ESAIM: COCV 28 (2022) 78
https://doi.org/10.1051/cocv/2022071

ESAIM: Control, Optimisation and Calculus of Variations
www.esaim-cocv.org

KERNEL REPRESENTATION OF KALMAN OBSERVER AND ASSOCIATED $\mathcal{H}$-MATRIX BASED DISCRETIZATION

Matthieu Aussal[1] and Philippe Moireau[2,*]

Abstract. In deterministic estimation, applying a Kalman filter to a dynamical model based on partial differential equations is theoretically seducing but solving the associated Riccati equation leads to a so-called curse of dimensionality for its numerical implementation. In this work, we propose to entirely revisit the theory of Kalman filters for parabolic problems where additional regularity results proves that the Riccati equation solution belongs to the class of Hilbert-Schmidt operators. The regularity of the associated kernel then allows to proceed to the numerical analysis of the Kalman full space-time discretization in adapted norms, hence justifying the implementation of the related Kalman filter numerical algorithm with $\mathcal{H}$-matrices typically developed for integral equations discretization.

Mathematics Subject Classification. 47D06, 49M41, 93B53.

Received December 17 , 2021. Accepted October 30, 2022.

1. A deterministic optimal control framework for PDE-model estimation

1.1. Introduction

Kalman filters have been very popular since their introduction in the 60s [31, 32] for estimating dynamical systems from available measurements. They were originally designed and justified for finite-dimensional linear systems of ordinary differential equations (ODE) with linearly generated observations. Moreover, the initial formalism encompasses stochastic formulations associated with Gaussian independent disturbances where the resulting estimator is proved to be the Best Linear Unbiased Estimator but also the Minimum Mean Square Error estimator or the Least Squares Estimator [8]. Surprisingly, the exact same formulation of the Kalman filter can be derived in a completely deterministic context [7], offering a theoretical alternative to adjoint based minimization of least squares functional [12, 18, 40] known in the data assimilation community as the 4D-Var approach [36, 37]. This time the Kalman estimator is defined as the observer equivalent to minimizing a moving-window least squares functional integrating a data fitting term, under the constraints of the dynamics. As a result, the initial dynamics is corrected by a feedback incorporating the available measurements, feedback computed from an operator solution of a Riccati dynamics. When considering PDEs evolution problems, new challenges appear to define adequate notions of solution of the initial problem, of the Riccati dynamics, and of

Keywords and phrases: Infinte dimensional systems, Riccati equation, data assimilation.

[1] Ecole Polytechnique (CMAP), Inria (Team Idefix), Institut Polytechnique, Paris Palaiseau 91128, France.
[2] Inria – LMS,Ecole Polytechnique, CNRS, Institut Polytechnique de Paris Team MΞDISIM, Inria Saclay - Ile-de-France, Turing Building, Palaiseau 91128, France.

[*] Corresponding author: philippe.moireau@inria.fr

the estimator – also called observer – dynamics, even in the deterministic context, and with additional subtlety in the stochastic context, see [7] or the survey [19]. Moreover, refined question of regularity arises, in particular concerning the Riccati operator. These questions have been widely studied in the literature, see [9, 21, 34] and references therein, with new results in the class of Hilbert-Schmidt [14, 15, 20] obtained recently for certain classes of parabolic equations. Building on these new results, we here proposed a complete vision of Kalman theory for PDEs, from its deterministic definition and its essential properties up to its complete full space-time numerical analysis in the space of Hilbert-Schmidt operators, completing in the sense the initial works of [14, 26]. Moreover here, in circumstances of adequate regularity, we aim at proving that the continuous Riccati operator is associated with a regular kernel which can be discretized using Hierarchical matrix algebra [10, 30], as well known when discretizing integral equations [2] but also experimented with Algebraic Riccati Equations [27]. Here, we mathematically justify in the context of the Kalman filter previous numerical experiments [38], while proposing numerical improvements in the $\mathcal{H}$-matrix treatment of the Riccati operators. This allows to overcome – here for parabolic problems – the classical curse of dimensionality that faces the Kalman filter for PDEs, sometimes limiting its use with respect to the least squares minimization using adjoint based methods [43]. In fact, with our approach, we can consider discretization with millions of degrees of freedom and approximate the corresponding Riccati operators, whereas classical research directions are more inclined to follow reduced basis methods model approximation [5, 17, 42, 44] or reduced covariance strategies [16, 35, 46] to be computationally effective.

