (6.12). The particle filter updates the weights by formulas derived directly from the Bayes' formula (see (6.13)), which is often followed by a resampling step. Both of these algorithms lead to an updated empirical filtering distribution $\mathbb{P}(u^{n+1}|Z^{n+1})$ given by $\{\lambda^{n+1}_{(j)}, u^{\mathrm{DLR},n+1}_{(j)}\}_{j=1}^{\hat{N}}$. We will present two lemmas that highlight the advantages and disadvantages of the analysis step performed in the aforementioned way.

The forecasted particles are of the form

$$\hat{u}^{\mathrm{DLR},n+1}_{(j)} = \hat{\bar{u}}^{\mathrm{DLR},n+1} + \sum_{r=1}^{R} \hat{U}^{n+1}_r \hat{Y}^{n+1}_{r,(j)}, \tag{7.8}$$

where $\hat{Y}^{n+1}_r$, $r = 1, \ldots, R$ are the stochastic coefficients Y_r resulting from (7.5) at time t^{n+1} summed with the projected model error samples $(\xi^n)_{\hat{U}^{n+1}_r} := \langle \hat{U}^{n+1}_r, \xi^n\rangle$.

Lemma 7.1.2. *Performing the ensemble Kalman filter update with perturbed observations in the full state space with the forecasted distribution given by* (7.8) *and observations* z^{n+1} *is equivalent to applying*

$$u^{\mathrm{DLR},n+1}_{(j)} = \hat{\bar{u}}^{\mathrm{DLR},n+1} + \sum_{r=1}^{R} \hat{U}^{n+1}_r Y^{n+1\intercal}_{r,(j)} \tag{7.9}$$

where $Y^{n+1\intercal}_{r,(j)}$ *are obtained by ensemble Kalman filter update formulas in an* R*-dimensional stochastic subspace via*

$$\begin{aligned}
\tilde{H} &:= H\hat{U}^{n+1}, \quad C_{\hat{N},\hat{Y}^{n+1}} = \mathbb{E}_{\hat{N}}[\hat{Y}^{n+1\intercal}\hat{Y}^{n+1}] \\
\tilde{S}^{n+1} &= \tilde{H}C_{\hat{N},\hat{Y}^{n+1}}\tilde{H}^\intercal + \Gamma, \\
\tilde{K}^{n+1} &= C_{\hat{N},\hat{Y}^{n+1}}\tilde{H}^\intercal(\tilde{S}^{n+1})^{-1} \\
Y^{n+1\intercal}_{r,(j)} &= (I - \tilde{K}^{n+1}\tilde{H})\hat{Y}^{n+1\intercal}_{r,(j)} + \tilde{K}^{n+1}z^{n+1}_{(j)}, \quad j = 1, \ldots, \hat{N};\ r = 1, \ldots, R, \\
z^{n+1}_{(j)} &= z^{n+1} - H\hat{\bar{u}}^{\mathrm{DLR},n+1} + \eta^{n+1}_{(j)}, \quad j = 1, \ldots, \hat{N}.
\end{aligned} \tag{7.10}$$

Proof. We start by noticing the following relations. The forecasted covariance matrix satisfies

$$\hat{C}^{n+1} = \hat{U}^{n+1} C_{\hat{N},\hat{Y}^{n+1}} \hat{U}^{n+1\intercal},$$

and for the Kalman gain it holds $K^{n+1} = \hat{U}^{n+1}\tilde{K}^{n+1}$, which from the orthonormality of $\hat{U}^{n+1}$ gives

$$\tilde{K}^{n+1} = \hat{U}^{n+1\intercal} K^{n+1}.$$

We follow by deriving

$$\begin{aligned}
u_{(j)}^{\mathrm{DLR},n+1} &= \hat{\bar{u}}^{\mathrm{DLR},n+1} + \hat{U}^{n+1} Y_{(j)}^{n+1\intercal} \\
&= \hat{\bar{u}}^{\mathrm{DLR},n+1} + \hat{U}^{n+1}\Big((I - \tilde{K}^{n+1}\tilde{H})\hat{Y}_{(j)}^{n+1\intercal} + \tilde{K}^{n+1} z_{(j)}^{n+1}\Big) \\
&= \hat{\bar{u}}^{\mathrm{DLR},n+1} + \hat{U}^{n+1}\hat{Y}_{(j)}^{n+1\intercal} + \hat{U}^{n+1}\tilde{K}^{n+1}\Big(z_{(j)}^{n+1} - \tilde{H}\hat{Y}_{(j)}^{n+1\intercal}\Big) \\
&= \hat{\bar{u}}^{\mathrm{DLR},n+1} + \hat{U}^{n+1}\hat{Y}_{(j)}^{n+1\intercal} + K^{n+1}\Big(z^{n+1} + \eta_{(j)}^{n+1} - H\hat{\bar{u}}^{\mathrm{DLR},n+1} - \tilde{H}\hat{Y}_{(j)}^{n+1\intercal}\Big) \\
&= \hat{u}_{(j)}^{\mathrm{DLR},n+1} + K^{n+1}\Big(z^{n+1} + \eta_{(j)}^{n+1} - H\hat{\bar{u}}^{\mathrm{DLR},n+1} - H\hat{U}^{n+1}\hat{Y}_{(j)}^{n+1\intercal}\Big) \\
&= \hat{u}_{(j)}^{\mathrm{DLR},n+1} + K^{n+1}\Big(z^{n+1} + \eta_{(j)}^{n+1} - H\hat{u}_{(j)}^{\mathrm{DLR},n+1}\Big),
\end{aligned}$$

i.e. we recovered exactly the formulas for the ensemble Kalman filter analysis step introduced in (6.12). ☐

We stress that being able to carry out the analysis step within the subspace is very favourable. The EnKF analysis step in the full state space requires computing the covariance matrix $\hat{C}^{n+1} \in \mathbb{R}^{N_h \times N_h}$, which for a high N_h is unfeasible. We shall note that for all the quantities in (7.10) it holds

$$\tilde{H} \in \mathbb{R}^{l\times R}, \quad C_{\hat{N},\hat{Y}^{n+1}} \in \mathbb{R}^{R\times R}, \quad \tilde{S}^{n+1} \in \mathbb{R}^{l\times l}, \quad \tilde{K}^{n+1} \in \mathbb{R}^{R\times l},$$

i.e. the update formulas do not see the squared full-state dimension $N_h \times N_h$, if $l \ll N_h$.

Corollary 1. The particles $\{u_{(j)}^{\mathrm{DLR},n+1}\}_{j=1}^{n+1}$ approximating the filtering distribution $\mathbb{P}(u^{n+1}|Z^{n+1})$ obtained by either EnKF or particle filter formulas satisfy

$$u_{(j)}^{\mathrm{DLR},n+1} \in \hat{\bar{u}}^{\mathrm{DLR},n+1} \oplus \mathrm{span}(\hat{U}_1^{n+1}, \dots, \hat{U}_R^{n+1}), \qquad \forall j = 1, \dots, \hat{N},$$

i.e. the mean value as well as the R-dimensional DLR subspace spanning the particle locations is not updated in the analysis step

$$\bar{u}^{\mathrm{DLR},n+1} = \hat{\bar{u}}^{\mathrm{DLR},n+1}, \quad U_r^{n+1} = \hat{U}_r^{n+1}, \qquad \forall r = 1, \dots, R.$$

Proof. For the EnKF the statement is a simple consequence of Lemma 7.1.2. For particle filter it comes from the fact that the locations of the particles are not updated in the analysis step and the resampling step does not enlarge the subspace spanning the particle

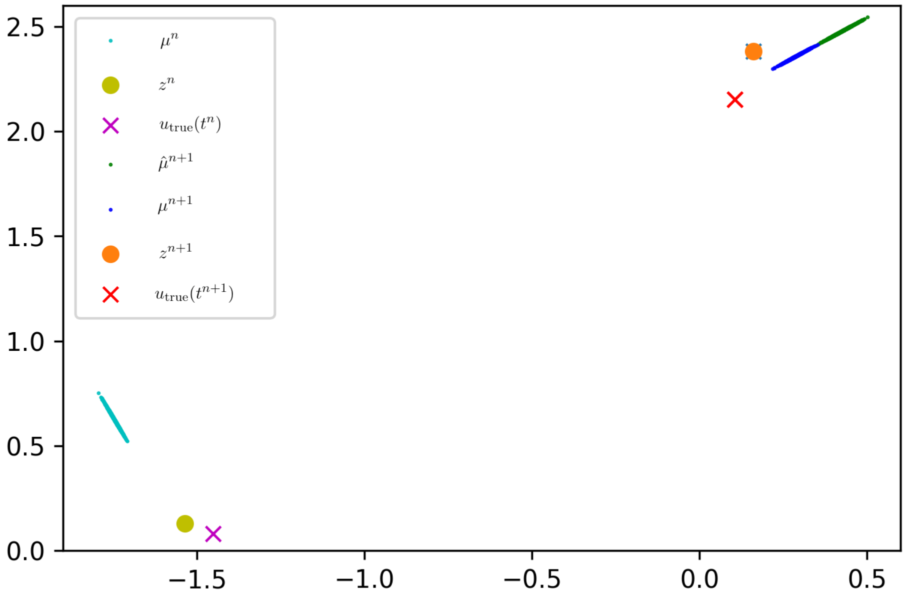


Figure 7.1 – Evolution of an empirical measure of the first two components of a Lorenz-96 system, computed at the beginning of a forecast step (μ^n), at the end of the forecast step ($\hat{\mu}^{n+1}$) and at the end of the analysis step (μ^{n+1}). The forecast step was computed via DLRA with $R = 1$.

locations. □

This behavior is clearly illustrated in Figure 7.1, which shows the evolution of the empirical measure of the first two components of a signal, using DLR with $R = 1$. All measures concentrate along lines (1-dimensional subspaces), since the particle locations are spanned by a 1-dimensional DLR subspace. We see that the measure at the beginning of the forecast step μ^n (lightblue) evolves into $\hat{\mu}^{n+1}$ (green). This update changes the subspace U which is depicted as a change in the corresponding line. The analysis step then updates the forecasted measure $\hat{\mu}^{n+1}$ into μ^{n+1} (dark blue) without updating the DLR subspace, nor the mean value, i.e. the line along which the two measures are spread remains unchanged.

We conclude that not having to update the R-dimensional subspace is advantageous in terms of computational complexity. However, in the following lemma we can see the drawback of not updating the subspace, at least in the case of full observations and an observation error with a specified distribution.

Lemma 7.1.3. *Consider a case with full observations* $H = \mathrm{Id}$. *Let us assume that there* *is a state* u_{true}*, such that*

$$z^{n+1} = H\, u_{\text{true}}(t^{n+1}) = u_{\text{true}}(t^{n+1}).$$

By $\mathcal{P}^{\Gamma}_{\hat{U}^{n+1}}[u_{\text{true}}(t^{n+1})]$ *we denote the orthogonal projection of* $u_{\text{true}}(t^{n+1})$ *on* $\operatorname{span}(\hat{U}_1^{n+1}, \ldots, \hat{U}_R^{n+1})$ *w.r.t. the inner product* $\langle u, v\rangle_\Gamma := \langle u, \Gamma^{-1} v\rangle$ *induced by the covariance matrix* Γ*. In addition, by* z_R^{n+1} *we denote the observation associated to the projected state*

$$z_R^{n+1} = H\, \mathcal{P}^{\Gamma}_{\hat{U}^{n+1}}[u_{\text{true}}(t^{n+1})] = \mathcal{P}^{\Gamma}_{\hat{U}^{n+1}}[u_{\text{true}}(t^{n+1})].$$

Then, for the filtering distribution $\mathbb{P}(u^{n+1}|Z^n, z^{n+1})$*, obtained by either EnKF or particle filter analysis step, it holds*

$$\mathbb{P}(u^{n+1}|Z^n, z^{n+1}) = \mathbb{P}(u^{n+1}|Z^n, z_R^{n+1}).$$

Proof. We start with the analysis step performed via the particle filter. As described in Section 6.2.3, in the analysis step, the locations remain unchanged and the weights get updated via the formula (6.13). Since the observations are obtained through (6.2), the computation of $g^n(\hat{u}_{(j)}^{\text{DLR},n+1})$ from (6.15) involves

$$\begin{aligned}
g^n(\hat{u}_{(j)}^{\text{DLR},n+1}) &\propto \exp\Big(-\frac{1}{2}\|\hat{u}_{(j)}^{n+1} - z^{n+1}\|_\Gamma^2\Big) = \exp\Big(-\frac{1}{2}\|\hat{\bar{u}}^{\text{DLR},n+1} + \hat{U}^{n+1}\hat{Y}_{(j)}^{n+1} - z^{n+1}\|_\Gamma^2\Big)\\
&= \exp\Big(-\frac{1}{2}\Big\|\hat{\bar{u}}^{\text{DLR},n+1} + \hat{U}^{n+1}\hat{Y}_{(j)}^{n+1} - \mathcal{P}^{\Gamma}_{\hat{U}^{n+1}}[z^{n+1}] - \mathcal{P}^{\Gamma}_{\hat{U}^{n+1}}{}^{\perp}[z^{n+1}]\Big\|_\Gamma^2\Big)\\
&= \exp\Big(-\frac{1}{2}\Big(\Big\|\hat{\bar{u}}^{\text{DLR},n+1} + \hat{U}^{n+1}\hat{Y}_{(j)}^{n+1} - \mathcal{P}^{\Gamma}_{\hat{U}^{n+1}}[z^{n+1}]\Big\|_\Gamma^2\\
&\qquad\qquad - 2\Big\langle\hat{\bar{u}}^{\text{DLR},n+1}, \mathcal{P}^{\Gamma\perp}_{\hat{U}^{n+1}}[z^{n+1}]\Big\rangle_\Gamma + \Big\|\mathcal{P}^{\Gamma\perp}_{\hat{U}^{n+1}}[z^{n+1}]\Big\|_\Gamma^2\Big)\Big)\\
&= \exp\Big(-\frac{1}{2}\Big\|\hat{\bar{u}}^{\text{DLR},n+1} + \hat{U}^{n+1}\hat{Y}_{(j)}^{n+1} - \mathcal{P}^{\Gamma}_{\hat{U}^{n+1}}[z^{n+1}]\Big\|_\Gamma^2\Big)\\
&\qquad\qquad \exp\Big(\Big\langle\hat{\bar{u}}^{\text{DLR},n+1}, \mathcal{P}^{\Gamma\perp}_{\hat{U}^{n+1}}[z^{n+1}]\Big\rangle_\Gamma\Big)\ \exp\Big(-\frac{1}{2}\Big\|\mathcal{P}^{\Gamma\perp}_{\hat{U}^{n+1}}[z^{n+1}]\Big\|_\Gamma^2\Big)
\end{aligned}$$

As the last two terms are constant across particles, they can be included in the proportionality constant and we conclude

$$\begin{aligned}
g^n(\hat{u}_{(j)}^{\text{DLR},n+1}) &\propto \exp\Big(-\frac{1}{2}\Big\|\hat{\bar{u}}^{\text{DLR},n+1} + \hat{U}^{n+1}\hat{Y}_{(j)}^{n+1} - \mathcal{P}^{\Gamma}_{\hat{U}^{n+1}}[z^{n+1}]\Big\|_\Gamma^2\Big)\\
&= \exp\Big(-\frac{1}{2}\Big\|\hat{u}_{(j)}^{\text{DLR},n+1} - z_R^{n+1}\Big\|_\Gamma^2\Big).
\end{aligned}$$

As for the EnKF analysis step, we proceed in the following way. From the update formulas (7.10) we can see the data gets incorporated through the term

$$\begin{aligned}
\tilde{K}^{n+1} z_{(j)}^{n+1} &= \tilde{K}^{n+1}(z^{n+1} - H\hat{\bar{u}}^{\text{DLR},n+1} + \eta_{(j)}^{n+1})\\
&= \tilde{K}^{n+1}(\mathcal{P}^{\Gamma}_{\hat{U}^{n+1}}[z^{n+1}] + \mathcal{P}^{\Gamma\perp}_{\hat{U}^{n+1}}[z^{n+1}] - H\hat{\bar{u}}^{\text{DLR},n+1} + \eta_{(j)}^{n+1}).
\end{aligned}$$

Using the Woodbury Matrix Identity ([LSZ15, Lemma 4.4]), we further derive

$$\begin{aligned}\tilde{K}^{n+1}\mathcal{P}^{\Gamma\,\perp}_{\hat{U}^{n+1}}[z^{n+1}] &= C_{\hat{N},\hat{Y}^{n+1}}\tilde{H}^{\intercal}(\tilde{S}^{n+1})^{-1}\mathcal{P}^{\Gamma\,\perp}_{\hat{U}^{n+1}}[z^{n+1}]\\ &= C_{\hat{N},\hat{Y}^{n+1}}\tilde{H}^{\intercal}\Big(\mathrm{Id}-\Gamma^{-1}\tilde{H}(C^{-1}_{\hat{N},\hat{Y}^{n+1}}+\tilde{H}^{\intercal}\Gamma^{-1}\tilde{H})^{-1}\tilde{H}^{\intercal}\Big)\Gamma^{-1}\mathcal{P}^{\Gamma\,\perp}_{\hat{U}^{n+1}}[z^{n+1}]\\ &= 0,\end{aligned}$$

which proves the statement in the Lemma. ☐

In other words, if the forecasted particles belong to a certain subspace, in the analysis step, they only learn from the part of the observations laying in the subspace and omit completely its orthogonal complement. Depending on the scenario, a major portion of the information might get lost. In the following section, we will see that these conclusions play an important role when applying the proposed algorithms to the Lorenz-96 system.

7.1.3 Numerical examples: ensemble Kalman filter and particle filter

To assess the quality of these filtering algorithms, we consider the same test case as described in Section 6.3. We set the following parameters: final time $T = 100$, time between observations $\triangle t = 0.05$, the model error covariance matrix $\Sigma = 10^{-4} \cdot \mathrm{Id}$, the observation error covariance matrix $\Gamma = 10^{-2} \cdot \mathrm{Id}$ and the observation operator $H = \mathrm{Id}$. We set the number of particles $\hat{N} = 1000$ sufficiently high, to see a clear impact of the DLR approximation on the results.

Figure 7.2 depicts the RMS errors obtained by applying the DLR+EnKF with $R = 5, 10, 15$. Increasing the rank, we see a clear improvement in the performance of the filter. However, a rather high rank $R = 15$ is barely enough to keep track of the signal with a sufficient accuracy.

From Figure 7.3 we can observe again the strong dependence of the reliability of the PF on the number of particles. With $\hat{N} = 1000$, the DLR-PF approach is able to track correctly the state only when the rank is $R = 30$. For smaller rank, the filter procedure at some point looses track of the state. Increasing the number of particles to $\hat{N} = 10000$, we manage to get satisfactory results with $R = 20$.