1.2. Model setting

1.2.1. A general parabolic initial value problem

Let $\mathcal{Z}$ and $\mathcal{V}$ two separable Hilbert spaces with $\mathcal{V}$ dense in $\mathcal{Z}$. Denoting $\mathcal{Z}'$ and $\mathcal{V}'$ their respective topological dual spaces, we assume a compact injection $i : \mathcal{V} \to \mathcal{Z}$ such that we identify

$$\mathcal{V} \subsetneq \mathcal{Z} \equiv \mathcal{Z}' \subsetneq \mathcal{V}',$$

and we denote $\|\cdot\|_{\mathcal{Z}}$ and $\|\cdot\|_{\mathcal{V}}$ with associated scalar product $(\cdot,\cdot)_{\mathcal{Z}}$ and $(\cdot,\cdot)_{\mathcal{V}}$. By contrast, the duality pairing on $\mathcal{V}' \times \mathcal{V}$ is denoted by $\langle \cdot,\cdot \rangle_{\mathcal{V}}$.

We consider an operator $A \in \mathcal{L}(\mathcal{V}', \mathcal{V})$ associated with a continuous and coercive bilinear form a in $\mathcal{V} \times \mathcal{V} \to \mathbb{R}$

$$\forall (v,w) \in \mathcal{V}, \quad \langle Av, w \rangle_{\mathcal{V}} = a(v,w),$$

with

$$\exists c_{\mathrm{st}}, \mid \forall v \in \mathcal{V}, \ a(v,v) \geq c_{\mathrm{st}} \|v\|_{\mathcal{V}}^2.$$

With a slight abuse of notation, we also consider A as an unbounded operator from $\mathcal{D}(A)$ to $\mathcal{Z}$ with

$$\mathcal{D}(A) = \big\{ v \in \mathcal{V} \, : \, \exists \beta \in \mathcal{Z} \text{ s.t. } \forall w \in \mathcal{V}, a(v,w) = (\beta, w)_{\mathcal{Z}} \big\}.$$

We denote A^* of domain $\mathcal{D}(A^*)$, the adjoint operator. The fractional power A^ρ, $0 \leq \rho \leq 1$, are defined following [33] – see also Section II-1.4 of [9] and we further restrict our study to the classical case $\mathcal{D}(A^{\frac{1}{2}}) = \mathcal{D}(A^{\frac{1}{2}*}) = \mathcal{V}$, studied and illustrated in [41]. Given $T > 0$, we consider a dynamical system in $\mathcal{Z}$ represented by the following dynamics

$$\begin{cases} \dot{z} + Az = B\nu, & \text{in } (0,T) \\ z(0) = z_0, \end{cases} \tag{1.1}$$

where we use the short notation $\dot{z} = \frac{\mathrm{d}}{\mathrm{d}t} z$. The quantity $(0, T) \ni t \mapsto \nu(t) \in \mathcal{U}$ is typically an unknown contribution to the dynamics – namely a so-called model error – and $\mathcal{U}$ is also assumed to be a Hilbert space with associated norm $\|\cdot\|_{\mathcal{U}}$. For the sake of simplicity, we consider a model-error operator $B \in \mathcal{L}(\mathcal{U}, \mathcal{Z})$.