7.1.4 Model error and analysis step: Gaussian mixture filter

Another technique to investigate consists in applying the DLR approximation within the forecast step of the GMF introduced in Subsection 6.2.5. This approach addresses the conjecture that projecting the model error on a DLR subspace and consequently not updating the subspace in the analysis step causes a significant loss of accuracy.

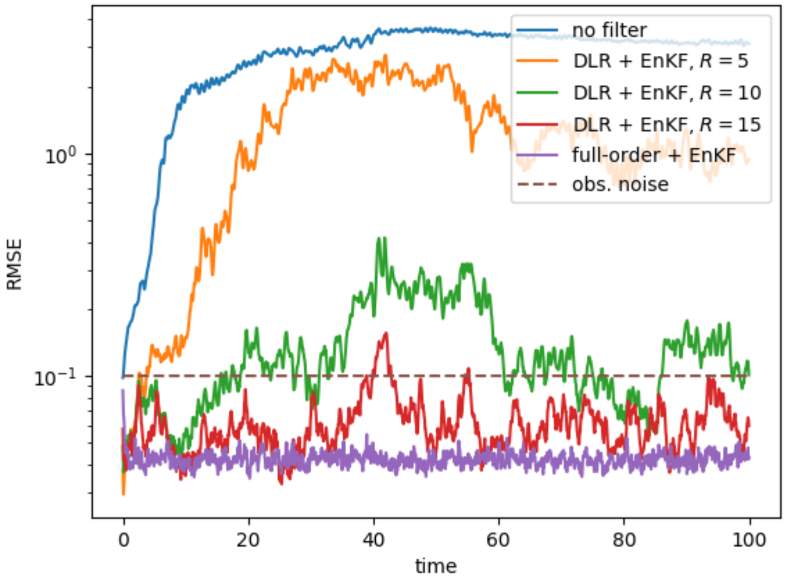


Figure 7.2 – Comparison of RMSEs when applying DLR with rank 5, 10, 15 vs. full-order with 1000 particles in the prediction step and EnKF in the analysis step.

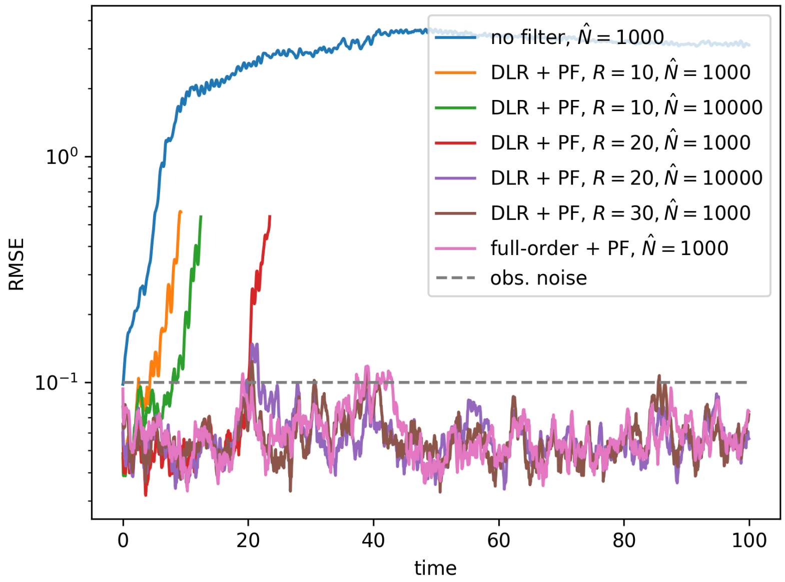


Figure 7.3 – Comparison of RMSEs when applying DLR with rank 10, 20, 30 vs. full order with 1000 or 10000 particles in the prediction step and particle filter in the analysis step.

The forecast step consists of two parts: evolving the system through the forward operator Ψ and incorporating the model error. The first part is executed in the same way as the DLR-EnKF or the DLR-PF of the previous section — the forward operator Ψ is approximated by the DLR method described in Section 7.1.1, resulting in a set of $\hat{N}$ particle locations and weights $\left\{\left(\Psi^{\mathrm{DLR}}(u^{\mathrm{DLR},n})\right)_{(j)}, \hat{\lambda}_{(j)}^{n+1}\right\}_{j=1}^{\hat{N}}$. The second part of the forecast step and the analysis step are performed analogously to the GMF algorithm described in Section 6.2.5. The particles provide the mixture means while the model error is used to build the mixture covariance of the forecasted distribution

$$\mathbb{P}(u^{n+1}|Z^n) = \sum_{j=1}^{\hat{N}} \hat{\lambda}_{(j)}^{n+1} \times N\Big(u^{n+1}; \left(\Psi^{\mathrm{DLR}}(u^{\mathrm{DLR},n})\right)_{(j)}, \Sigma\Big). \tag{7.11}$$

We recall that the mixture means are of the form

$$\left(\Psi^{\mathrm{DLR}}(u^{\mathrm{DLR},n})\right)_{(j)} = \hat{\bar{u}}^{\mathrm{DLR},n+1} + \sum_{r=1}^{R} \hat{U}_r^{n+1} \hat{Y}_{r,(j)}^{n+1}$$

where $\hat{\bar{u}}^{\mathrm{DLR},n+1}, \hat{U}_r^{n+1}, \hat{Y}_r^{n+1}$, $r = 1, \ldots, R$ are the mean, deterministic basis functions and stochastic coefficients, respectively, resulting from (7.7) at time t^{n+1}. Applying Lemma 6.2.2 we see that the analysis step results in a Gaussian mixture

$$\mathbb{P}(u^{n+1}|Z^{n+1}) \sim \sum_{j=1}^{\hat{N}} \lambda_{(j)}^{n+1} \times N(u^{n+1}; m_j^{n+1}, C^{n+1}) \tag{7.12}$$

with $\lambda_{(j)}^{n+1}, m_j^{n+1}, C^{n+1}$ computed with formulas (6.23).

Note that this algorithm does not project the model error on the DLR subspace. Realizations from the forecasted distribution (7.11), as well as from the filtering distribution (7.12), are generally not restricted to a low-dimensional subspace. To restart the forecast step for the new time step, we need obtain $\hat{N}$ new particles. This is performed analogously to (6.24). In addition, we need to recover a low-rank format (7.1). This is realized via an SVD decomposition which results in an updated DLR subspace that provides the best R-rank approximation (in the $\mathbb{E}_{\hat{N}}[\|\cdot\|]$-norm) to the samples of the filtered signal at time t^{n+1}

$$u_{(j)}^{\mathrm{DLR},n} = \bar{u}^{\mathrm{DLR},n+1} + \sum_{r=1}^{R} U_r^{n+1} Y_{r,(j)}^{n+1}. \tag{7.13}$$

Thanks to projecting the model error when applying the DLR-EnKF or DLR-PF, the filtering algorithms of Sections 7.1.1, 7.1.2 never see the squared full-state dimension $N_h \times N_h$ nor $N_h \times \hat{N}$ (assuming $l \ll N_h$). However, the GMF algorithm builds a covariance matrix of size $N_h \times N_h$ at the end of the forecast step and updates it in the analysis step. This could be avoided by an approximation of the model error covariance matrix by a matrix with a lower rank $R_C \ll N_h$, which consequently leads to an updated

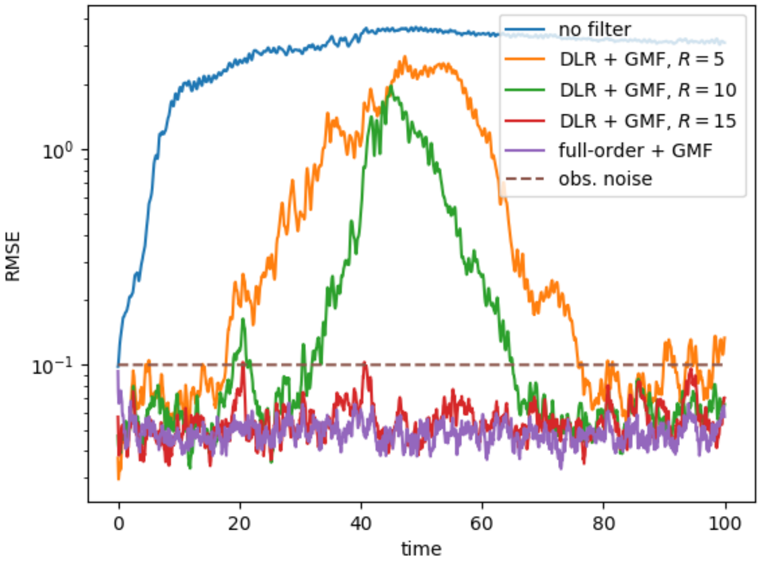


Figure 7.4 – Comparison of RMSEs when applying DLRA with rank 5, 10, 15 vs. full-order model with 1000 particles in the prediction step and GMF in the analysis step.

mixture covariance matrix of rank $\leq 2R_C$ (can be deduced from (6.20)).

7.1.5 Numerical examples: Gaussian mixture filter

Similarly to the previous numerical experiments, we test the performance of the proposed combination of DLRA with GMF by setting the following parameters: final time $T = 100$, time between observations $\triangle t = 0.05$, the model error covariance matrix $\Sigma = 10^{-4} \cdot \mathrm{Id}$, the observation error covariance matrix $\Gamma = 10^{-2} \cdot \mathrm{Id}$,the observation operator $H = \mathrm{Id}$ and the number of particles $\hat{N} = 1000$.

In Figure 7.4 we see that applying the GMF vastly improves the performance of the particle filter in its simple form with numerical results provided in Figure 7.3. Rank $R = 15$ is sufficient to track the signal. Comparing GMF and EnKF with results provided in Figure 7.2, we observe that both filters require $R = 15$ but GMF attains lower mean square error for such rank.

Having tested multiple techniques in combination with a simple DLRA in the prediction step, we can conclude that the algorithms are capable of tracking the signal in a sufficient way (at least w.r.t. the RMSE) but still a rather high rank ($R = 15$ for EnKF and GMF, $R = 20$ for PF) is required. In the following section, we propose two methods that manage to alleviate this demand by complementing the DLRA with a Gaussian term.

7.2 Complemented dynamical low rank approximation

The preceding section examined the idea of applying simple DLR approximation in the forecast step combined with various filtering techniques for the analysis step. We observed that completely omitting the portion of the signal not captured in the DLR subspace causes an accuracy loss, which for small ranks R becomes rather significant. The techniques presented in this section stem from the idea of approximating the omitted modes by a Gaussian approximation. In the probabilistic description of these methods we will focus on the evolution of the measures from time t^n to t^{n+1}. The signal at the beginning of each forecast step will be distributed as a Gaussian mixture

$$\mathbb{P}(u^n|Z^n) = \sum_{j=1}^{\hat{N}} \lambda^n_{(j)} \times N(u^n; m^n_j, C^n). \tag{7.14}$$

As the operator Ψ is deterministic, all the uncertainty in the system (6.3) arises from the initial condition u^n. The underlying abstract probability space $(\Omega, \mathcal{S}, \mu)$ can therefore be parametrized as

$$\Omega = \mathbb{R}^{N_h} \times \{m^n_1, \dots, m^n_{\hat{N}}\}, \quad \mathcal{S} = \mathcal{B}(\mathbb{R}^{N_h}) \times 2^{\{m^n_1, \dots, m^n_{\hat{N}}\}}$$
$$\mu = N(0, \mathrm{Id}_{N_h \times N_h}) \otimes \lambda,$$

where $\lambda = \{\lambda^n_1, \dots, \lambda^n_{\hat{N}}\}$ is a probability mass function over the (discrete) space $\{m^n_1, \dots, m^n_{\hat{N}}\}$ and the signal at the beginning of the forecast step satisfies

$$u^n(\omega) = m + (C^n)^{1/2}\gamma^{\intercal}, \qquad \text{with } \omega = (\gamma, m) \in \Omega. \tag{7.15}$$

In particular, m is a discrete random variable taking values $\{m^n_1, \dots, m^n_{\hat{N}}\}$ with probability λ and γ is a standard normal vector independent of m. Throughout the evolution within the forecast step, the signal can be split into two parts: u^{DLR} and the remaining portion which we will denote as u^{rest}. Evolving u^{rest} exactly is computationally as expensive as evolving the full state u_{true}. To make the problem tractable we will enforce that u^{rest} follows a zero mean Gaussian distribution at all times. As such, it can be represented as a linear combination of the Gaussian random vector γ

$$\begin{aligned} u_{\mathrm{true}}(t, \omega) &= \bar{u}(t) + u^{\mathrm{DLR}^*}(t, \omega) + u^{\mathrm{rest}}(t, \omega) \\ &\approx \bar{u}(t) + u^{\mathrm{DLR}^*}(t, m, \gamma) + A(t)\gamma^{\intercal}, \end{aligned} \tag{7.16}$$

where $A(t) \in \mathbb{R}^{N_h \times N_h}$ is unknown (as well as u^{DLR}) and has to be determined through the dynamics (6.3). The last term in (7.16) is then distributed as $u^{\mathrm{rest}} \sim N(0, T)$ with $T = AA^{\intercal}$. The difficult part is to derive evolution equations for u^{DLR^*} and T which would avoid redundancy, i.e. a situation in which a portion of the signal is captured in both terms. In Section 7.2.1 and 7.2.2, we propose two methods to propagate an approximate solution in the form (7.16) for the system (6.3). Then, in Section 7.2.3

we detail how to include the model error at the end of the forecast step and in Section 7.2.4, we detail the analysis step. Finally, Section 7.2.5 is dedicated to numerical results comparing the proposed methods.

7.2.1 Forecast step: DLRA with an independent Gaussian component

We are looking for a signal $u \in L^2(\Omega; \mathbb{R}^{N_h})$ approximating the solution of (6.3) for $t \in (t^n, t^{n+1})$ in the form

$$u(t,\omega) = \bar{u}(t) + u^{\mathrm{DLR}^*}(t,\omega) + u^{\mathrm{rest}}(t,\omega) = \bar{u}(t) + \sum_{r=1}^{R} U_r(t) Y(t,m,\gamma) + A(t)\gamma^{\intercal}.$$

As mentioned above, the challenge is to derive evolution equations for $\bar{u}, \{U_r\}_{r=1}^R, \{Y_r\}_{r=1}^R, A$, so that no portion of the signal is captured in both u^{DLR^*} and u^{rest}. This scenario is clearly avoided if u^{DLR^*} and u^{rest} are kept independent. In our first algorithm we therefore impose these terms to be independent at the initial time t^n and to remain independent over time. We denote this approximation by $u^{\mathrm{DLR+i}}$

$$u^{\mathrm{DLR+i}}(t,\omega) := \bar{u}(t) + \sum_{r=1}^{R} U_r(t) Y_r(t,m,\gamma) + A(t)\gamma^{\intercal},$$
$$\text{with} \quad Y(t,m,\gamma),\ A(t)\gamma^{\intercal} \quad \text{independent.} \tag{7.17}$$

Note that tracking the signal in this form with Y approximated by particles $\{Y_{(j)}, \lambda_{(j)}\}_{j=1}^{\hat{N}}$ means tracking the signal as a Gaussian mixture at every time t

$$u^{\mathrm{DLR+i}}(t) \sim \sum_{j=1}^{\hat{N}} \lambda_{(j)} \times N(\bar{u}(t) + U(t)Y_{(j)}(t), T(t))$$

with $T = AA^{\intercal}$.

Initial condition

The initial condition at time $t = t^n$ resulting from the analysis step of the previous time step is of the form (7.15). To proceed with our new-proposed method, we need to construct an approximation of it in the form (7.17). Notice that (7.15) is not in the form (7.17) because the mixture means $m_1^n, \dots, m_{\hat{N}}^n$ do not live, in general, in an R-dimensional subspace. Even if they were, we would like to split the Gaussian part $(C^n)^{1/2}\gamma^{\intercal}$ into a component in the R-dimensional subspace (so that to have a GMM in the DLR subspace) and an independent one in the orthogonal complement, which is not always possible. We need therefore to make some approximation steps to recast (7.15) into the form (7.17). The most natural idea is to assign $u^{\mathrm{DLR}^*}(t^n,\omega)$ to be the R-truncation of the Karhunen–Loève expansion of m from (7.15) and $A(t^n) = (C^n)^{1/2}$.

As a consequence, the u^{DLR^*} depends only on the projected discrete variable m, which is by construction independent of γ. This approach, however, does not lead to satisfactory results. With time, the DLR part u^{DLR^*}, which follows the nonlinear dynamics of the problem, might not capture the dominant part of the solution and the independence condition causes loss in accuracy. To mitigate this, we allow the DLR part $u^{\mathrm{DLR}^*}(t^n)$ to depend on $\omega = (\gamma, m)$, capturing the most dominant part of u^n from (7.15) (not only m).