The operator $(A, \mathcal{D}(A))$ is obviously maximal accretive, hence – by Lummer-Philips theorem [45] – is the generator of a C^0-semigroup of contraction $\varPhi$ such that

$$\forall z \in \mathcal{D}(A), \quad Az = -\frac{\mathrm{d}}{\mathrm{d}t}(\varPhi(t)z)_{|t=0^+}.$$

Moreover, $\varPhi$ is analytical Theorem 3.6.1 of [49] – see Definition 2.3 and Theorem 2.11 of [9].

For the sake of completeness, we recall the various notions of solution of the non-homogeneous linear evolution equations (1.1) that will be used in this work – see for instance ([9], Part II).

Definition 1.1 (Notions of solution)**.** Given $[0, T] \ni t \mapsto \beta(t)$ and z_0 regular enough, we list 4 different types of solution of

$$\begin{cases} \dot{z} + Az = \beta(t), & \text{in } (0, T) \\ z(0) = z_0. \end{cases} \tag{1.2}$$

(i) z is a *strict* solution of Problem (1.2) in $\mathrm{L}^2((0, T); \mathcal{Z})$ if z belongs to $\mathrm{H}^1((0, T); \mathcal{Z}) \cap \mathrm{L}^2((0, T); \mathcal{D}(A))$ and (1.2) is satisfied in the strong sense.

(ii) z is a *mild* solution of Problem (1.2) in $\mathrm{L}^2((0, T); \mathcal{Z})$ if $z \in \mathrm{C}^0([0, T]; \mathcal{Z})$ is given by the Duhamel formula

$$z(t) = \varPhi(t)z_0 + \int_0^t \varPhi(t-s)\beta(t) \ \mathrm{d}s. \tag{1.3}$$

(iii) z is a *weak* solution of Problem (1.2) if (1.) $z \in \mathrm{L}^2((0, T); \mathcal{Z})$, (2.) for all $q \in \mathcal{D}(A^*)$, $(q, z(\cdot))_{\mathcal{Z}}$ belongs to $\mathrm{H}^1(0, T)$ and (3.) for almost all $t \in (0, T)$,

$$\forall q \in \mathcal{D}(A^*), \quad \frac{\mathrm{d}}{\mathrm{d}t}(q, z(t))_{\mathcal{Z}} + (A^* q, z(t))_{\mathcal{Z}} = (q, \beta(t))_{\mathcal{Z}}. \tag{1.4}$$

(iv) z is a *variational* solution of Problem (1.2) if (1.) $z \in \mathrm{L}^2((0, T); \mathcal{V})$ and (2.) $\frac{\mathrm{d}z}{\mathrm{d}t} \in \mathrm{L}^2((0, T); \mathcal{V}')$ and (3.) for almost all $t \in (0, T)$,

$$\forall w \in \mathcal{V}, \quad \left\langle \frac{\mathrm{d}z}{\mathrm{d}t} + Az - \beta, w \right\rangle_{\mathcal{V}} = 0.$$

We now recall the classical existence results.

Theorem 1.2 (Solution existence)**.** *We list different cases of solution of Problem* (1.1) *or* (1.2)*:*

(i) Given $z_0 \in \mathcal{D}(A)$ and $\beta \in \mathrm{L}^2((0, T); \mathcal{D}(A))$ (or $\beta \in \mathrm{H}^1((0, T); \mathcal{Z})$ resp.), Problem (1.2) *has a unique strict solution which belongs to $\mathrm{H}^1((0, T); \mathcal{Z}) \cap \mathrm{C}^0([0, T]; \mathcal{D}(A))$ (or to $\mathrm{C}^1([0, T]; \mathcal{Z}) \cap \mathrm{C}^0([0, T]; \mathcal{D}(A))$ resp.).*

(ii) Given $z_0 \in \mathcal{Z}$ and $\beta \in \mathrm{L}^2((0, T); \mathcal{Z})$, Problem (1.2) *has a unique weak solution which coincides with the mild solution given by the Duhamel formula and the variational solution.*