From the Karhunen–Loève expansion of u^{n*}

$$u^{n*} = \sum_{k=1}^{N_h} \xi_k e_k,$$

with ξ_k ordered with decreasing variance, it holds that $\xi = \{\xi_1, \dots, \xi_{N_h}\}$ are uncorrelated random variables and $\{e_k\}_{k=1}^{N_h}$ are pairwise orthonormal w.r.t. the Euclidean inner product $\langle \cdot, \cdot \rangle$. In particular, $\{e_k\}_{k=1}^{N_h}$ are the eigenvectors of the covariance matrix $\tilde{C}^n = \sum_{j=1}^{\hat{N}} \lambda_{(j)}^n (m_j^n - \bar{m})(m_j^n - \bar{m})^\intercal + C^n$, with $\bar{m} = \sum_{j=1}^{\hat{N}} \lambda_{(j)}^n m_j^n$ and

$$\xi_k = \langle u^{n^*}, e_k \rangle = \langle m - \bar{m} + (C^n)^{1/2} \gamma^\intercal, e_k \rangle, \quad k = 1, \dots, N_h.$$

We denote the corresponding eigenvalues by $\{\vartheta_k\}_{k=1}^{N_h}$ so that $\mathbb{E}[\xi_k \xi_l] = \delta_{kl} \vartheta_k$. We then set

$$\bar{u}^n = \mathbb{E}[u^n] = \bar{m}, \quad U_r^n = e_r, \; Y_r^n(\omega) = \xi_r, \quad \forall 1 \leq r \leq R. \tag{7.18}$$

Notice that, by proceeding this way, the DLR part of the signal $u^{\mathrm{DLR},n} = \bar{u}^n + \sum_{r=1}^{R} U_r^n Y_r^n$ has a GMM distribution concentrated on the R-dimensional subspace spanning $U = (U_1^n, \dots, U_R^n)$

$$u^{\mathrm{DLR},n} \sim \sum_{j=1}^{\hat{N}} \lambda_j^n \times N(\bar{m} + UU^\intercal m_j^{n^*}, \, UU^\intercal C^n UU^\intercal),$$

where $m_j^{n^*} = m_j^n - \bar{m}$. As such, it can be re-parametrized by a discrete random variable $\tilde{m}$ taking values $\{\bar{m} + UU^\intercal m_1^{n^*}, \dots, \bar{m} + UU^\intercal m_{\hat{N}}^{n^*}\}$ with probability $\lambda = \{\lambda_1^n, \dots, \lambda_{\hat{N}}^n\}$, and an R-dimensional standard normal vector $\tilde{\gamma}_{1:R} = (\tilde{\gamma}_1, \dots, \tilde{\gamma}_R) \sim N(0, \mathrm{Id}_{R \times R})$ as

$$u^{\mathrm{DLR},n} = \tilde{m} + U(U^\intercal C^n U)^{1/2} \tilde{\gamma}_{1:R}^\intercal.$$

Concerning the u^{rest} part,

$$u^{\mathrm{rest}} = u^n - u^{\mathrm{DLR},n} = \sum_{k=R+1}^{N_h} \xi_k e_k$$

we proceed by simply replacing ξ_k by $\sqrt{\vartheta_k} \tilde{\gamma}_k$ with $\tilde{\gamma}_{R+1:N_h} = (\tilde{\gamma}_{R+1}, \dots, \tilde{\gamma}_{N_h}) \sim N(0, \mathrm{Id}_{N_h - R \times N_h - R})$ and independent of $\tilde{\gamma}_{1:R}$. This choice preserves the mean and covariance of u^{rest}. In conclusion, our final approximation of the initial condition reads

$$u^{\text{DLR+i},n} = \tilde{m} + U(U^\intercal C^n U)^{1/2}\tilde{\gamma}_{1:R}^\intercal + U^\perp\Big((U^\perp)^\intercal C^n U^\perp\Big)^{1/2}\tilde{\gamma}_{R+1:N_h}^\intercal, \tag{7.19}$$

where $U^\perp = (e_{R+1}, \dots, e_{N_h})$. We set

$$A^n = U^\perp\Big((U^\perp)^\intercal C^n U^\perp\Big)^{1/2}.$$

We work then in the probability space Ω parametrized by the discrete random variable $\tilde{m}$ and the Gaussian vector $\tilde{\gamma} = (\tilde{\gamma}_{1:R}, \tilde{\gamma}_{R+1:N_h})$. In what follows, we rename $\tilde{m}$ as m and $\tilde{\gamma}$ as γ. In Theorem 7.2.1 we will see, that when propagating the signal $u^{\text{DLR+i}}$, there is no need to track explicitly the dependence on $\tilde{\gamma}_{R+1:N_h}$ and we will work only with $T = AA^\intercal$.

Note that the initial condition (6.1) at time $t = 0$ is distributed normally $u^0 \sim N(m^0, C^0)$ and thus is a special case of (7.15) with $\lambda_{(1)}^n = 1$, $\lambda_{(j)}^n = 0$, $j = 2, \dots, \hat{N}$. In this case, the initial condition for the u^{rest} is simplified. Indeed, the vectors $\{e_k\}_{k=1}^{N_h}$ are eigenvectors of C^0 and the matrix $(U^\perp)^\intercal C^0 U$ is diagonal. Such initial condition is in fact exact. It is clear that in this case, the u^{rest} term is independent of the u^{DLR^*} term.

We follow by defining the complemented DLR solution for $t \in (t^n, t^{n+1})$.

Definition 7.2.1. We define the DLR solution complemented by an independent Gaussian term of the problem (6.3) as

$$u^{\text{DLR+i}}(t,\omega) = \bar{u}(t) + \sum_{r=1}^{R} U_r(t) Y_r(t, m, \gamma_{1:R}) + A(t)\gamma_{R+1:N_h}^\intercal,$$

where $\bar{u}$ is the solution of

$$\dot{\bar{u}} = \mathbb{E}[\mathcal{F}(u^{\text{DLR+i}})], \qquad t \in (t^n, t^{n+1}) \tag{7.20}$$

and $\{U_r\}_{r=1}^R, \{Y_r\}_{r=1}^R, A$ are solutions of the following variational formulation for $t \in (t^n, t^{n+1})$

$$\begin{aligned}\mathbb{E}\Big[\Big\langle \dot{U}Y^\intercal + U\dot{Y}^\intercal + \dot{A}\gamma_{R+1:N_h}{}^\intercal,\, v\Big\rangle\Big] &= \mathbb{E}\Big[\Big\langle \mathcal{F}^*\Big(u^{\text{DLR+i}}\Big),\, v\Big\rangle\Big]\\ \forall v \in \{f = \delta U Y^\intercal + U\delta Y^\intercal + \delta A \gamma_{R+1:N_h}^\intercal &\text{ with } \delta U \in \mathbb{R}^{N_h\times R},\ \langle \delta U, U\rangle = 0,\\ \delta A \in \mathbb{R}^{N_h\times(N_h-R)},\ \delta Y \in L_0^2(\Omega;\mathbb{R}^R),\ &\delta Y \text{ independent of } \gamma_{R+1:N_h}\}.\end{aligned} \tag{7.21}$$

We see that the independence of the two terms is enforced in the variational formulation by constraining the set of test functions. We only allow the variations in Y which are independent of the random variables $\gamma_{R+1:N_h}$. The detailed evolution equations for the DLR modes and the matrix A will be specified in the following theorem for the case of the operator $\mathcal{F}$ of the form (6.26).

Theorem 7.2.1. *Consider the problem* (6.26)*, i.e. the operator $\mathcal{F}$ is of the form*

$$\mathcal{F}(u) = (L + D)u + B(u, u) + F.$$

The variational formulation (7.21) *results in the following system of equations for the mean value $\dot{\bar{u}}$ and the matrix $T = AA^\intercal$:*

$$\begin{aligned}
\dot{\bar{u}} &= (L + D)\bar{u} + F + B(\bar{u}, \bar{u}) + \sum_{j,k=1}^{N_h} S_{jk} B(v_j, v_k) \\
\dot{T} &= L_u T + T L_u{}^\intercal \\
&\quad \text{with } (L_u)_{ij} = (L_{ij} + D_{ij}) + \langle B(\bar{u}, v_j), v_i\rangle + \langle B(v_j, \bar{u}), v_i\rangle,
\end{aligned} \tag{7.22}$$

where $\{v_k\}_{k=1}^{N_h}$ is the canonical basis spanning $\mathbb{R}^{N_h}$ and $S \in \mathbb{R}^{N_h \times N_h}$ is a matrix approximating the covariance matrix of the full signal $u^{\mathrm{DLR+i}}$ obtained as

$$S = \mathbb{E}[(u^{\mathrm{DLR+i}} - \bar{u})(u^{\mathrm{DLR+i}} - \bar{u})^\intercal] = U\mathbb{E}[Y^\intercal Y]U^\intercal + T.$$

The DLR modes $\{U_r, Y_r\}_{r=1}^{R}$ satisfy

$$\begin{aligned}
\dot{Y}_r &= \sum_{k=1}^{R} \langle (L + D)U_k, U_r\rangle Y_k + \sum_{k=1}^{R} \langle B(\bar{u}, U_k), U_r\rangle Y_k + \langle B(U_k, \bar{u}), U_r\rangle Y_k \\
&\quad + \sum_{j,k=1}^{R} \langle B(U_j, U_k), U_r\rangle \Big(Y_j Y_k - \mathbb{E}[Y_j Y_k]\Big) \\
&= \langle \mathcal{F}^*(u^{\mathrm{DLR}}), U_r\rangle, \qquad r = 1, \dots, R,
\end{aligned} \tag{7.23}$$

$$\sum_{j=1}^{R} \dot{U}_j (C_Y)_{jr} = \Big(\mathbb{E}[\mathcal{F}^*(u^{\mathrm{DLR}}) Y_r] - \mathcal{P}_U\Big[\mathbb{E}[\mathcal{F}^*(u^{\mathrm{DLR}}) Y_r]\Big]\Big), \qquad r = 1, \dots, R$$

$$\text{where } C_Y = \mathbb{E}[Y^\intercal Y].$$

Proof. For the sake of simplicity, we denote by $u^{\mathrm{DLR}} = \bar{u}(t) + \sum_{r=1}^{R} U_r(t) Y_r(t, \gamma_{1:R})$ the DLR portion of the signal, by $u^{\mathrm{DLR}^*} = \sum_{r=1}^{R} U_r(t) Y_r(t, \gamma_{1:R})$ the stochastic part of the DLR portion of the signal and by $u^{\mathrm{rest}} = A\gamma_{R+1:N_h}^\intercal$ the complementary independent term.

The mean value can be obtained as a solution of the following equation

$$\begin{aligned}
\dot{\bar{u}} &= \mathbb{E}[(L + D)(\bar{u} + u^{\mathrm{DLR}^*} + u^{\mathrm{rest}})] + F \\
&\quad + \mathbb{E}[B(\bar{u}, \bar{u})] + \mathbb{E}[B(\bar{u}, u^{\mathrm{DLR}^*})] + \mathbb{E}[B(\bar{u}, u^{\mathrm{rest}})] \\
&\quad + \mathbb{E}[B(u^{\mathrm{DLR}^*}, \bar{u})] + \mathbb{E}[B(u^{\mathrm{DLR}^*}, u^{\mathrm{DLR}^*})] + \mathbb{E}[B(u^{\mathrm{DLR}^*}, u^{\mathrm{rest}})]
\end{aligned}$$

$$
\begin{aligned}
&\quad + \mathbb{E}[B(u^{\text{rest}},\bar{u})] + \mathbb{E}[B(u^{\text{rest}},u^{\text{DLR}^*})] + \mathbb{E}[B(u^{\text{rest}},u^{\text{rest}})]\\
&= (L+D)\bar{u} + F + B(\bar{u},\bar{u}) + \mathbb{E}[B(u^{\text{DLR}^*},u^{\text{DLR}^*})]\\
&\quad + \mathbb{E}[B(u^{\text{rest}},u^{\text{rest}})]\\
&= (L+D)\bar{u} + F + B(\bar{u},\bar{u}) + \sum_{j,k=1}^{R} \mathbb{E}[Y_j Y_k] B(U_j,U_k)\\
&\quad + \sum_{j,k=R+1}^{N_h} \mathbb{E}[\gamma_j\gamma_k] B(A_j,A_k)\\
&= (L+D)\bar{u} + F + B(\bar{u},\bar{u}) + \sum_{j,k=1}^{N_h} S_{jk} B(v_j,v_k),
\end{aligned}
$$

where $\{v_k\}_{k=1}^{N_h}$ is the canonical basis spanning $\mathbb{R}_h^N$ and $S \in \mathbb{R}^{N_h \times N_h}$ is a matrix approximating the covariance matrix of the full signal $u^{\text{DLR+i}}$ obtained as

$$
S = \mathbb{E}[(u^{\text{DLR+i}} - \bar{u})(u^{\text{DLR+i}} - \bar{u})^\intercal] = U\mathbb{E}[Y^\intercal Y]U^\intercal + AA^\intercal.
$$

We now proceed with deriving evolution equations for the stochastic DLR modes $\{Y_r\}_{r=1}^R$. Consider a test function of the form $v = U_r \delta Y_r$. Stemming from the variational formulation (7.21), we derive

$$
\begin{align}
\mathbb{E}[\dot{Y}_r \delta Y_r] &= \sum_{k=1}^{R} \langle \dot{U}_k, U_r\rangle \mathbb{E}[Y_k \delta Y_r] + \langle U_k, U_r\rangle \mathbb{E}[\dot{Y}_k \delta Y_r] + \sum_{k=R+1}^{N_h} \langle \dot{A}_k, U_r\rangle \mathbb{E}[\gamma_k, \delta Y_r] \notag\\
&= \mathbb{E}\Big[\Big\langle \mathcal{F}^*\Big(u^{\text{DLR}} + u^{\text{rest}}\Big), U_r\delta Y_r\Big\rangle\Big] \notag\\
&= \mathbb{E}\Big[\Big\langle (L+D)u^{\text{DLR}}, U_r\Big\rangle \delta Y_r\Big] + \mathbb{E}\Big[\Big\langle B(u^{\text{DLR}},u^{\text{DLR}}), U_r\Big\rangle \delta Y_r\Big] \notag\\
&\quad + \mathbb{E}\Big[\Big\langle (L+D)u^{\text{rest}}, U_r\Big\rangle \delta Y_r\Big] + \mathbb{E}\Big[\Big\langle B(u^{\text{DLR}},u^{\text{rest}}), U_r\Big\rangle \delta Y_r\Big] \tag{7.24}\\
&\quad + \mathbb{E}\Big[\Big\langle B(u^{\text{rest}},u^{\text{DLR}}), U_r\Big\rangle \delta Y_r\Big] + \mathbb{E}\Big[\Big\langle B(u^{\text{rest}},u^{\text{rest}}), U_r\Big\rangle \delta Y_r\Big] \tag{7.25}\\
&= \sum_{k=1}^{R} \langle (L+D)U_k, U_r\rangle \mathbb{E}[Y_k \delta Y_r] + \sum_{k=1}^{R} \langle B(\bar{u},U_k), U_r\rangle \mathbb{E}[Y_k \delta Y_r] \notag\\
&\quad + \langle B(U_k,\bar{u}), U_r\rangle \mathbb{E}[Y_k \delta Y_r] \tag{7.26}\\
&\quad + \sum_{j,k=1}^{R} \langle B(U_j,U_k), U_r\rangle \Big(\mathbb{E}\Big[(Y_j Y_k - \mathbb{E}[Y_j Y_k])\delta Y_r\Big]\Big). \notag
\end{align}
$$

All four terms in (7.24), (7.25) vanished since the considered δY_r is a random function independent of u^{rest}. The resulting differential equation for the stochastic modes is

$$\begin{aligned}
\dot{Y}_r &= \sum_{k=1}^{R} \langle (L+D)U_k, U_r\rangle Y_k + \sum_{k=1}^{R} \langle B(\bar{u}, U_k), U_r\rangle Y_k + \langle B(U_k, \bar{u}), U_r\rangle Y_k \\
&\quad + \sum_{j,k=1}^{R} \langle B(U_j, U_k), U_r\rangle \Big(Y_j Y_k - \mathbb{E}[Y_j Y_k]\Big) \\
&= \langle \mathcal{F}^*(u^{\text{DLR}}), U_r\rangle, \qquad r = 1, \dots, R,
\end{aligned}$$

which are the standard DLR equations for the stochastic modes when evaluating the operator only in the DLR portion of the signal u^{DLR}.

To derive the evolution equations for the deterministic modes, we consider the test function $v = \mathcal{P}_U^{\perp}[\delta U_r] Y_r$ with $\delta U_r \in \mathbb{R}^{N_h}$ arbitrary. We follow by

$$\begin{align*}
&\sum_{k=1}^{R} \mathbb{E}[\dot{Y}_k Y_r] \langle U_k, \mathcal{P}_U^{\perp}[\delta U_r]\rangle + \sum_{k=1}^{R} \mathbb{E}[Y_k Y_r] \langle \dot{U}_k, \mathcal{P}_U^{\perp}[\delta U_r]\rangle \\
&\quad = \sum_{k=1}^{R} \mathbb{E}[\dot{Y}_k Y_r] \langle U_k, \mathcal{P}_U^{\perp}[\delta U_r]\rangle + \sum_{k=1}^{R} \mathbb{E}[Y_k Y_r] \langle \dot{U}_k, \mathcal{P}_U^{\perp}[\delta U_r]\rangle + \langle \mathbb{E}[u^{\text{rest}} Y_r] \mathcal{P}_U^{\perp}[\delta U_r]\rangle \\
&\quad = \mathbb{E}\Big[\Big\langle \mathcal{F}^*\Big(u^{\text{DLR}} + u^{\text{rest}}\Big),\ \mathcal{P}_U^{\perp}[\delta U_r] Y_r\Big\rangle\Big] \\
&\quad = \mathbb{E}\Big[\Big\langle (L+D)u^{\text{DLR}}, \mathcal{P}_U^{\perp}[\delta U_r]\Big\rangle Y_r\Big] + \mathbb{E}\Big[\Big\langle B(u^{\text{DLR}}, u^{\text{DLR}}), \mathcal{P}_U^{\perp}[\delta U_r]\Big\rangle Y_r\Big] \\
&\qquad + \mathbb{E}\Big[\Big\langle (L+D)u^{\text{rest}}, \mathcal{P}_U^{\perp}[\delta U_r]\Big\rangle Y_r\Big] + \mathbb{E}\Big[\Big\langle B(u^{\text{DLR}}, u^{\text{rest}}), \mathcal{P}_U^{\perp}[\delta U_r]\Big\rangle Y_r\Big] \tag{7.27} \\
&\qquad + \mathbb{E}\Big[\Big\langle B(u^{\text{rest}}, u^{\text{DLR}}), \mathcal{P}_U^{\perp}[\delta U_r]\Big\rangle Y_r\Big] + \mathbb{E}\Big[\Big\langle B(u^{\text{rest}}, u^{\text{rest}}), \mathcal{P}_U^{\perp}[\delta U_r]\Big\rangle Y_r\Big] \tag{7.28} \\
&\quad = \sum_{k=1}^{R} \mathbb{E}[Y_k Y_r] \langle (L+D)U_k, \mathcal{P}_U^{\perp}[\delta U_r]\rangle + \sum_{k=1}^{R} \langle B(\bar{u}, U_k), \mathcal{P}_U^{\perp}[\delta U_r]\rangle \mathbb{E}[Y_k Y_r] \\
&\qquad + \langle B(U_k, \bar{u}), \mathcal{P}_U^{\perp}[\delta U_r]\rangle \mathbb{E}[Y_k Y_r] + \sum_{j,k=1}^{R} \langle B(U_j, U_k), \mathcal{P}_U^{\perp}[\delta U_r]\rangle \mathbb{E}[Y_j Y_k Y_r].
\end{align*}$$

Analogously, the four terms in (7.27) and (7.28) vanished since Y_r are independent of

u^{rest}. The evolution equation for the deterministic modes becomes

$$\begin{aligned}\sum_{k=1}^{R}\mathbb{E}[Y_kY_r]\langle\dot{U}_k,\mathcal{P}_U^{\perp}[\delta U_r]\rangle &= \Big\langle\sum_{k=1}^{R}\mathbb{E}[Y_kY_r](L+D)U_k+\sum_{k=1}^{R}B(\bar{u},U_k)\mathbb{E}[Y_kY_r]\\ &\quad + B(U_k,\bar{u})\mathbb{E}[Y_kY_r]+\sum_{j,k=1}^{R}B(U_j,U_k)\mathbb{E}[Y_jY_kY_r],\,\mathcal{P}_U^{\perp}[\delta U_r]\Big\rangle \\ &= \Big\langle\mathbb{E}\Big[\mathcal{F}^*(u^{\mathrm{DLR}})Y_r\Big],\,\mathcal{P}_U^{\perp}[\delta U_r]\Big\rangle,\quad \forall r=1,\dots,R.\end{aligned}\tag{7.29}$$

Using the symmetry of the projection operator $\mathcal{P}_U^{\perp}$ w.r.t. the inner product $\langle\cdot,\cdot\rangle$, we obtain

$$\sum_{k=1}^{R}\mathbb{E}[Y_kY_r]\langle\mathcal{P}_U^{\perp}[\dot{U}_k],\delta U_r\rangle=\sum_{k=1}^{R}\mathbb{E}[Y_kY_r]\langle\dot{U}_k,\delta U_r\rangle=\Big\langle\mathcal{P}_U^{\perp}\Big[\mathbb{E}[\mathcal{F}^*(u^{\mathrm{DLR}})Y_r]\Big],\,\delta U_r\Big\rangle,$$
$$\forall r=1,\dots,R.$$

Set in the physical space, this is equivalent to

$$\sum_{k=1}^{R}\mathbb{E}[Y_kY_r]\dot{U}_r=\mathbb{E}\Big[\Big(\mathcal{F}^*(u^{\mathrm{DLR}})-\sum_{k=1}^{R}\langle\mathcal{F}^*(u^{\mathrm{DLR}}),U_k\rangle U_k\Big)Y_r\Big],\qquad\forall r=1,\dots,R,$$

which is the standard DLR system of deterministic equations to obtain the basis $\{U_r\}_{r=1}^R$, evaluated only in the DLR portion of the signal u^{DLR}.

As the last step, we develop equations for the matrix A tracking the linear dependence on $\gamma_{R+1:N_h}$. Let us consider the test function $v=\delta A\gamma_l$ for some $l\in\{R+1,\dots,N_h\}$, $\delta A\in\mathbb{R}^{N_h}$ in the variational formulation (7.21). We derive

$$\begin{aligned}\langle\dot{A}_l,\delta A\rangle &= \mathbb{E}\Big[\langle\sum_{k=R+1}^{N_h}\dot{A}_k\gamma_k,\delta A\gamma_l\rangle\Big]+\mathbb{E}[\langle\dot{u}^{\mathrm{DLR}},\delta A\rangle\gamma_l]=\mathbb{E}\Big[\langle\mathcal{F}^*(u^{\mathrm{DLR+i}}),\delta A\rangle\gamma_l\Big]\\ &= \mathbb{E}\Big[\langle(L+D)u^{\mathrm{rest}},\delta A\rangle\gamma_l\Big]+\mathbb{E}\Big[\langle B(\bar{u},u^{\mathrm{rest}}),\delta A\rangle\gamma_l\Big]\\ &\quad+\mathbb{E}\Big[\langle B(u^{\mathrm{rest}},u^{\mathrm{rest}}),\delta A\rangle\gamma_l\Big]+\mathbb{E}\Big[\langle B(u^{\mathrm{rest}},\bar{u}),\delta A\rangle\gamma_l\Big]\\ &= \Big\langle(L+D)A_l,\delta A\Big\rangle+\Big\langle\sum_{k=R+1}^{N_h}B(\bar{u},A_k),\delta A\Big\rangle\mathbb{E}[\gamma_k\gamma_l]\\ &\quad+\Big\langle\sum_{k=R+1}^{N_h}B(A_k,\bar{u}),\delta A\Big\rangle\mathbb{E}[\gamma_k\gamma_l]+\Big\langle\sum_{j,k=R+1}^{N_h}B(A_k,A_j),\delta A\Big\rangle\underbrace{\mathbb{E}[\gamma_j\gamma_k\gamma_l]}_{=0}\\ &= \Big\langle(L+D)A_l,\delta A\Big\rangle+\Big\langle B(\bar{u},A_l),\delta A\Big\rangle+\Big\langle B(A_l,\bar{u}),\delta A\Big\rangle.\end{aligned}$$

In the third equality, we used the fact the u^{DLR} and u^{rest} are independent. In the fourth inequality, we applied the fact that the third moments of Gaussian random variables are equal to zero. The resulting equation for each of the columns of the sought matrix A set in the physical space becomes

$$\dot{A}_l = (L + D)A_l + B(\bar{u}, A_l) + B(A_l, \bar{u}). \tag{7.30}$$

The evolution equation for the covariance matrix of the independent term u^{rest} can be obtained as

$$\dot{T} = \dot{A}A^\intercal + A\dot{A}^\intercal = L_u T + T L_u{}^\intercal \qquad \text{with } (L_u)_{ij} = (L_{ij} + D_{ij}) + \langle B(\bar{u}, v_j), v_i\rangle + \langle B(v_j, \bar{u}), v_i\rangle.$$

□

We see that the DLR modes $\{U_r, Y_r\}_{r=1}^R$ follow exactly the standard DLR evolution equations (7.5), where the operator $\mathcal{F}$ is evaluated only in the DLR portion of the signal u^{DLR} (not in the full signal $u^{\text{DLR+i}}$). This results in the stochastic modes dependent only on the random variables $\gamma_{1:R}$. For their evolution we apply a particle approximation $\{Y_{(j)}, \lambda_{(j)}\}_{j=1}^{\hat{N}}$. The evolution of the complement is traced only through its covariance matrix T. The two terms u^{DLR^*} and u^{rest} communicate only through the mean value $\bar{u}$ whose evolution involves the full covariance matrix S. There is therefore no need to track the matrix A.

For high-dimensional problems, keeping track of the matrix $T \in \mathbb{R}^{N_h \times N_h}$ is not feasible and one needs to rely on some further low-rank or a particle approximation. This approach is not analysed further in this work and remains a possible future research direction. For completeness, we recall the dimension of all terms which determine the distribution of $u^{\text{DLR+i}}$:

$$\bar{u} \in \mathbb{R}^{N_h}, \quad U \in \mathbb{R}^{N_h \times R},\ Y \in \hat{N} \times R, \quad T \in \mathbb{R}^{N_h \times N_h}.$$

7.2.2 Forecast step: DLRA with a linear complement

Similarly as before, in our second method we are looking for a signal $u \in L^2(\Omega; \mathbb{R}^{N_h})$ approximating the solution of (6.3) for $t \in (t^n, t^{n+1})$ in the form

$$u(t, \omega) = \bar{u}(t) + u^{\text{DLR}^*}(t, m, \gamma) + A(t)\gamma^\intercal.$$

However, in this case we do not require independence between u^{DLR} and $u^{\text{rest}} = A\gamma^\intercal$. First, let us introduce some new notation.

We define the space of random functions linear w.r.t. $\gamma = (\gamma_1, \dots, \gamma_{N_h}) \in \mathbb{R}^{N_h}$

$$\mathcal{W}_\gamma^{\text{lin}} = \{f \in L^2_0(\Omega;\mathbb{R}) |\, f = B\gamma^\intercal,\, B \in \mathbb{R}^{1\times N_h}\}.$$

In addition, we define the projection on $\mathcal{W}_\gamma^{\text{lin}}$

$$\begin{aligned}\mathcal{P}_{\mathcal{W}_\gamma^{\text{lin}}} &: L^2_0(\Omega;\mathbb{R}) \to \mathcal{W}_\gamma^{\text{lin}}\\ \mathcal{P}_{\mathcal{W}_\gamma^{\text{lin}}}[f] &= \sum_{k=1}^{N_h} \mathbb{E}[f\gamma_k]\,\gamma_k.\end{aligned}$$

Note that the space $L^2_0(\Omega;\mathbb{R})$ can be split into

$$L^2_0(\Omega;\mathbb{R}) = \mathcal{W}_\gamma^{\text{lin}} \oplus \mathcal{W}_\gamma^{\text{lin}\perp}. \tag{7.31}$$

Based on this decomposition, we split the signal $u^*_{\text{true}} \in \left(L^2_0(\Omega;\mathbb{R})\right)^{N_h}$ in two parts, the linear component evolving only in $(\mathcal{W}_\gamma^{\text{lin}})^{N_h}$ and the DLR component evolving in its orthogonal complement $u^{\text{DLR}^*} \in \left((\mathcal{W}_\gamma^{\text{lin}})^{N_h}\right)^\perp$ and we denote this approximate solution by $u^{\text{DLR}+\ell}$

$$u^{\text{DLR}+\ell}(t,\omega) := \bar{u}(t) + \sum_{r=1}^{R} U_r(t)Y_r(t,\omega) + \sum_{k=1}^{N_h} A_k(t)\gamma_k, \qquad \text{with } \{Y_r\}_{r=1}^R \subset \left((\mathcal{W}_\gamma^{\text{lin}})^R\right)^\perp. \tag{7.32}$$

The linear part $A\gamma^\intercal = \sum_{k=1}^{N_h} A_k(t)\gamma_k \in (\mathcal{W}_\gamma^{\text{lin}})^{N_h}$ captures the linear dependence on γ of the full signal $u^{\text{DLR}+\ell}$, not only of the remaining part u^{rest}. On the other hand, the DLR term remains well separated from $A\gamma^\intercal$ by enforcing $u^{\text{DLR}^*} \in \left((\mathcal{W}_\gamma^{\text{lin}})^{N_h}\right)^\perp$.

Initial condition

Again, as a first step we describe how to obtain the initial condition in the form (7.32). The initial condition at time t^n is distributed as a GM (7.14), which can be parametrized as

$$u^n = m + (C^n)^{1/2}\gamma^\intercal$$

with m a discrete random variable taking values in $\{m_1^n, \dots, m_{N_h}^n\}$ with probability $\lambda = \{\lambda_1^n, \dots, \lambda_{N_h}^n\}$ and $\gamma \sim N(0, \text{Id}_{N_h\times N_h})$ independent of m. In general, u^n is not in the form (7.32), so an approximation step is necessary. Differently from Section 7.2.1, this time we perform a Karhunen–Loève expansion only of the variable m (instead of u^n):

$$m = \mathbb{E}[m] + \sum_{k=1}^{N_h} \xi_k e_k$$

with $\{e_k\}_{k=1}^{N_h}$ eigenvectors of the covariance matrix $\sum_{j=1}^{\hat{N}} \lambda_j^n m_j^n m_j^{n\intercal}$ and $\xi_k = \langle m - \mathbb{E}[m], e_k \rangle$.

We then set

$$\begin{gathered}\bar{u}^n = \mathbb{E}[u^n], \quad U_r^n = e_r, \ Y_r^n = \xi_r, \ r = 1, \dots, R \\ A^n = (C^n)^{1/2},\end{gathered}$$

i.e. the means $\{m_j^n\}_{j=1}^{\hat{N}}$ are used to obtain the DLR portion of the signal and the GM covariance C^n is used to compute the new matrix A^n, so that $A^n A^{n\intercal} = C^n$.

For the time $t = 0$, the initial condition is distributed as $u^0 \sim N(m^0, C^0)$. It is therefore natural to set the following initial conditions for the terms in (7.32)

$$\begin{aligned}&\bar{u}^0 = m^0, \ A^0 = (C^0)^{1/2}, \\ &U^0 \text{ - arbitrary orthonormal set of } R \text{ vectors in } \mathbb{R}^{N_h}, \quad Y^0 = 0.\end{aligned} \tag{7.33}$$

We follow by defining the complemented DLR solution for $t \in (t^n, t^{n+1})$.

Definition 7.2.2. We define the DLR solution complemented by a linear term of the problem (6.3) as

$$u^{\mathrm{DLR}+\ell}(t, \omega) = \bar{u}(t) + \sum_{r=1}^{R} U_r(t) Y_r(t, m, \gamma) + A(t)\gamma^\intercal$$

where $\bar{u}$ is the solution of

$$\dot{\bar{u}} = \mathbb{E}[\mathcal{F}(u^{\mathrm{DLR}+\ell})], \qquad t \in (t^n, t^{n+1}) \tag{7.34}$$

and $\{U_r\}_{r=1}^R, \{Y_r\}_{r=1}^R, A$ are solutions of the following variational formulation for $t \in (t^n, t^{n+1})$

$$\begin{aligned}&\mathbb{E}\Big[\Big\langle \dot{U}Y^\intercal + U\dot{Y}^\intercal + \dot{A}\gamma^\intercal, \, v\Big\rangle\Big] = \mathbb{E}\Big[\Big\langle \mathcal{F}^*\Big(u^{\mathrm{DLR}+\ell}\Big), \, v\Big\rangle\Big] \\ &\qquad \forall v \in \{w = \delta U Y^\intercal + U \delta Y^\intercal + \delta A \gamma^\intercal \text{ with } \delta U \in \mathbb{R}^{N_h \times R}, \ \langle \delta U^\intercal, U \rangle = 0 \\ &\qquad\qquad \delta A \in \mathbb{R}^{N_h \times N_h}, \ \delta Y \in \Big((\mathcal{W}_\gamma^{\mathrm{lin}})^R\Big)^\perp\}.\end{aligned} \tag{7.35}$$

The detailed evolution equations for the DLR modes and the matrix A will be specified in the following theorem.

Theorem 7.2.2. *The variational formulation* (7.35) *results in the following system of* *equations for the mean value* $\dot{\bar{u}}$*, the DLR modes* $\{U_r, Y_r\}_{r=1}^R$ *and the matrix* A*:*

$$
\begin{aligned}
\dot{\bar{u}} &= \mathbb{E}[\mathcal{F}(u^{\mathrm{DLR}+\ell})] \\
\dot{A}_k &= \mathbb{E}[\mathcal{F}^*(u^{\mathrm{DLR}+\ell})\gamma_k], \quad k = 1, \dots, N_h \\
\dot{Y}_r &= \Big\langle U_r, \mathcal{P}^{\perp}_{\mathcal{W}^{\mathrm{lin}}_{\gamma}}[\mathcal{F}^*(u^{\mathrm{DLR}+\ell})]\Big\rangle, \quad r = 1, \dots, R \\
\sum_{j=1}^{R} (C_Y)_{jr} \dot{U}_j &= \Big(\mathbb{E}[\mathcal{F}^*(u^{\mathrm{DLR}+\ell})Y_r] - \mathcal{P}_U\Big[\mathbb{E}[\mathcal{F}^*(u^{\mathrm{DLR}+\ell})Y_r]\Big]\Big), \quad r = 1, \dots, R \\
& \qquad \textit{with } C_Y = \mathbb{E}[Y^\intercal Y].
\end{aligned}
\tag{7.36}
$$

Proof. The equation for the matrix A follows from (7.35) by considering $v = \delta A \gamma_k$ for some $\delta A \in \mathbb{R}^{N_h}$, $k \in \{1, \dots, N_h\}$. We derive

$$
\begin{aligned}
\langle \dot{A}_k, \delta A \rangle = \langle \dot{U}, \delta A \rangle \mathbb{E}[Y^\intercal \gamma_k] + \langle U, \delta A \rangle \mathbb{E}[\dot{Y}^\intercal \gamma_k] + \sum_{l=1}^{N_h} \langle \dot{A}_l, \delta A \rangle \mathbb{E}[\gamma_l \gamma_k] \\
= \Big\langle \mathbb{E}[\mathcal{F}^*(u^{\mathrm{DLR}+\ell})\gamma_k],\, \delta A \Big\rangle,
\end{aligned}
$$

from which the equation for A_k follows.

As the next step, we derive equations for the deterministic modes U. Consider a test function $v = \mathcal{P}^{\perp}_U[\delta U_r] Y_r$ for an arbitrary $\delta U_r \in \mathbb{R}^{N_h}$. From the variational formulation, we obtain

$$
\begin{aligned}
\sum_{j=1}^{R} \langle \dot{U}_j, \mathcal{P}^{\perp}_U[\delta U_r] \rangle \mathbb{E}[Y_j Y_r] + \langle U_j, \mathcal{P}^{\perp}_U[\delta U_r] \rangle \mathbb{E}[\dot{Y}_j Y_r] + \sum_{k=1}^{N_h} \langle \dot{A}_k, \mathcal{P}^{\perp}_U[\delta U_r] \rangle \mathbb{E}[\gamma_k Y_r] \\
= \langle \mathbb{E}[\mathcal{F}^*(u^{\mathrm{DLR}+\ell})], \mathcal{P}^{\perp}_U[\delta U_r] \rangle.
\end{aligned}
\tag{7.37}
$$

As $Y \in \Big((\mathcal{W}^{\mathrm{lin}}_{\gamma})^R\Big)^{\perp}$, we have that $\mathbb{E}[\gamma_k Y_r] = 0$. Using the symmetry of $\mathcal{P}^{\perp}_U$ w.r.t. the inner product $\langle \cdot, \cdot \rangle$ and the DO condition $\langle \dot{U}_l, U_k \rangle = 0$, $\forall l, k$, we obtain

$$
\sum_{j=1}^{R} \langle \dot{U}_j, \delta U_r \rangle \mathbb{E}[Y_j Y_r] = \sum_{j=1}^{R} \langle \mathcal{P}^{\perp}_U[\dot{U}_j], \delta U_r \rangle \mathbb{E}[Y_j Y_r] = \langle \mathcal{P}^{\perp}_U[\mathbb{E}[\mathcal{F}^*(u^{\mathrm{DLR}+\ell})]], \delta U_r \rangle,
$$

which is equivalent to the equation in the theorem. Lastly, we focus on the equation for the stochastic DLR modes $\{Y_r\}_{r=1}^R$. We start by the following observation

$$
\mathbb{E}[f\, \mathcal{P}_{\mathcal{W}^{\mathrm{lin}}_{\gamma}}[v]] = \mathbb{E}[f \sum_{k=1}^{N_h} \mathbb{E}[v\gamma_k]\gamma_k] = \mathbb{E}[\sum_{k=1}^{N_h} \mathbb{E}[f\gamma_k]\gamma_k\, v] = \mathbb{E}[\mathcal{P}_{\mathcal{W}^{\mathrm{lin}}_{\gamma}}[f]\, v], \qquad \forall f, v \in L^2(\Omega; \mathbb{R}),
$$

which consequently implies $\mathbb{E}[f\, \mathcal{P}^{\perp}_{\mathcal{W}^{\mathrm{lin}}_{\gamma}}[v]] = \mathbb{E}[\mathcal{P}^{\perp}_{\mathcal{W}^{\mathrm{lin}}_{\gamma}}[f]\, v]$.