(iii) Given $z_0 \in \mathcal{Z}$ and $\beta \in \mathrm{L}^2((0, T); \mathcal{V}')$, Problem (1.2) *has a unique variational solution.*

Proof. We here aggregate several classical results. For (i), we refer for instance to II-1 Proposition 3.3 of [9], For (ii), we refer to II-1 Proposition 3.2 of [9] and for (iii), we refer to II-2 - Theorem 1.1 of [9]. □

1.3. Uncertainties and observation modeling

We are now going to introduce some model uncertainties in a deterministic framework of estimation. Let us consider that we are interested in the prediction of a natural system that follows the dynamics (1.1). We denote the target trajectory $\{\check{z}(t),\, t \in [0,T]\}$, obtained from (1.1) with unknown initial condition and model error $\check{\nu}$. More precisely concerning the initial condition, we separate the unknown part ζ from the known part $\hat{z}_0$ of $z(0)$ such that

$$\check{z}(0) = \hat{z}_0 + \check{\zeta}.$$

We typically assume to have an *a priori* on the level of uncertainty in $\check{\zeta}$ namely $\|\check{\zeta}\|_{\mathcal{Z}} = O(\alpha)$ with α known. Another choice could be to assume, as it is for ill-posed inverse problems, that the initial condition belongs to a more regular space $\mathcal{V}_s$, typically

$$\mathcal{V}_s \subsetneq \mathcal{V} \subsetneq \mathcal{Z},$$

with the injection $i_s : \mathcal{V}_s \to \mathcal{V}$ at least continuous. In this case, the estimation procedure should benefit from knowing that $\|\zeta\|_{\mathcal{V}_s} = O(\alpha)$.

The source error is typically assumed to belong to $\mathrm{L}^2((0,T);\mathcal{U})$ or more strongly to $\mathrm{L}^\infty((0,T);\mathcal{U})$ with for instance

$$\forall a.e.\, t \in [0,T],\ \|\check{\nu}(t)\|_{\mathcal{U}} = O(\kappa) \text{ or } \|\check{\nu}\|^2_{\mathrm{L}^2((0,T);\mathcal{U})} = O(\kappa^2 T).$$

To circumvent this lack of information on this system, we assume to observe the given target trajectory, hence we expect to estimate the associated initial condition and the model error from the available measurements. We model the measurement procedure, by an observation operator C, such that a given measurement $y \in \mathcal{Y}$ is modeled from the application of C on a given $z \in \mathcal{Z}$, namely

$$C : \mathcal{Z} \ni z \mapsto y \in \mathcal{Y}. \tag{1.5}$$

For the sake of simplicity, we restrict ourselves to bounded observation operators. The available noisy measurements are y^δ and they are a perturbation of the unavailable perfect measurements $\check{y} = C\check{z}$ such that, $\eta = y^\delta - \check{y}$ belongs to $\mathrm{L}^\infty((0,T);\mathcal{Y})$ or more strongly to $\mathrm{L}^\infty((0,T);\mathcal{Y})$ with for instance

$$\forall a.e.\, t \in [0,T],\ \|\eta(t)\|_{\mathcal{Y}} = O(\delta) \text{ or } \|\eta\|^2_{\mathrm{L}^2((0,T);\mathcal{Y})} = O(\delta^2 T).$$

We recall that compensating the lack of knowledge on $(\check{\zeta}, \check{\nu})$ by the known data y^δ, consists in being able to invert the operator

$$\Psi \,:\, \mathcal{Z} \times \mathrm{L}^2((0,T)) \ni (\zeta,\nu) \mapsto y : \Big[[0,T] \ni t \mapsto C\Phi(t)\zeta + \int_0^t C\Phi(t-s)\nu(s)\ \mathrm{d}s\Big] \in \mathcal{Z}_T. \tag{1.6}$$

The operator Ψ can be injective but is not surjective as A is analytical hence Φ is regularizing. Therefore, inverting Ψ is ill-posed.