Now, consider $v = U_r \delta Y_r$ with δY_r s.t. $\delta Y_r = \mathcal{P}^{\perp}_{\mathcal{W}^{\mathrm{lin}}_{\gamma}}[\delta Z]$ for arbitrary $\delta Z \in L^2(\Omega; \mathbb{R}^R)$. We derive from the variational formulation

$$\sum_{j=1}^{R} \langle \dot{U}_j, U_r \rangle \mathbb{E}[Y_j \mathcal{P}^{\perp}_{\mathcal{W}^{\mathrm{lin}}_{\gamma}}[\delta Z]] + \langle U_j, U_r \rangle \mathbb{E}[\dot{Y}_j\, \mathcal{P}^{\perp}_{\mathcal{W}^{\mathrm{lin}}_{\gamma}}[\delta Z]] + \sum_{k=1}^{N_h} \langle \dot{A}_k, U_r \rangle \mathbb{E}[\gamma_k\, \mathcal{P}^{\perp}_{\mathcal{W}^{\mathrm{lin}}_{\gamma}}[\delta Z]]$$
$$= \langle \mathbb{E}[\mathcal{F}^*(u^{\mathrm{DLR}+\ell}) \mathcal{P}^{\perp}_{\mathcal{W}^{\mathrm{lin}}_{\gamma}}[\delta Z]], U_r \rangle.$$

Applying the following relations

$$\langle \dot{U}, U \rangle = 0, \quad \langle U_j, U_r \rangle = \delta_{jr}, \quad \mathbb{E}[\gamma_k\, \mathcal{P}^{\perp}_{\mathcal{W}^{\mathrm{lin}}_{\gamma}}[\delta Z]] = 0, \quad \mathbb{E}[\dot{Y}_j\, \mathcal{P}^{\perp}_{\mathcal{W}^{\mathrm{lin}}_{\gamma}}[\delta Z]] = \mathbb{E}[\dot{Y}_j\, \delta Z]$$
$$\mathbb{E}[\mathcal{F}^*(u^{\mathrm{DLR}+\ell}) \mathcal{P}^{\perp}_{\mathcal{W}^{\mathrm{lin}}_{\gamma}}[\delta Z]] = \mathbb{E}[\mathcal{P}^{\perp}_{\mathcal{W}^{\mathrm{lin}}_{\gamma}}[\mathcal{F}^*(u^{\mathrm{DLR}+\ell})] \delta Z]$$

we obtain

$$\mathbb{E}[\dot{Y}_r\, \delta Z] = \mathbb{E}[\langle \mathcal{P}^{\perp}_{\mathcal{W}^{\mathrm{lin}}_{\gamma}}[\mathcal{F}^*(u^{\mathrm{DLR}+\ell})], U_r \rangle \delta Z], \qquad \forall j = 1, \ldots, R, \quad \forall \delta Z \in L^2(\Omega; \mathbb{R}^R),$$

which is equivalent to the equation stated in the theorem.

□

An important difference between DLR with independent and linear term is the function in which the operator $\mathcal{F}$ is evaluated. In the first method, in order to compute the DLR modes, the operator is evaluated only in the DLR component u^{DLR^*} (see equations (7.22)–(7.23)). This clearly ensures the independence condition but brings along a further approximation which might result in an accuracy loss during the time evolution. The second algorithm does not impose any restrictive condition during the evolution, apart from the standard DO condition $\langle \dot{U}_i, U_j \rangle = 0$, $\forall i, j = 1, \ldots, R$. The linear dependence on γ is tracked exactly. The operator $\mathcal{F}$ is, however, evaluated in the full signal $u^{\mathrm{DLR}+\ell}$. In order to deal with the quadratic term $B(u^{\mathrm{DLR}+\ell}, u^{\mathrm{DLR}+\ell})$, we need to compute the third moments of $u^{\mathrm{DLR}+\ell}$, which is avoided in the first algorithm.

Computational aspects

The evolution of $u^{\mathrm{DLR}+\ell}$ involves computing the third moments of the signal, including the mixed terms between u^{DLR^*} and $A\gamma^{\intercal}$. To make these accessible, we apply particle approximation for both terms, i.e. tracking $\{Y_{(j)}, \gamma_{(j)}\}_{j=1}^{\hat{N}}$, and replace all expectations $\mathbb{E}[\cdot]$ needed for (7.36) by $\mathbb{E}_{\hat{N}}[\cdot]$.

For high dimensional problems, tracking the full matrix $A \in \mathbb{R}^{N_h \times N_h}$ becomes unfeasible. In what follows, we address this issue by a low-rank approximation technique. Instead of tracking the linear dependence on γ, we only keep track of a linear dependence on $VV^{\intercal}\gamma$, formed by an orthonormal basis $V \in \mathbb{R}^{N_h \times M}$, with $M << N_h$.

Denoting $\tilde{B} = BV \in \mathbb{R}^{N_h \times M}$, $\tilde{\gamma}^\intercal = V^\intercal \gamma^\intercal = (\tilde{\gamma}_1, \dots, \tilde{\gamma}_M)^\intercal \in \mathbb{R}^M$, we define the space of all random functions linear w.r.t. $\tilde{\gamma}$

$$\mathcal{W}^{\text{lin}}_{\tilde{\gamma}} = \{f \in L^2_0(\mathbb{R}^{N_h}; \mathbb{R}) |\ f = B\tilde{\gamma}^\intercal,\ B \in \mathbb{R}^M\},$$

and analogously the projectors $\mathcal{P}_{\mathcal{W}^{\text{lin}}_{\tilde{\gamma}}}$, $\mathcal{P}^\perp_{\mathcal{W}^{\text{lin}}_{\tilde{\gamma}}}$. The complemented DLR solution $u^{\text{DLR}+\ell}$ will be of the form

$$u^{\text{DLR}+\ell}(t,\omega) = \bar{u}(t) + U(t)Y(t,\omega) + \tilde{A}(t)\tilde{\gamma}, \qquad \text{with } \tilde{A} \in \mathbb{R}^{N_h \times M},\ \{Y_r\}_{r=1}^R \subset \left((\mathcal{W}^{\text{lin}}_{\tilde{\gamma}})^R \right)^\perp. \tag{7.38}$$

As a consequence, the complexity of tracking the linear term is decreased to $O(N_h \times M)$. This is achieved by sacrificing the accuracy of the linear component of the signal. Disregarding the particle approximation error, the linear dependence of the signal on γ in the former definition was tracked exactly. In the new definition, linear dependence of the signal on $\tilde{\gamma}$ only is tracked exactly, the rest is approximated by DLR method. For completeness, we recall the dimension of all terms in $\{u^{\text{DLR}+\ell}_{(j)}\}_{j=1}^{\hat{N}}$

$$\bar{u} \in \mathbb{R}^{N_h}, \quad U \in \mathbb{R}^{N_h \times R},\ Y \in \hat{N} \times R, \quad \tilde{A} \in \mathbb{R}^{N_h \times M},\ \tilde{\gamma} \in \mathbb{R}^{M \times \hat{N}}.$$

Note that by setting $M = N_h$, we recover the former definition, and by setting $M = 0$, we recover the simple DLR approximation u^{DLR} of Section 7.1.4.

There are many options for the choice of the basis V. We propose M eigenvectors of matrix C^n corresponding to the M most dominant eigenvalues, where C^n is the covariance matrix from the GM at the start of the forecast step. With this choice we prescribe the initial condition of matrix $\tilde{A}$ in the following way

$$\tilde{A}^n = V\sqrt{\Lambda},$$

where $\sqrt{\Lambda} \in \mathbb{R}^{M \times M}$ is a diagonal matrix with the square roots of the M most dominant eigenvalues of C^n on its diagonal and each column of V constitutes a corresponding eigenvector. This results in the best M-rank approximation of A^n.

7.2.3 Including the model error

In the previous subsection we described two approximate methods that evolve the signal through (6.3). The first method, that keeps the u^{DLR} term and the linear term $A\gamma^\intercal$ independent, naturally constitutes a GMM

$$\mathbb{P}(u^{n+1}|Z^n) = \sum_{j=1}^{\hat{N}} \hat{\lambda}^{n+1}_{(j)} \times N(\hat{u}^{\text{DLR},n+1}_{(j)}, \hat{T}^{n+1}),$$

$$\hat{u}_{(j)}^{\mathrm{DLR},n+1} = \hat{\bar{u}}^{n+1} + \hat{U}^{n+1}\hat{Y}_{(j)}^{n+1\intercal},$$

where $\hat{\bar{u}}^{n+1}, \hat{U}^{n+1}, \{\hat{Y}_{(j)}^{n+1}\}_{j=1}^{\hat{N}}, \hat{T}^{n+1}$ are solutions of the equations (7.22)–(7.23) at time $t = t^{n+1}$.

The second method applies a particle approximation for both DLR and linear term and results in

$$\{\hat{u}_{(j)}^{n+1}, \hat{\lambda}_{(j)}^{n+1}\}_{j=1}^{\hat{N}} \quad \text{with} \quad \hat{u}_{(j)}^{n+1} = \hat{\bar{u}}^{n+1} + \hat{U}^{n+1}\hat{Y}_{(j)}^{n+1} + \hat{A}^{n+1}\gamma_{(j)},$$

where $\hat{\bar{u}}^{n+1}, \hat{U}^{n+1}, \{\hat{Y}_{(j)}^{n+1}\}_{j=1}^{\hat{N}}, \hat{A}^{n+1}$ are solutions of the equations (7.36) at time $t = t^{n+1}$. Similarly to the first method, we build a GM at the end of the forecast step as

$$\begin{aligned} \mathbb{P}(u^{n+1}|Z^n) &\approx \sum_{j=1}^{\hat{N}} \hat{\lambda}_{(j)}^{n+1} \times N(\hat{u}_{(j)}^{\mathrm{DLR},n+1}, \hat{A}^{n+1}\hat{A}^{n+1\intercal}), \\ \hat{u}_{(j)}^{\mathrm{DLR},n+1} &= \hat{\bar{u}}^{n+1} + \hat{U}^{n+1}\hat{Y}_{(j)}^{n+1\intercal}. \end{aligned}$$

However, note that this step is only approximate, since we treat the u^{DLR} term and the linear term $A\gamma^{\intercal}$ as if they were independent at the end of the evolution.

For both algorithms, the model error is then incorporated into the mixture covariance and the forecast step results in the Gaussian mixture

$$\mathbb{P}(u^{n+1}|Z^n) \approx \sum_{j=1}^{\hat{N}} \hat{\lambda}_{(j)}^{n+1} \times N(\hat{u}_{(j)}^{\mathrm{DLR},n+1}, \hat{C}^{n+1}), \tag{7.39}$$

where

$$\begin{aligned} \hat{u}_{(j)}^{\mathrm{DLR},n+1} &= \hat{\bar{u}}^{n+1} + \hat{U}^{n+1}\hat{Y}_{(j)}^{n+1\intercal} \\ \hat{C}^{n+1} &= \hat{A}^{n+1}\hat{A}^{n+1\intercal} + \Sigma. \end{aligned}$$

7.2.4 Analysis step

The analysis step takes in a Gaussian mixture of the form (7.39). Applying Lemma 6.2.2, we obtain the filtering distribution at time t^{n+1} which is again a Gaussian mixture

$$\mathbb{P}(u^{n+1}|Z^{n+1}) \approx \sum_{j=1}^{\hat{N}} \lambda_{(j)}^{n+1} \times N(m_j^{n+1}, C^{n+1}), \tag{7.40}$$

with $\lambda_{(j)}^{n+1}, m_j^{n+1}, C^{n+1}$ computed with formulas (6.23).

Remark 8. The first method - the DLRA complemented by an independent term strongly resembles a method introduced in [MQS14; QJM15; SM13], called the QG-DO method.

The authors proposed a method for the forecast step of the filtering problem, where they run evolution equations for the mean and covariance matrix of the full signal u, alongside with evolution equations for the DLR modes $\{U_r, Y_r\}_{r=1}^R$, which are used to approximate the third moments required in the evolution of the full covariance matrix. There is a natural question arising, whether these two methods are not, in fact, equivalent. The evolution equations for the mean value $\bar{u}$ as well as the evolution equations for the DLR modes in QG-DO are equivalent to the equations stated in Theorem 7.2.1. In this remark, we compare the evolution equation for the full covariance matrix of QG-DO, denoted by $\tilde{S} = \mathbb{E}[(u-\bar{u})(u-\mathbb{E}[u])^\intercal]$, with the full covariance matrix for the signal $u^{\mathrm{DLR+i}}$, denoted by S. We will see that for the quadratic problem considered here, these matrices differ.

The evolution of $\tilde{S}$ is given by

$$\dot{\tilde{S}} = L_{\bar{u}}\tilde{S} + \tilde{S}L_{\bar{u}}^\intercal + Q_F \text{ with } (L_{\bar{u}})_{ij} = (L_{ij} + D_{ij}) + \langle B(\bar{u}, v_j), v_i\rangle + \langle B(v_j, \bar{u}), v_i\rangle$$

and

$$\begin{aligned}(Q_F)_{ij} &= \sum_{r,j,l} \langle v_i, B(U_j, U_l)\rangle \mathbb{E}[Y_j Y_l Y_r]\langle U_r, v_j\rangle + \langle v_j, B(U_j, U_l)\rangle \mathbb{E}[Y_j Y_l Y_r]\langle U_r, v_i\rangle \\ &= \Big(\mathbb{E}[B(u^{\mathrm{DLR}^*}, u^{\mathrm{DLR}^*}) u^{\mathrm{DLR}^{*\intercal}}]\Big)_{ij} + \Big(\mathbb{E}[u^{\mathrm{DLR}^*} B(u^{\mathrm{DLR}^*}, u^{\mathrm{DLR}^*})^\intercal]\Big)_{ij}.\end{aligned}$$

Now, let us denote by $\Pi_{UY^\intercal} : L^2(\Omega; \mathbb{R}^{N_h}) \to L^2(\Omega; \mathbb{R}^{N_h})$ the operator

$$\begin{aligned}&\Pi_{UY^\intercal}[v] = \mathcal{P}_U[v] + \mathcal{P}_U^\perp[\mathcal{P}_Y[v]], \\ &\qquad\qquad\text{where } \mathcal{P}_U[v] = \sum_{r=1}^R \langle U_r, v\rangle U_r,\ \mathcal{P}_Y[v] = \sum_{k,r=1}^R \mathbb{E}[Y_k v](C_Y)_{kr}^{-1} Y_r.\end{aligned}$$

More details on the geometrical interpretation of the operator $\Pi_{UY^\intercal}$ can be found in Chapter 1. It is easy to see from Theorem 7.2.1, that the evolution equation for $u^{\mathrm{DLR}^*} = UY^\intercal$ satisfies

$$(\dot{UY^\intercal}) = \Pi_{UY^\intercal}[\mathcal{F}^*(u^{\mathrm{DLR}})].$$

Now we follow with the evolution equation of the full covariance matrix S of the signal $u^{\mathrm{DLR+i}}$:

$$\begin{aligned}\dot{S} &= \mathbb{E}[(\dot{UY^\intercal})(UY^\intercal)^\intercal] + \mathbb{E}[(UY^\intercal)(\dot{UY^\intercal})^\intercal] + L_u T + T L_u^\intercal \\ &= \mathbb{E}\Big[\Pi_{UY^\intercal}[\mathcal{F}^*(u^{\mathrm{DLR}})](UY^\intercal)^\intercal\Big] + \mathbb{E}\Big[(UY^\intercal)\Pi_{UY^\intercal}[\mathcal{F}^*(u^{\mathrm{DLR}})]^\intercal\Big] + L_u T + T L_u^\intercal \\ &= \mathbb{E}\Big[\Big((L+D)UY^\intercal + B(\bar{u}, UY^\intercal) + B(UY^\intercal, \bar{u}) + \Pi_{UY^\intercal}[B(UY^\intercal, UY^\intercal)]\Big)(UY^\intercal)^\intercal\Big] \\ &\quad + \mathbb{E}\Big[(UY^\intercal)\Big((L+D)UY^\intercal + B(\bar{u}, UY^\intercal) + B(UY^\intercal, \bar{u}) + \Pi_{UY^\intercal}[B(UY^\intercal, UY^\intercal)]\Big)^\intercal\Big] \\ &\quad + L_u T + T L_u^\intercal \\ &= L_u U \mathbb{E}[Y^\intercal Y] U^\intercal + U \mathbb{E}[Y^\intercal Y] U^\intercal L_u^\intercal + L_u T + T L_u^\intercal\end{aligned}$$

$$\begin{aligned}&+\mathbb{E}\Big[\Pi_{UY^\intercal}[B(UY^\intercal,UY^\intercal)](UY^\intercal)^\intercal\Big]+\mathbb{E}\Big[(UY^\intercal)\Pi_{UY^\intercal}[B(UY^\intercal,UY^\intercal)]^\intercal\Big]\\ &=L_uS+SL_u^\intercal+Q,\end{aligned}$$

where

$$(Q)_{ij}=\Big(\mathbb{E}\Big[\Pi_{UY^\intercal}[B(u^{\mathrm{DLR}*},u^{\mathrm{DLR}*})]u^{\mathrm{DLR}*^\intercal}\Big]\Big)_{ij}+\Big(\mathbb{E}\Big[u^{\mathrm{DLR}*}\Pi_{UY^\intercal}[B(u^{\mathrm{DLR}*},u^{\mathrm{DLR}*})]^\intercal\Big]\Big)_{ij}.$$

Comparing the equation for $\tilde{S}$ and S we see, that the third order terms are approximated differently. Both methods use the knowledge of the DLR modes but in our method, the quadratic operator gets projected, whilst in the QG-DO method, it does not. On the other hand, our approach allows for a variational formulation which brings along a geometrical insight. Furthermore, interpreting the evolved signal as a DLR approximation complemented by an independent term results in a natural characterization via GM with a mixture covariance constant across the mixtures. In fact, to proceed with the analysis step, the authors of QG-DO in [MQS14] propose to build a conditional Gaussian particle distribution, with a constant covariance matrix across particles. The build-up is, however, much less straightforward. The analysis step then applies the Bayes' formula.

7.2.5 Numerical results

To asses the quality of the proposed algorithms, we provide two test cases with different sets of parameters.

Test case I: frequent observations, small observation error

In the first experiment, we examine the proposed algorithms on the test case from Sections 7.1.3 and 7.1.5, i.e. setting the final time $T = 100$, time between observations $\triangle t = 0.05$, the model error covariance matrix $\Sigma = 10^{-4} \cdot \mathrm{Id}$, the observation error covariance matrix $\Gamma = 10^{-2} \cdot \mathrm{Id}$, the observation operator $H = \mathrm{Id}$ and number of particles $\hat{N} = 1000$.

In Figure 7.5 we show the obtained RMS errors. We see that applying DLR with $R = 10$ complemented by an independent Gaussian approximation (orange) helps to correct the signal obtained by a simple DLR (blue) and achieves a sufficient accuracy. In fact, reducing the DLR approximation to $R = 5$ still maintains the desired accuracy. Applying the DLR complemented by a full-rank linear term (red) does not further improve the performance of the filter. In conclusion, tracking the third moments of the signal is not necessary for this test case and DLR complemented by an independent Gaussian approximation provides results with a satisfactory accuracy.

In Figure 7.6 we focus on the second approach and quantify the effect of different low-rank approximations of the linear term. We see that providing a good approximation for the linear part ($M \geq 10$) of the signal seems to be more important than an accurate

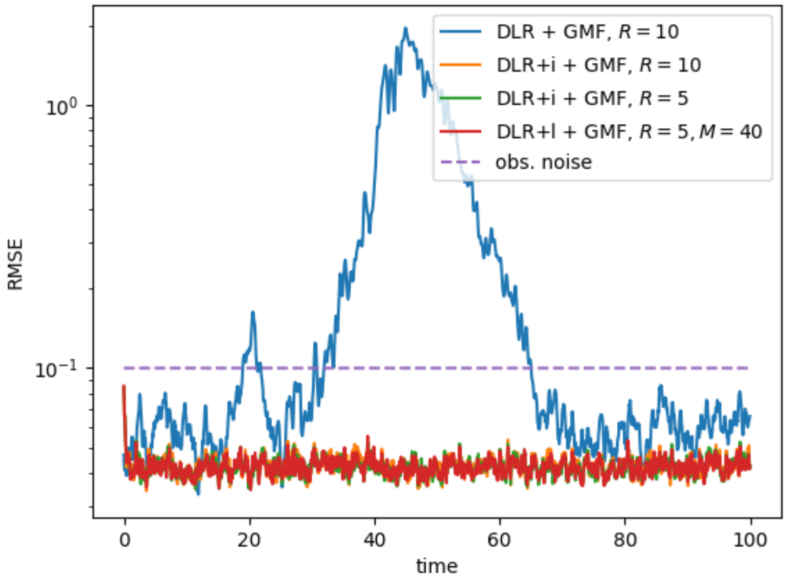


Figure 7.5 – Comparison of RMSEs for filtering algorithms with prediction step realized by simple DLR with $R = 10$ vs DLR with $R = 5, 10$ complemented by independent (DLR+i) or full-rank linear (DLR+l) term. The analysis step applies GMF in all cases.

approximation of the higher-order terms. Since the observations are full, with a small observation error and very frequent in time, they provide highly informative data. With the time step being very small ($\triangle t = 0.05$), the nonlinear interactions in the dynamics do not play a significant role and consequently the signal is sufficiently well approximated by a Gaussian approximation and a DLR term with small R.

Test case II: rare observations, high observation error

Our second test case tries to examine the performance of both algorithms in a very different scenario. We have less frequent full observations with a high observation error. More specifically, we set the final time $T = 100$, time between observations $\triangle t = 0.5$, the model error covariance matrix $\Sigma = 0$, the observation error covariance matrix $\Gamma = 3. \cdot \mathrm{Id}$, the observation operator $H = \mathrm{Id}$ and number of particles $\hat{N} = 1000$. This setting provides not very informative data and the filtering of such problem is difficult. With $\Sigma = 0$ the optimal proposal particle filter becomes the bootstrap particle filter, whose behaviour is highly dependent on the number of particles (see Fig. 6.3). Observing a significant difference between the DLR with $R = 10$ complemented by independent term (orange) and by full-rank liner term (brown) in Figure 7.7, we conclude that in this scenario tracking the third moments has a notable impact on the resulting approximation. The signal can be comparably well approximated by $R = 10$ and $M = 15$.

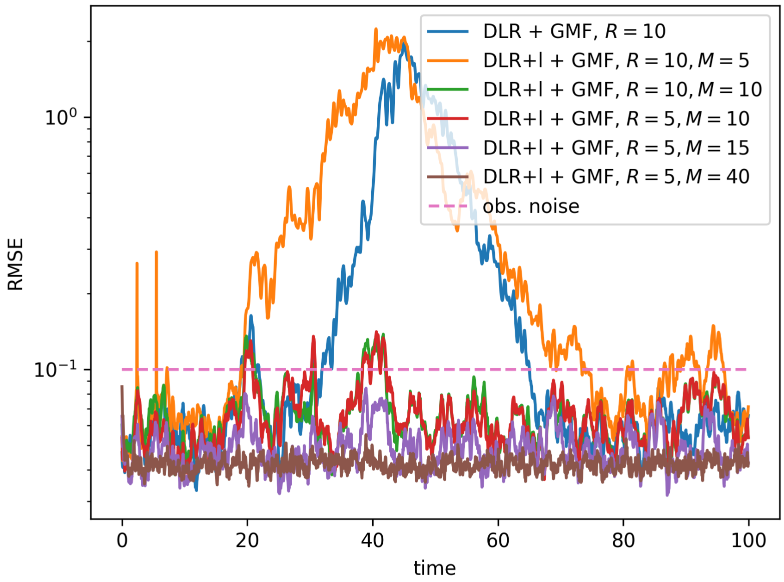


Figure 7.6 – Comparison of RMSEs for filtering algorithms with prediction step realized by simple DLR with $R = 10$ vs DLR with $R = 5, 10$ complemented by low-rank approximation of the linear term (DLR+l). The analysis step applies GMF in all cases.

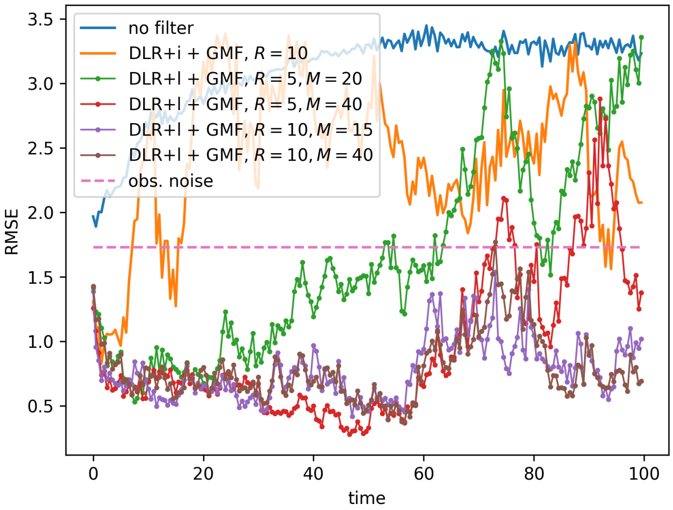


Figure 7.7 – Comparison of RMS errors for filtering algorithms with prediction step realized by DLR with $R = 5, 10$ complemented by independent (DLR+i) or both full-rank and low-rank linear (DLR+l) term. The analysis step applies GMF in all cases.

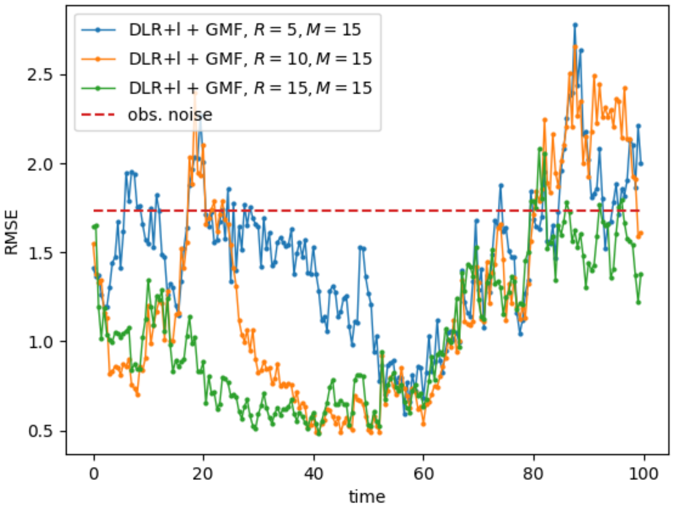


Figure 7.8 – Comparison of RMS errors for filtering algorithms with prediction step realized by DLR with $R = 5, 10, 15$ complemented by low-rank linear term (DLR+l) with $M = 15$. The dimension of the observations is 17. The analysis step applies GMF in all cases.

In Figure 7.8 we report numerical results obtained by running the same test case with partial observations. The operator H observes every third value of the 40-dimensional signal u, resulting in the dimension of observations $l = 17$. The signal can be tracked by 15-rank DLR approximation and 15-rank approximation of the linear term. Note that for a test case with partial observations, there is no expectation for the RMSE to be under the observation noise.

8 Conclusions and perspectives

This thesis covers most of the research work that the author has carried out during her Ph.D. studies at EPFL. The work is divided into two parts; the first one concerns an analysis of discretization schemes for dynamical low-rank approximation (DLRA), the second one deals with applying DLRA in data assimilation.

In the first work, we proposed and analyzed three discretization schemes, namely explicit, implicit and semi-implicit, to obtain a numerical solution of the DLR system of evolution equations for the deterministic and stochastic modes. Such discrete DLR solution was obtained by projecting the discretized dynamics on the tangent space of the low-rank manifold at an intermediate point. This point was built using the new-computed deterministic modes and old stochastic modes. We found this projection property to be useful when investigating the stability of the DLR solution. The solution obtained by the implicit scheme remains unconditionally bounded by the data in suitable norms. Concerning the explicit and semi-implicit schemes, we derived stability conditions on the time step, independent of the smallest singular value, under which the solution remains bounded. Remarkably, applying the proposed semi-implicit scheme to a random heat equation with diffusion coefficient affine with respect to random variables results in a scheme unconditionally stable, with the same computational complexity as the explicit scheme. Our theoretical derivations are supported by numerical tests applied to a random heat equation with a zero-forcing term. In the semi-implicit case, we observed that the norm of the solution consistently decreases for every time step considered. In the explicit case, our numerical results suggest that our theoretical stability condition on the time step is, in fact, sharp. Our future work includes investigating if the proposed approach can be extended to higher-order projector-splitting integrators or used to show stability properties for other types of equations.

Our second work involved deriving a-priori and a-posteriori error estimates for the fully discrete DLR solution obtained by the discretization schemes proposed in the first project. The projection property turns out to be the key element in both estimations. Concerning the a-priori error estimation, the problem is discretized by FEM in space, with piece-wise polynomials of degree $\leq r$, Monte Carlo (MC) method in stochastic space and follows the staggered time-marching scheme (described in Section 2.2). Under the approximability assumption, that the operator $\mathcal{F}(u)$ maps onto the tangent space of $\mathcal{M}_R$ at u up to a small reminder of size ε, we proved the first-order convergence in time, r-th order convergence in space and the standard $(-\frac{1}{2})$-order of convergence w.r.t. the number of MC samples. All the considered constants are independent of the smallest singular value of the solution. The future directions include proving higher-order convergence rates for higher-order projector-splitting integrators. Another open question is the possibility of alleviating the ε-approximability assumption.

Concerning the a-posteriori error estimation, we started with a residual-based a posteriori error estimation for a heat equation with a random forcing term and a random diffusion coefficient dependent on a finite number of independent random variables, with no DLR approximation involved. Moreover, the dependency of the diffusion coefficient is assumed to be affine. This problem was discretized by a θ-scheme in time, FEM in physical space, and sparse grid collocation method in stochastic space, which required the use of nested collocation points. The estimate consisted of three parts accounting for the FEM error, time discretization error, and stochastic error, respectively. The derivation is valid for the case of time-varying FE meshes and time-varying sparse grids, allowing for both refinement and coarsening. We proposed an adaptive algorithm for the choice of time discretization, FE mesh, and sparse grid, where the mesh and sparse grid are fixed in time which simplifies the computation of the estimators. The estimators are localized on each element of the FE mesh, each time step, and each index from the margin of the sparse grid index set, and we perform a refinement whenever the localized estimate is higher than a prescribed condition (see Algorithm 1). We studied the effectiveness of the estimators over non-uniform time discretizations, non-uniform meshes, and anisotropic sparse grids, fixed in time and applied the adaptive algorithm to a problem with a deterministic time independent forcing term. This algorithm is one possible strategy. Several other versions could be considered as well, for instance, to allow for coarsening in the adaptive process for a more uniform distribution of the error. We believe that the derived error estimates could provide a reliable basis for error estimation and adaptation strategies that include time-varying FE meshes and sparse grids. One could, for example, drive an adaptive choice of time-varying meshes and sparse girds by localizing the spatial and stochastic estimator for a specific time step, as was proposed in [Pic98; BR03] for time-varying FE or DG meshes in the case of a deterministic heat equation.

This work was then extended to treat the a-posteriori error estimation for a DLRA of a random heat equation, again with the diffusion coefficient affine w.r.t. a finite number of random variables. This problem is discretized by a FEM in physical space, tensor grid collocation method in stochastic space, and the staggered time-marching scheme (described in Section 2.2) for time integration. The estimate consists of parts accounting for the FEM error, time discretization error, stochastic discretization error, and rank truncation error. The derivation only holds for fixed-in-time FE meshes and tensor grids but could be extended to time-varying FE meshes and tensor grids. The rank is allowed to change between the time intervals. The estimators are further localized on each element of the FE mesh for the space estimator, each time step for the time estimator, each dimension of the stochastic space for the stochastic error estimator, and each time step for the rank truncation estimator. These localized error estimates can be used to drive an adaptive algorithm for the choice of time discretization, FE mesh, tensor grid, and time-varying rank. We proposed an algorithm that performs a refinement whenever the localized estimate is higher than a prescribed condition (see Algorithm 2) and discussed various ways of updating the rank. The implementation of this algorithm is a part of an ongoing project, which includes a comparison of the different approaches when the rank gets increased.

The last project corresponds to the second part of the thesis and is concerned with applying DLR to data assimilation, particularly the filtering problem. The filtering problem can be split into two steps: the forecast and the analysis. The forecast step involves solving an often high-dimensional dynamical system of (stochastic) equations; the analysis step incorporates data into the solution (also called the signal). In real-world applications, computing the full system is unfeasible, and one needs to rely on some model-order reduction technique. In our work, we examined the idea of applying the DLRA in the forecast step. We started with the simple DLRA in the forecast step, combined with ensemble Kalman filter or particle filter in the analysis step. We observed that completely disregarding the omitted modes in the DLR approximation leads to unsatisfactory results. To alleviate this issue, we proposed two new algorithms that complement the DLRA with a Gaussian component, expressed as a part of the signal, which is linear w.r.t. normally distributed random variables. The first algorithm assumed the Gaussian component to be independent of the DLR part, the second one did not. We applied both algorithms to deal with the filtering problem for a 40-dimensional Loren-96 system of equations (a simplified mathematical model of atmospheric processes) and compared their performance in different scenarios. The first algorithm involves evolving a full covariance matrix. The future directions comprise a low-rank approximation or a particle approximation of the full covariance matrix. Furthermore, it would be very interesting to see numerical results applying the proposed algorithm to a very high-dimensional problem arising from real-world applications.

Bibliography

[Aba13] H. Abarbanel. *Predicting the Future: Completing Models of Observed Complex Systems.* Springer-Verlag New York, 2013. DOI: 10.1007/978-3-642-03711-5.

[Aln+15a] M. S. Alnæs et al. “The FEniCS Project Version 1.5”. In: *Arch. of Numer. Softw.* 3.100 (2015). DOI: 10.11588/ans.2015.100.20553.

[Aln+15b] M. S. Alnæs et al. “The FEniCS Project Version 1.5”. In: *Archive of Numerical Software* 3.100 (2015). DOI: 10.11588/ans.2015.100.20553.

[AMN06] G. Akrivis, C. Makridakis, and R. H. Nochetto. “A posteriori error estimates for the Crank-Nicolson method for parabolic equations”. In: *Math. Comput.* 75 (2006), pp. 511–531.

[And01] J. L. Anderson. “An Ensemble Adjustment Kalman Filter for Data Assimilation”. In: *Monthly Weather Review* 129.12 (2001), pp. 2884 –2903.

[Bau+15] U. Baur et al. “Comparison of methods for parametric model order reduction of instationary problems”. In: 2015.

[Bau11] H. Bauer. *Probability Theory.* De Gruyter, 2011. DOI: 10.1515/9783110814668.

[Bec+12] J. Beck et al. “On the optimal polynomial approximation of stochastic PDEs by galerkin and collocation methods”. In: *Mathematical Models and Methods in Applied Sciences* (2012). DOI: 10.1142/S0218202512500236.

[Bec+99] M. Beck et al. “The multiconfiguration time-dependent Hartree (MCTDH) method: A highly efficient algorithm for propa”. In: 1999.

[BEM01] C. H. Bishop, B. J. Etherton, and S. J. Majumdar. “Adaptive Sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical Aspects”. In: *Monthly Weather Review* 129.3 (2001), pp. 420 –436.

[Ben02] A. Bennett. *Inverse Modeling of the Ocean and Atmosphere.* Cambridge University Press, 2002. DOI: 10.1017/CBO9780511535895.

[BFFN21] M. Billaud-Friess, A. Falcó, and A. Nouy. “A new splitting algorithm for dynamical low-rank approximation motivated by the fibre bundle structure of matrix manifolds”. In: *BIT Numerical Mathematics* (2021).

[BGW07] J. V. Burkardt, M. D. Gunzburger, and C. Webster. “Reduced order modeling of some nonlinear stochastic partial differential equations”. In: 2007.

[BKU21] M. Bachmayr, E. Kieri, and A. Uschmajew. “Existence of dynamical low-rank approximations to parabolic problems”. In: *Math. Comput.* 90 (2021), pp. 1799–1830.

[BL12] A. Barth and A. Lang. “Multilevel Monte Carlo method with applications to stochastic partial differential equations”. In: *International Journal of Computer Mathematics* 89.18 (2012), pp. 2479–2498. DOI: 10.1080/00207160.2012.701735.

[BNR00] V. Barthelmann, E. Novak, and K. Ritter. “High dimensional polynomial interpolation on sparse grids”. In: *Advances in Computational Mathematics* 12.4 (2000), pp. 273–288. DOI: 10.1023/A:1018977404843.

[BNT10] I. Babuška, F. Nobile, and R. Tempone. “A Stochastic Collocation Method for Elliptic Partial Differential Equations with Random Input Data”. In: *SIAM Review* 52.2 (2010), pp. 317–355. DOI: 10.1137/100786356.

[Boy+10] S. Boyaval et al. “Reduced Basis Techniques for Stochastic Problems”. In: *Archives of Computational Methods in Engineering* 17 (2010), pp. 435–454.

[BPS14] A. Bespalov, C. Powell, and D. Silvester. “Energy Norm A Posteriori Error Estimation for Parametric Operator Equations”. In: *SIAM Journal on Scientific Computing* 36.2 (2014), A339–A363. DOI: 10.1137/130916849.

[BR03] G Bangerth and R. Rannacher. *Adaptive Finite Element Methods for Solving Differential Equations*. Birkhäuser, Basel, 2003.