1.4. The advection-diffusion example

As an illuminating example all along this article, we consider an advection-diffusion problem. We introduce a bounded domain $\Omega \subset \mathbb{R}^d$ of C^2 boundary where we will define solutions of an advection-diffusion equation with an unknown source term error. We introduce a strictly positive known continuous function $f \in \mathrm{C}^0(\Omega)$ with $\forall x \in \Omega$, $f(x) > 0$ but a potentially unknown time dependent $\nu \in \mathrm{L}^2(0,T)$

$$\begin{cases} \partial_t z(x,t) - b(x)\cdot\nabla z(x,t) - \Delta z(x,t) = f(x)\nu(t), & (x,t) \in \Omega \times (0,T), \\ z(x,t) = 0, & (x,t) \in \partial\Omega \times (0,T), \\ z(x,0) = z_0(x), & x \in \Omega, \end{cases} \tag{1.7}$$

where $b \in \mathrm{H}^1(\Omega) \cap \mathrm{L}^\infty(\Omega)$ is a given velocity field such that $\nabla \cdot b = 0$. The model defined by the dynamics (1.7) enters the framework introduced in Section 1.2.1 with $\mathcal{Z} = \mathrm{L}^2(\Omega)$, $\mathcal{V} = \mathrm{H}^1_0(\Omega)$,

$$A = \upsilon \cdot \nabla + \Delta, \quad \mathcal{D}(A) = \mathrm{H}^2(\Omega) \cap \mathrm{H}^1_0(\Omega),$$

and a model error operator given by

$$B \,:\, \mathbb{R} \ni \nu \mapsto f(x)\nu \in \mathrm{L}^2(\Omega),$$

of corresponding adjoint operator

$$B^* \,:\, \mathrm{L}^2(\Omega) \ni \psi(x) \mapsto \int_\Omega f(x)\psi(x) \in \mathbb{R}.$$

Note in particular that $\mathcal{D}(A^{\frac{1}{2}}) = \mathcal{D}(A^{\frac{1}{2}*}) = \mathcal{V}$ as justified in [41].

About the measurements, we typically consider to observe the system over a subdomain ω. Therefore, we have

$$C \,:\, L^2(\Omega) \ni \varphi \mapsto \varphi|_\omega \in L^2(\omega), \quad C^* \,:\, L^2(\omega) \ni \mu \mapsto \mathbb{1}_\omega(x)\mu(x) \in L^2(\Omega).$$

In this case, Ψ – defined by (1.6) – is injective. Indeed, let us consider $(\zeta,\nu) \in \mathcal{Z} \times \mathrm{L}^2(0,T)$ such that $y = \Psi(\zeta,\nu) \equiv 0$. We first have, using the equation (1.7) satisfied in $\omega \subset \Omega$, that

$$f(x)\nu(t) = 0, \quad (x,t) \in \omega \times (0,T).$$

As f is strictly positive, we deduce that $\nu \equiv 0$. Then, from classical observability inequalities for parabolic equation from interior measurements – see [28] and references therein – we have

$$\|z(T)\|^2_{\mathrm{L}^2(\Omega)} \leq c_{\mathrm{st}} \int_0^T \int_\omega |z(x,t)|^2 \,\mathrm{d}x\,\mathrm{d}t = 0.$$

Finally, by backward uniqueness of the solution $z(T) = 0 \Rightarrow \zeta = 0$.

1.5. Least squares estimation

1.5.1. least squares criterion

A classical estimation approach consists in formulating the estimation problem as an optimal control problem, hence minimizing a least squares criterion balancing the uncertainties. We thus introduce, for $\gamma \in \mathbb{R}^+$,

$$\mathscr{J}_T(\zeta,\nu) = \frac{1}{2} a_s(\zeta,\zeta) + \frac{1}{2}\int_0^T \left[\gamma \|y^\delta(s) - C z_{|\zeta,\nu}(s)\|^2_{\mathcal{Y}} + \kappa^2 \|\nu(s)\|^2_{\mathcal{U}}\right] \,\mathrm{d}s, \tag{1.8}$$

where we denote by $z_{|\zeta,\nu}(s)$, a trajectory of (1.1) for a corresponding initial condition $z(0) = \hat{z}_0 + \zeta$.

On the one hand, the bilinear and symmetric form a_s is a penalization terms aiming at controlling the regularity of the estimated initial condition of such ill-posed problem. Moreover a_s will be assumed to be coercive and bounded in $\mathcal{V}_s$, with typically the existence of $0 < \varepsilon < \alpha^{-1}$ such that

$$\varepsilon^2 \|\zeta\|^2_{\mathcal{V}_s} \leq a_s(\zeta,\zeta) \leq \alpha^{-2}\|\zeta\|^2_{\mathcal{V}_s}.$$

Denoting A_s the *Friedrichs* extension of the triple $(\mathcal{Z}, \mathcal{V}_s, a_s)$, we introduce

$$\mathcal{D}(A_s) = \Big\{ v \in \mathcal{V}_s \,\Big|\, \exists f \in \mathcal{Z} \text{ s .t. } a_s(v,w) = (f,w),\, w \in \mathcal{V}_s \Big\},$$

and – again with a slight abuse of notation – the operator $\mathit{\Pi}_0 = A_s^{-1}$ can be either consider as a bounded application from $\mathcal{Z}$ to $\mathcal{D}(A_s)$ or from $\mathcal{V}_s'$ to $\mathcal{V}_s$. The term

$$a_s(\zeta,\zeta) = \langle \mathit{\Pi}_0^{-1}\zeta, \zeta\rangle_{\mathcal{V}_s},$$

where $\langle\cdot,\cdot\rangle_{\mathcal{V}_s}$ stands for the duality product, is a generalized Tikhonov regularization term [23] enforcing a regularity in $\mathcal{V}_s \subset \mathcal{Z}$. The operator $\mathit{\Pi}_0$ will be called the *a priori* initial covariance operator, as we typically expect that the target trajectory satisfies

$$\langle \mathit{\Pi}_0^{-1}\check{\zeta}, \check{\zeta}\rangle_{\mathcal{V}_s} \leq \alpha^{-2}\|\check{\zeta}\|^2_{\mathcal{V}_s} \leq 1.$$

On the other hand, the parameter γ is a scaling positive parameters to balance uncertainty in the data information with respect to uncertainty in the source and in the initial condition. In practice for a given $\mathit{\Pi}_0^{-1}$ and κ, γ is adjusted with respect to the estimated observation noise scale δ.

In this setting, our objective is typically to find the trajectory associated with

$$\min_{\substack{\zeta\in\mathcal{V}_s \\ \nu\in \mathrm{L}^2((0,T);\mathcal{U})}} \mathscr{J}_T(\zeta,\nu),$$

and we emphasize that minimizing $\mathscr{J}_T$ must be understood as a minimization under the constraint that $z_{|\zeta,\nu}$ follows the dynamics (1.1).

Theorem 1.3. *There exists one, and only one minimizer couple $(\bar{\zeta}_T, \bar{\nu}_T) \in \mathcal{V}_s \times \mathrm{L}^2((0,T);\mathcal{U})$ of the criterion $\mathscr{J}_T$,*

$$(\bar{\zeta}_T, \bar{\nu}_T) = \operatorname*{argmin}_{\mathcal{V}_s\times \mathrm{L}^2((0,T);\mathcal{U})} \mathscr{J}_T(\zeta,\nu).$$