[BR14] C. Bedregal and M.-C. Rivara. “Longest-edge algorithms for size-optimal refinement of triangulations”. In: *Computer-Aided Design* 46 (2014). 2013 SIAM Conference on Geometric and Physical Modeling, pp. 246 –251. DOI: 10.1016/j.cad.2013.08.040.

[BTZ04] I. Babuška, R. Tempone, and G. E. Zouraris. “Galerkin Finite Element Approximations of Stochastic Elliptic Partial Differential Equations”. In: *SIAM Journal on Numerical Analysis* 42.2 (2004), pp. 800–825. DOI: 10.1137/S0036142902418680.

[Buf+12] A. Buffa et al. “A priori convergence of the greedy algorithm for the parametrized reduced basis method”. In: *Mathematical Modelling and Numerical Analysis* 46 (2012), pp. 595–603.

[BX20] A. Bespalov and F. Xu. “A posteriori error estimation and adaptivity in stochastic Galerkin FEM for parametric elliptic PDEs: Beyond the affine case”. In: *Comput. Math. Appl.* 80 (2020), pp. 1084–1103.

[Bäc+11] J. Bäck et al. “Stochastic Spectral Galerkin and Collocation Methods for PDEs with Random Coefficients: A Numerical Comparison”. In: *Spectral and High Order Methods for Partial Differential Equations.* Ed. by J. S. Hesthaven and E. M. Rønquist. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 43–62.

[Caf98] R. E. Caflisch. “Monte Carlo and quasi-Monte Carlo methods”. In: *Acta Numerica* 7 (1998), p. 149.

[Cas+20] C. Q. Casas et al. “A Reduced Order Deep Data Assimilation model”. In: *Physica D: Nonlinear Phenomena* 412 (2020), p. 132615. DOI: 10.1016/j.physd.2020.132615.

[CB01] G. Casella and R. Berger. *Statistical Inference. 2nd Edition.* Duxbury Press, 2001.

[CCS14] A. Chkifa, A. Cohen, and C. Schwab. “High-Dimensional Adaptive Sparse Polynomial Interpolation and Applications to Parametric PDEs”. In: *Foundations of Computational Mathematics* 14.4 (2014), pp. 601–633.

[CD02] D. Crisan and A. Doucet. “A survey of convergence results on particle filtering methods for practitioners”. In: *IEEE Transactions on Signal Processing* 50.3 (2002), pp. 736–746.

[CF11] K. Carlberg and C. Farhat. “A low-cost, goal-oriented compact proper orthogonal decomposition basis for model reduction of static systems”. In: *Int. J. Num. Methods Eng.* 86.3 (2011), pp. 381–402.

[CHZ13a] M. Cheng, T. Y. Hou, and Z. Zhang. “A dynamically bi-orthogonal method for time-dependent stochastic partial differential equations I: Derivation and algorithms”. In: *J. Comput. Phys.* 242 (2013), pp. 843–868. DOI: 10.1016/j.jcp.2013.02.033.

[CHZ13b] M. Cheng, T. Y. Hou, and Z. Zhang. “A dynamically bi-orthogonal method for time-dependent stochastic partial differential equations II: Adaptivity and generalizations”. In: *J. Comput. Phys.* 242 (2013), pp. 753–776. DOI: 10.1016/j.jcp.2013.02.020.

[CKL22] G. Ceruti, J. Kusch, and C. Lubich. “A rank-adaptive robust integrator for dynamical low-rank approximation”. In: *BIT Numerical Mathematics* (2022). DOI: 10.1007/s10543-021-00907-7.

[CL10] D. Conte and C. Lubich. “An error analysis of the multi-configuration time-dependent Hartree method of quantum dynamics”. In: *Mathematical Modelling and Numerical Analysis* 44 (2010), pp. 759–780.

[CL19] G. Ceruti and C. Lubich. “Time integration of symmetric and anti-symmetric low-rank matrices and Tucker tensors”. In: *BIT Numerical Mathematics* (2019), pp. 1–24.

[CL21] G. Ceruti and C. Lubich. “An unconventional robust integrator for dynamical low-rank approximation”. In: *BIT Numerical Mathematics* 0.1007/s10543-021-00873-0 (2021), pp. 1572–9125.

[Cli+11] K. A. Cliffe et al. “Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients - Computing and Visualization in Science”. In: *Comput. Visual Sci.* 14.3 (2011). DOI: 10.1007/s00791-011-0160-x.

[CLW21] G. Ceruti, C. Lubich, and H. Walach. “Time Integration of Tree Tensor Networks”. In: *SIAM J. Numer. Anal.* 59 (2021), pp. 289–313.

[CM47] R. H. Cameron and W. T. Martin. “The Orthogonal Development of Non-Linear Functionals in Series of Fourier-Hermite Functionals”. In: *Annals of Mathematics* 48 (1947), p. 385.

[Con20] D. Conte. “Dynamical low-rank approximation to the solution of parabolic differential equations”. In: *Applied Numerical Mathematics* 156 (2020), pp. 377–384. DOI: https://doi.org/10.1016/j.apnum.2020.05.011.

[CPB19] A. J. Crowder, C. E. Powell, and A. Bespalov. “Efficient Adaptive Multilevel Stochastic Galerkin Approximation Using Implicit A Posteriori Error Estimation”. In: *SIAM Journal on Scientific Computing* 41.3 (2019), A1681–A1705. DOI: 10.1137/18M1194420.

[CQR13] P. Chen, A. Quarteroni, and G. Rozza. “A Weighted Reduced Basis Method for Elliptic Partial Differential Equations with Random Input Data”. In: *SIAM J. Numer. Anal.* 51 (2013), pp. 3163–3185.

[CQR14] P. Chen, A. Quarteroni, and G. Rozza. “Comparison Between Reduced Basis and Stochastic Collocation Methods for Elliptic Problems”. In: *Journal of Scientific Computing* 59 (2014), pp. 187–216.

[CQR15] P. Chen, A. Quarteroni, and G. Rozza. “Reduced order methods for uncertainty quantification problems”. In: *The annual research report* (2015).

[CS15] P. Chen and C. Schwab. “Model Order Reduction Methods in Computational Uncertainty Quantification”. In: 2015.

[CSK14] M. Choi, T. P. Sapsis, and G. E. Karniadakis. “On the equivalence of dynamically orthogonal and bi-orthogonal methods: Theory and numerical simulations”. In: *J. Comput. Phys.* 270 (2014), pp. 1–20. DOI: 10.1016/j.jcp.2014.03.050.

[Dau+99] R. Dautray et al. *Mathematical Analysis and Numerical Methods for Science and Technology: Volume 5 Evolution Problems I*. Mathematical Analysis and Numerical Methods for Science and Technology. Springer Berlin Heidelberg, 1999.

[EG04a] A. Ern and J.-L. Guermond. "Finite Element Interpolation". In: *Theory and Practice of Finite Elements.* New York, NY: Springer New York, 2004, pp. 3–80. DOI: 10.1007/978-1-4757-4355-5_1.

[EG04b] A. Ern and J.-L. Guermond. "Time-Dependent Problems". In: *Theory and Practice of Finite Elements.* New York, NY: Springer New York, 2004, pp. 279–334. DOI: 10.1007/978-1-4757-4355-5_6.

[EHW21] L. Einkemmer, J. Hu, and Y. Wang. "An asymptotic-preserving dynamical low-rank method for the multi-scale multi-dimensional linear transport equation". In: *J. Comput. Phys.* 439 (2021), p. 110353.

[EHY21] L. Einkemmer, J. Hu, and L. Ying. "An Efficient Dynamical Low-Rank Algorithm for the Boltzmann-BGK Equation Close to the Compressible Viscous Flow Regime". In: *SIAM Journal on Scientific Computing* 43.5 (2021), B1057–B1080. DOI: 10.1137/21M1392772.

[Eig+13] M. Eigel et al. "Residual-based a posteriori error estimation for stochastic Galerkin finite element methods". In: *CMAME* (Nov. 2013).

[Eig+14] M. Eigel et al. "Adaptive stochastic Galerkin FEM". In: *Computer Methods in Applied Mechanics and Engineering* 270 (2014), pp. 247 –269. DOI: 10.1016/j.cma.2013.11.015.

[Eig+15] M. Eigel et al. "A convergent adaptive stochastic Galerkin finite element method with quasi-optimal spatial meshes". In: *ESAIM: M2AN* 49.5 (2015), pp. 1367–1398. DOI: 10.1051/m2an/2015017.

[Eig+21] M. Eigel et al. *On the convergence of adaptive stochastic collocation for elliptic partial differential equations with affine diffusion.* arXiv:2008.07186. 2021.

[Ein19] L. Einkemmer. "A Low-Rank Algorithm for Weakly Compressible Flow". In: *SIAM J. Sci. Comput.* 41.5 (2019), A2795–A2814. DOI: 10.1137/18M1185417.

[EJ21] L. Einkemmer and I. Joseph. "A mass, momentum, and energy conservative dynamical low-rank scheme for the Vlasov equation". In: *J. Comput. Phys.* 443 (2021), p. 110495.

[EJ91] K. Eriksson and C. Johnson. "Adaptive Finite Element Methods for Parabolic Problems I: A Linear Model Problem". In: *SIAM Journal on Numerical Analysis* 28.1 (1991), pp. 43–77. DOI: 10.1137/0728003.

[EJ95] K. Eriksson and C. Johnson. "Adaptive Finite Element Methods for Parabolic Problems II: Optimal Error Estimates in $L_\infty L_2$ and $L_\infty L_\infty$". In: *SIAM Journal on Numerical Analysis* 32.3 (1995), pp. 706–740.

[EKP11] J. L. Eftang, D. J. Knezevic, and A. T. Patera. "An hp certified reduced basis method for parametrized parabolic partial differential equations". In: *Mathematical and Computer Modelling of Dynamical Systems* 17 (2011), pp. 395 –422.

[EL13] H. C. Elman and Q. Liao. “Reduced Basis Collocation Methods for Partial Differential Equations with Random Coefficients”. In: *SIAM/ASA J. Uncertain. Quantification* 1 (2013), pp. 192–217.

[EL18] L. Einkemmer and C. Lubich. “A Low-Rank Projector-Splitting Integrator for the Vlasov–Poisson Equation”. In: *SIAM J. Sci. Comput.* 40.5 (2018), B1330–B1360. DOI: 10.1137/18M116383X.

[EOP20] L. Einkemmer, A. Ostermann, and C. Piazzola. “A low-rank projector-splitting integrator for the Vlasov-Maxwell equations with divergence correction”. In: *J. Comput. Phys.* 403 (2020).

[Eve09] G. Evensen. *Data Assimilation: The ensemble Kalman filter.* Springer-Verlag Berlin Heidelberg, 2009. DOI: 10.1007/978-3-642-03711-5.

[FHN19] A. Falcó, W. Hackbusch, and A. Nouy. “On the Dirac–Frenkel Variational Principle on Tensor Banach Spaces”. In: *Found. Comput. Math.* 19.1 (2019), pp. 159–204. DOI: 10.1007/s10208-018-9381-4.

[Fis96] G. Fishman. *Monte Carlo: Concepts, Algorithms, and Applications.* Springer-Verlag, New York, 1996.

[FL18] F. Feppon and P. F. J. Lermusiaux. “A Geometric Approach to Dynamical Model Order Reduction”. In: *SIAM J. Matrix Anal. Appl.* 39.1 (2018), pp. 510–538. DOI: 10.1137/16M1095202.

[FN17] M. B. Friess and A. Nouy. “Dynamical Model Reduction Method for Solving Parameter-Dependent Dynamical Systems”. In: *SIAM J. Sci. Comput.* 39 (2017).

[FS21] M. Feischl and A. Scaglioni. “Convergence of adaptive stochastic collocation with finite elements”. In: *Computers and Mathematics with Applications* 98 (2021), pp. 139–156. DOI: 10.1016/j.camwa.2021.07.001.

[GDS03] R. Ghanem and P D Spanos. “Stochastic Finite Element: a Spectral Approach”. In: vol. 224. Springer, New York, Jan. 2003. DOI: 10.1007/978-1-4612-3094-6.

[Ger+10] M. Gerritsma et al. “Time-dependent generalized polynomial chaos”. In: *Journal of Computational Physics* 229.22 (2010), pp. 8333–8363. DOI: 10.1016/j.jcp.2010.07.020.

[GN18] D. Guignard and F. Nobile. “A Posteriori Error Estimation for the Stochastic Collocation Finite Element Method”. In: *SIAM Journal on Numerical Analysis* 56.5 (2018), pp. 3121–3143. DOI: 10.1137/17M1155454.

[GP05] M. A. Grepl and A. T. Patera. “A posteriori error bounds for reduced-basis approximations of parametrized parabolic partial differential equations”. In: *Mathematical Modelling and Numerical Analysis* 39 (2005), pp. 157–181.

[Gra+11] I. Graham et al. “Quasi-Monte Carlo methods for elliptic PDEs with random coefficients and applications”. In: *J. Comput. Phys.* 230.10 (2011), pp. 3668–3694. DOI: 10.1016/j.jcp.2011.01.023.

[GS91] R. G. Ghanem and P. Spanos. *Stochastic Finite Elements: a Spectral Approach.* Springer–Verlag, New York, 1991.

[Gui18] D. Guignard. “Partial Differential Equations with Random Input Data: A Perturbation Approach”. In: *Archives of Computational Methods in Engineering* (Sept. 2018). DOI: 10.1007/s11831-018-9275-2.

[GVL96] G. H. Golub and C. F. Van Loan. *Matrix Comput.* 3rd ed. Baltimore, MD, USA: Johns Hopkins University Press, 1996.

[Haa13] B. Haasdonk. “Convergence Rates of the POD–Greedy Method”. In: *Mathematical Modelling and Numerical Analysis* 47 (2013), pp. 859–873.

[HLW04] E. Hairer, C. Lubich, and G. Wanner. “Geometric Numerical Integration: Structure Preserving Algorithms for Ordinary Differential Equations”. In: 2004.

[HWL06] E. Hairer, G. Wanner, and C. Lubich. “Numerical Integrators”. In: *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations.* Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 27–50. DOI: 10.1007/3-540-30666-8_2.

[Jaz70] A. Jazwinski. *Stochastic Processes and Filtering Theory.* Vol. 63. Academic Pr, 1970. DOI: 10.1017/CBO9781107706804.

[JOP+01] E. Jones, T. Oliphant, P. Peterson, et al. *SciPy: Open source scientific tools for Python.* 2001.

[Kal02] E. Kalnay. *Atmospheric Modeling, Data Assimilation and Predictability.* Cambridge University Press, 2002. DOI: 10.1017/CBO9780511802270.

[Kal60] R. Kalman. “A New Approach to Linear Filtering and Prediction Problems”. In: *Journal of Basic Engineering* 82.1 (Mar. 1960), pp. 35–45. DOI: 10.1115/1.3662552.

[Kha+20] A. Khan et al. “Robust a posteriori error estimation for stochastic Galerkin formulations of parameter-dependent linear elasticity equations”. In: *Mathematics of Computation* 90 (2020). DOI: 10.1090/mcom/3572.

[KKS06] O. Koch, W. Kreuzer, and A. Scrinzi. “Approximation of the Time-dependent Electronic SchröDinger Equation by MCTDHF”. In: *Appl. Math. Comput.* 173.2 (Feb. 2006), pp. 960–976. DOI: 10.1016/j.amc.2005.04.027.

[KL07a] O. Koch and C. Lubich. “Dynamical Low Rank Approximation”. In: *SIAM J. Matrix Anal. Appl.* 29.2 (2007), pp. 434–454. DOI: 10.1137/050639703.

[KL07b] O. Koch and C. Lubich. “Regularity of the multi-configuration time-dependent Hartree approximation in quantum molecular dynamics”. In: *Mathematical Modelling and Numerical Analysis* 41 (2007), pp. 315–331.

[KL10] O. Koch and C. Lubich. “Dynamical Tensor Approximation”. In: *SIAM J. Matrix Anal. Appl.* 31 (2010), pp. 2360–2375.

[KLW16] E. Kieri, C. Lubich, and H. Walach. “Discretized Dynamical Low Rank Approximation in the Presence of Small Singular Values”. In: *SIAM J. on Numer. Anal.* 54.2 (2016), pp. 1020–1038. DOI: 10.1137/15M1026791.

[KN21] Y. Kazashi and F. Nobile. “Existence of dynamical low rank approximations for random semi-linear evolutionary equations on the maximal interval”. In: *Stochastic Partial Differential Equations* 9 (2021), pp. 603 –629.

[KNV21] Y. Kazashi, F. Nobile, and E. Vidličková. “Stability properties of a projector-splitting scheme for dynamical low rank approximation of random parabolic equations”. In: *Numerische Mathematik* (2021). DOI: 10.1007/s00211-021-01241-4.

[KPB18] A. Khan, C. E. Powell, and A. Bespalov. “Robust error estimation for lowest-order approximation of nearly incompressible elasticity”. In: (2018). arXiv:1801.04122.

[KV18] E. Kieri and B. Vandereycken. “Projection Methods for Dynamical Low-Rank Approximation of High-Dimensional Problems”. In: *Comput. Methods Appl. Math.* 19.1 (2018), pp. 73–92. DOI: 10.1515/cmam-2018-0029.

[LE96a] P. J. van Leeuwen and G. Evensen. “Data Assimilation and Inverse Methods in Terms of a Probabilistic Formulation”. In: *Monthly Weather Review* 124.12 (1996), pp. 2898 –2913. DOI: 10.1175/1520-0493.

[LE96b] P. van Leeuwen and G. Evensen. “Data Assimilation and Inverse Methods in Terms of a Probabilistic Formulation”. In: *Monthly Weather Review* 124.12 (Dec. 1996), pp. 2898–2913. DOI: 10.1175/1520-0493(1996)124<2898:DAAIMI>2.0.CO;2.

[Lee+19] P. van Leeuwen et al. “Particle filters for high-dimensional geoscience applications: A review”. In: *Quarterly Journal of the Royal Meteorological Society* 145.723 (2019), pp. 2335–2365. DOI: 10.1002/qj.3551.

[Leo17] G. Leoni. *A first course in Sobolev spaces (2nd Ed.)* American Mathematical Society, Providence, Rhode Island, 2017.

[LGMT09] F. Le Gland, V. Monbet, and V.-D. Tran. *Large sample asymptotics for the ensemble Kalman filter.* Research Report RR-7014. INRIA, 2009, p. 25.

[LMK10] O. Le Maître and O. Knio. *Spectral Methods for Uncertainty Quantification: With Applications to Computational Fluid Dynamics.* Springer Netherlands, Jan. 2010, p. 536. DOI: 10.1007/978-90-481-3520-2.

[LO14] C. Lubich and I. V. Oseledets. “A projector-splitting integrator for dynamical low-rank approximation”. In: *BIT Numer. Math.* 54.1 (2014), pp. 171–188. DOI: 10.1007/s10543-013-0454-0.

[Lor96] E. Lorenz. "Predictability: a problem partly solved". In: *Seminar on Predictability.* Vol. 1. ECMWF. Shinfield Park, Reading: ECMWF, 1996, pp. 1–18.

[LOV15a] C. Lubich, I. Oseledets, and B. Vandereycken. "Time Integration of Tensor Trains". In: *SIAM J. Numer. Anal.* 53 (2015), pp. 917–941.

[LOV15b] C. Lubich, I. V. Oseledets, and B. Vandereycken. "Time Integration of Tensor Trains". In: *SIAM J. Numer. Anal.* 53.2 (2015), pp. 917–941. DOI: 10.1137/140976546.

[LPP09] A. Lozinski, M. Picasso, and V. Prachittham. "An Anisotropic Error Estimator for the Crank–Nicolson Method: Application to a Parabolic Problem". In: *SIAM Journal on Scientific Computing* 31.4 (2009), pp. 2757–2783. DOI: 10.1137/080715135.

[LPS14] G. J. Lord, C. E. Powell, and T. Shardlow. *An Introduction to Computational Stochastic PDEs.* Cambridge Texts in Applied Mathematics. Cambridge University Press, 2014. DOI: 10.1017/CBO9781139017329.

[LSZ15] K. Law, A. Stuart, and K. Zygalakis. *Data Assimilation: A Mathematical Introduction.* Springer International Publishing Switzerland, 2015. DOI: 10.1007/978-3-319-20325-6.

[Lub+13] C. Lubich et al. "Dynamical Approximation by Hierarchical Tucker and Tensor-Train Tensors". In: *SIAM J. Matrix Anal. Appl.* 34 (2013), pp. 470–494.

[LX08] J. Li and D. Xiu. "On numerical properties of the ensemble Kalman filter for data assimilation". In: *Computer Methods in Applied Mechanics and Engineering* 197.43 (2008). Stochastic Modeling of Multiscale and Multiphysics Problems, pp. 3574 –3583. DOI: 10.1016/j.cma.2008.03.022.

[Men+18] H. Mena et al. "Numerical low-rank approximation of matrix differential equations". In: *J. Comput. Appl. Math.* 340 (2018), pp. 602–614.

[MH12] A. Majda and J. Harlim. *Filtering Complex Turbulent Systems.* Cambridge University Press, 2012. DOI: 10.1017/CBO9781139061308.

[MK05] H. G. Matthies and A. Keese. "Galerkin methods for linear and nonlinear elliptic stochastic partial differential equations". In: *Computer Methods in Applied Mechanics and Engineering* 194.12 (2005). Special Issue on Computational Methods in Stochastic Mechanics and Reliability Analysis, pp. 1295–1331. DOI: 10.1016/j.cma.2004.05.027.

[MK10] O. P. L. Maître and O. M. Knio. "Spectral Methods for Uncertainty Quantification". In: 2010.

[MMC90] H.-D. Meyer, U. Manthe, and L. S. Cederbaum. "The multi-configurational time-dependent Hartree approach". In: *Chemical Physics Letters* 165 (1990), pp. 73–78.

[MN18] E. Musharbash and F. Nobile. “Dual Dynamically Orthogonal approximation of incompressible Navier Stokes equations with random boundary conditions”. In: *J. Comput. Phys.* 354 (2018), pp. 135–162.

[MNV17] E. Musharbash, F. Nobile, and E. Vidlicková. “Symplectic dynamical low rank approximation of wave equations with random parameters”. In: *BIT Numerical Mathematics* (2017), pp. 1–49.

[MNZ15] E. Musharbash, F. Nobile, and T. Zhou. “Error Analysis of the Dynamically Orthogonal Approximation of Time Dependent Random PDEs”. In: *SIAM J. Sci. Comput.* 37.2 (2015), A776–A810. DOI: 10.1137/140967787.

[MQS14] A. J. Majda, D. Qi, and T. P. Sapsis. “Blended particle filters for large-dimensional chaotic dynamical systems”. In: *Proceedings of the National Academy of Sciences* 111.21 (2014), pp. 7511–7516. DOI: 10.1073/pnas.1405675111.

[MS13] A. Majda and T. Sapsis. “Blending Modified Gaussian Closure and Non-Gaussian Reduced Subspace Methods for Turbulent Dynamical Systems”. In: *Journal of Nonlinear Science* 23 (2013), pp. 1039–1071. DOI: 10.1007/s00332-013-9178-1.

[MX06] A. Majda and W. Xiaoming. *Nonlinear Dynamics and Statistical Theories for Basic Geophysical Flows*. Cambridge University Press, 2006.

[Nee08] J. van Neerven. *Stochastic Evolution Equations - Lecture Notes of the 11th Internet Seminar, 2007/08*. Jan. 2008.

[Nie92] H. Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*. Society for Industrial and Applied Mathematics, 1992. DOI: 10.1137/1.9781611970081.

[NRP09] N. C. Nguyen, G. Rozza, and A. T. Patera. “Reduced basis approximation and a posteriori error estimation for the time-dependent viscous Burgers’ equation”. In: *Calcolo* 46 (2009), pp. 157–185.

[NTT16] F. Nobile, L. Tamellini, and R. Tempone. “Convergence of quasi-optimal sparse-grid approximation of Hilbert-space-valued functions: application to random elliptic PDEs”. In: *Numerische Mathematik* 134.2 (2016), pp. 343–388. DOI: 10.1007/s00211-015-0773-y.

[NTW08a] F. Nobile, R. Tempone, and C. Webster. “A Sparse Grid Stochastic Collocation Method for Partial Differential Equations with Random Input Data”. In: *SIAM Journal on Numerical Analysis* 46.5 (2008), pp. 2309–2345.

[NTW08b] F. Nobile, R. Tempone, and C. Webster. “An Anisotropic Sparse Grid Stochastic Collocation Method for Partial Differential Equations with Random Input Data”. In: *SIAM Journal on Numerical Analysis* 46.5 (2008), pp. 2411–2442.

[NV19] F. Nobile and E. Vidličková. “MATHICSE Technical Report: A posteriori error estimation for the stochastic collocation finite element approximation of the heat equation with random coefficients”. In: (2019). MATHICSE Technical Report, 10.5075/epfl-MATHICSE-265791.

[OPW19] A. Ostermann, C. Piazzola, and H. Walach. “Convergence of a Low-Rank Lie-Trotter Splitting for Stiff Matrix Differential Equations”. In: *SIAM J. Numer. Anal.* 57 (2019), pp. 1947–1966.

[Pic98] M. Picasso. “Adaptive finite elements for a linear parabolic problem”. In: *Computer Methods in Applied Mechanics and Engineering* 167.3-4 (1998), pp. 223–237.

[QJM15] D. Qi and A. J. Majda. “Blended particle methods with adaptive subspaces for filtering turbulent dynamical systems”. In: *Physica D: Nonlinear Phenomena* 298-299 (Feb. 2015). DOI: 10.1016/j.physd.2015.02.002.

[Qua09] A. Quarteroni. “Numerical Models for Differential Problems”. In: 2009.

[QV08] A. Quarteroni and A. Valli. “Numerical Approximation of Partial Differential Equations”. In: 2008.

[RC15] S. Reich and C. Cotter. *Probabilistic Forecasting and Bayesian Data Assimilation.* Cambridge University Press, 2015. DOI: 10.1017/CBO9781107706804.

[RH15] P. Rebeschini and R. van Handel. “Can local particle filters beat the curse of dimensionality?” In: *Ann. Appl. Probab.* 25.5 (Oct. 2015), pp. 2809–2866. DOI: 10.1214/14-AAP1061.

[Sal98] R. Salmon. *Lectures on Geophysical Fluid Dynamics.* Oxford University Press, 1998.

[SG04] C. Soize and R. Ghanem. “Physical Systems with Random Uncertainties: Chaos Representations with Arbitrary Probability Measure”. In: *SIAM Journal on Scientific Computing* 26.2 (2004), pp. 395–410. DOI: 10.1137/S1064827503424505.

[SL09] T. P. Sapsis and P. F. J. Lermusiaux. “Dynamically orthogonal field equations for continuous stochastic dynamical systems”. In: *Phys. D: Nonlinear Phenom.* 238.23 (2009), pp. 2347–2360. DOI: 10.1016/j.physd.2009.09.017.

[SL12] T. P. Sapsis and P. F. J. Lermusiaux. “Dynamical criteria for the evolution of the stochastic dimensionality in flows with uncertainty”. In: *Phys. D: Nonlinear Phenom.* 241.1 (2012), pp. 60–76. DOI: 10.1016/j.physd.2011.10.001.

[SL13a] T. Sondergaard and P. Lermusiaux. “Data Assimilation with Gaussian Mixture Models Using the Dynamically Orthogonal Field Equations. Part I: Theory and Scheme”. In: *Monthly Weather Review* 141.6 (June 2013), pp. 1737–1760. DOI: 10.1175/MWR-D-11-00295.1.

[SL13b] T. Sondergaard and P. Lermusiaux. “Data Assimilation with Gaussian Mixture Models Using the Dynamically Orthogonal Field Equations. Part II: Applications”. In: *Monthly Weather Review* 141.6 (June 2013), pp. 1761–1785. DOI: 10.1175/MWR-D-11-00296.1.

[SM13] T. P. Sapsis and A. J. Majda. “Blended reduced subspace algorithms for uncertainty quantification of quadratic systems with a stable mean state”. In: *Physica D: Nonlinear Phenomena* 258 (2013), pp. 61 –76. DOI: 10.1016/j.physd.2013.05.004.

[Smo63] S. Smolyak. “Quadrature and interpolation formulas for tensor products of certain classes of functions”. In: *Dokl. Akad. Nauk SSSR* 148.5 (1963). cited By 108, pp. 1042–1045.

[Sny12] C. Snyder. “Particle filters, the optimal proposal and high-dimensional systems”. In: *Seminar on Data assimilation for atmosphere and ocean, 6-9 September 2011.* ECMWF. Shinfield Park, Reading: ECMWF, 2012, pp. 161–170.

[SSN15] R. Stefanescu, A. Sandu, and I. Navon. “POD/DEIM reduced-order strategies for efficient four dimensional variational data assimilation”. In: *Journal of Computational Physics* 295 (2015), pp. 569–595. DOI: 10.1016/j.jcp.2015.04.030.

[ST06] C. Schwab and R. A. Todor. “Karhunen Loève approximation of random fields by generalized fast multipole methods”. In: *Journal of Computational Physics* 217.1 (Sept. 2006), pp. 100–122. DOI: 10.1016/j.jcp.2006.01.048.

[ULS13] M. P. Ueckermann, P. F. Lermusiaux, and T. P. Sapsis. “Numerical schemes for dynamically orthogonal equations of stochastic fluid and ocean flows”. In: *J. Comput. Phys.* 233 (2013), pp. 272–294. DOI: 10.1016/j.jcp.2012.08.041.

[Ver03] R. Verfürth. “A posteriori error estimates for finite element discretizations of the heat equation”. In: *CALCOLO* 40.3 (2003), pp. 195–212. DOI: 10.1007/s10092-003-0073-2.

[Ver13] R. Verfürth. *A Posteriori Error Estimation Techniques for Finite Element Methods.* A Posteriori Error Estimation Techniques for Finite Element Methods. OUP Oxford, 2013.

[Ver99a] R. Verfürth. “Error estimates for some quasi-interpolation operators”. In: *ESAIM: Mathematical Modelling and Numerical Analysis* 33.4 (1999), 695–713. DOI: 10.1051/m2an:1999158.

[Ver99b] R. Verfürth. “Error estimates for some quasi-interpolation operators”. In: *ESAIM: Mathematical Modelling and Numerical Analysis* 33.4 (1999), 695–713. DOI: 10.1051/m2an:1999158.

[Wal18] H. Walach. *Time integration for the dynamical low-rank approximation of matrices and tensors.* Tübingen: Eberhard Karls Universität Tübingen, 2018.

[WH02] J. S. Whitaker and T. M. Hamill. “Ensemble Data Assimilation without Perturbed Observations”. In: *Monthly Weather Review* 130.7 (2002), pp. 1913 –1924.

[Wie38] N. Wiener. “The Homogeneous Chaos”. In: *Am. J. Math.* 60.4 (1938), pp. 897–936.

[WK06] X. Wan and G. E. Karniadakis. “Long-Term Behavior of Polynomial Chaos in Stochastic Flow Simulations”. In: *Computer Methods in Applied Mechanics and Engineering* 195 (2006), pp. 5582–5596.

[Wlo87] J. Wloka. *Partial Differential Equations.* Cambridge University Press, 1987.

[WP02] K. E. Willcox and J. Peraire. “Balanced Model Reduction via the Proper Orthogonal Decomposition”. In: *AIAA Journal* 40 (2002), pp. 2323–2330.

[WS07] X. Wang and I. Sloan. “Brownian bridge and principal component analysis: towards removing the curse of dimensionality”. In: *IMA J. Numer. Anal.* 27.4 (2007), pp. 631–654. DOI: 10.1093/imanum/drl044.

[XH05a] D. Xiu and J. S. Hesthaven. “High-order collocation methods for differential equations with random inputs”. In: *SIAM Journal on Scientific Computing* 27.3 (2005), pp. 1118–1139.

[XH05b] D. Xiu and J. S. Hesthaven. “High-Order Collocation Methods for Differential Equations with Random Inputs”. In: *SIAM Journal on Scientific Computing* 27.3 (2005), pp. 1118–1139. DOI: 10.1137/040615201.

[XK02] D. Xiu and G. E. Karniadakis. “The Wiener–Askey Polynomial Chaos for Stochastic Differential Equations”. In: *SIAM J. Sci. Comput.* 24.2 (2002), pp. 619–644. DOI: 10.1137/S1064827501387826.

[YY18] T. Yamada and K. Yamamoto. “A second-order discretization with Malliavin weight and Quasi-Monte Carlo method for option pricing”. In: *Quantitative Finance* (2018), pp. 1–13. DOI: 10.1080/14697688.2018.1430371.

[Zan+03] J. Zanghellini et al. “An MCTDHF approach to multielectron dynamics in laser fields”. In: *Laser Physics* 13.8 (2003), pp. 1064–1068.

[Zei90] E. Zeidler. *Linear monotone operators: Hilbert Space Methods and Linear Parabolic Differential Equations.* Springer-Verlag, New York, 1990. DOI: 10.1007/978-1-4612-0985-0.

Eva Vidličková

Avenue de Cour 15, Lausanne 1007, Switzerland
eva.vidlickova@epfl.ch

Education

École Polytechnic Fédérale de Lausanne *Lausanne, 08/2017 - now*
Department of Applied and Computational Mathematics

- Phd candidate, Scientific Computing and Uncertainty Quantification group
- main topic: Dynamical low rank approximation for uncertainty quantification of random time-dependent problems

Charles University *Prague, 10/2011 - 06/2017*
Faculty of Mathematics and Physics

- bachelor and master studies, major in Numerical and computational mathematics, graduated with awards
- master thesis: Fourier-Galerkin Method for Stochastic Homogenization of Elliptic Partial Differential Equations, awarded the Prof. Babuska prize

Heidelberg University *Heidelberg, 09/2014 - 02/2015*
Faculty of Mathematics and Computer Science

- Erasmus programme, focus on numerical solutions of PDEs

Further research experience

Manchester University *Manchester, 01/2019*

- scientific visit of Prof. Kody Law, Department of Mathematics

Czech Technical University *Prague, 05/2016 - 06/2017*

- research assistant in the Faculty of Civil Engineering, Department of Mechanics

Slovak Academy of Sciences *Bratislava, 08/2016 - 09/2016*

- summer internship at the Mathematical Institute, Department of Computer Science
- topics: modifications of Jacobi algorithm for computation of SVD decomposition

Heidelberg Institute for Theoretical Studies *Heidelberg, 07/2015 - 09/2015*

- summer internship in the Data Mining and Uncertainty Quantification group
- topics: uncertainty quantification, sparse grid collocation method, inverse problems

Conference or seminar presentations

- **Poster presentation:** *Dynamical low rank approximation and sparse grid adaptivity*, MoRePaS 2018 - Model reduction of parametrized systems IV, Nantes, France, 10.-13.4.2018
- **Poster presentation:** *Dynamical low rank approximation and sparse grid adaptivity*, Swiss Numerics Day 2018, ETH Zurich, Switzerland, 20.4.2018
- **Seminar talk:** *A posteriori error estimation for the stochastic collocation finite element approximation of a random heat equation,* MATHICSE retreat 2019, Switzerland
- **Conference talk:** *A posteriori error estimation for the stochastic collocation finite element approximation of a random heat equation,* MAFELAP 2019, Brunel university, London, UK
- **Conference talk:** *Time discretization and stability estimates for DLR approximation of a heat equation,* ICIAM 2019, Valencia, Spain
- **Conference talk:** *Time discretization and stability estimates for DLR approximation of a heat equation,* Modeling 2019, Olomouc, Czech Republic
- **Online minisymposium talk:** *Stability properties of a projector-splitting scheme for the dynamical low rank approximation of random parabolic equations,* 21.7.2020, Online Minisymposium on Low-rank Geometry and Computation 2020, Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany; Virtual event
- **Summer school talk:** *Dynamical low rank approximation and discretization schemes,* 10.9.2020, Model Order Reduction Summer School 2020, École polytechnique fédérale, Lausanne, Swizterland; Virtual event
- **Conference talk:** *Dynamical low rank approximation for data assimilation,* 28.-30.6.2021, UNCECOMP 2021, Athens, Greece; Virtual event
- **Conference talk:** *Dynamical low rank approximation for data assimilation,* 14.-18.3.2022, SIAM PD 2022, Berlin, Germany; Virtual event
- **Conference talk:** *Stability properties of a projector-splitting scheme for the dynamical low rank approximation of random parabolic equations,* 12.-15.4.2022, SIAM UQ 2022, Atlanta, USA; Hybrid event

Publications

- Y. Kazashi, F. Nobile, E. Vidličková, *Stability properties of a projector-splitting scheme for dynamical low rank approximation of random parabolic equations*, Numerische Mathematik, 2021
- F. Nobile, E. Vidličková, *A posteriori error estimation for the stochastic collocation finite element approximation of the heat equation with random coefficients*, Sparse Grids and Applications - Munich 2018 (to appear), 2022

Other skills and experiences

- **Summer school co-organization:** *Model Order Reduction Summer School 2020,* 7.-10.9.2020, École polytechnique fédérale, Lausanne, Switzerland; Virtual event
- **Languages:** C1 in English, B1 in German, B1 in French, native speaker in Slovak and Czech