Moreover,

$$\bar{\zeta}_T = \mathit{\Pi}_0 \bar{q}_T(0), \quad \bar{\nu}_T(t) = \kappa^{-2}B^*\bar{q}_T(t),\, t\in(0,T), \tag{1.9}$$

where $(\bar{z}_T, \bar{q}_T)$ is the unique solution

$$\begin{cases} \dot{\bar{z}}_T + A\bar{z}_T = \kappa^{-2}BB^*\bar{q}_T, & \text{in } (0,T) \\ \dot{\bar{q}}_T - A^*\bar{q}_T = -\gamma C^*(y^\delta - C\bar{z}_T(t)), & \text{in } (0,T) \\ \bar{z}_T(0) = \hat{z}_0 + \mathit{\Pi}_0\bar{q}_T(0), \\ \bar{q}_T(T) = 0, \end{cases} \tag{1.10}$$

and $\bar{z}_T$ *is thus the solution of* (1.1) *associated with* $(\bar{\zeta}_T, \bar{\nu}_T)$.

Proof. From (1.3), we see that (1.8) defines a quadratic functional in the Hilbert space $\mathcal{V}_s \times \mathrm{L}^2((0,T);\mathcal{U})$. Moreover, from Duhamel's formula

$$\|z_{|\zeta,\nu}(t)\|_{\mathcal{Z}} \leq \|\zeta\|_{\mathcal{Z}} + \sqrt{t}\|\nu\|_{\mathrm{L}^2((0,T);\mathcal{U})} \leq \|\zeta\|_{\mathcal{V}_s} + \sqrt{t}\|\nu\|_{\mathrm{L}^2((0,T);\mathcal{U})},$$

ensures that $\mathscr{J}_T(\zeta,\nu)$ is Fréchet differentiable and we directly infer that

$$\mathscr{J}_T(\zeta_2,\nu_2) \geq \mathscr{J}_T(\zeta_1,\nu_1) + \big\langle \mathrm{D}\mathscr{J}_T(\zeta_1,\nu_1), (\zeta_2-\zeta_1, \nu_2-\nu_1)\big\rangle_{\mathcal{V}_s\times \mathrm{L}^2((0,T);\mathcal{U})} \\ + \frac{1}{2}\big(\varepsilon^2\|\zeta_1-\zeta_2\|^2_{\mathcal{V}_s} + \kappa^2\|\nu_1-\nu_2\|^2_{\mathrm{L}^2((0,T);\mathcal{U})}\big),$$

where ε^2 is the coercivity constant of $\varPi_0^{-1}$. Namely, $\mathscr{J}_T$ is a strongly convex function.

Therefore, there exists one, and only one, optimal estimation $(\bar{\zeta}_T, \bar{\nu}_T)$ such that

$$(\bar{\zeta}_T, \bar{\nu}_T) = \underset{\mathcal{V}_s\times \mathrm{L}^2((0,T);\mathcal{U})}{\operatorname{argmin}} \mathscr{J}_T(\zeta,\nu).$$

Note that $(\bar{\zeta}_T, \bar{\nu}_T)$ are indexed by T as $\mathscr{J}_T$ is.

Let us now introduce for all $z \in \mathrm{L}^2((0,T);\mathcal{Z})$ and $y \in \mathrm{L}^2((0,T);\mathcal{Y})$ the adjoint dynamics

$$\begin{cases} \dot{q}_T - A^* q_T = -\gamma C^*(y - Cz), & \text{in } (0,T) \\ q_T(T) = 0 \end{cases} \tag{1.11}$$

which is also well posed as it is considered backward in time. Namely, we have $q_T \in \mathrm{C}^0([0,T],\mathcal{Z})$ from Theorem 1.2. The adjoint variable allows to easily compute the Fréchet derivatives with respect to ζ and ν. We find for a given $(\zeta,\nu) \in \mathcal{V}_s \times \mathrm{L}^2([0,T],\mathcal{U})$

$$\forall \xi \in \mathcal{V}_s \quad \big\langle \mathrm{D}_\zeta \mathscr{J}_T(\zeta,\nu), \xi\big\rangle_{\mathcal{V}_s} = \big\langle \zeta, \varPi_0^{-1}\xi\big\rangle_{\mathcal{V}_s} + (q_T(0),\xi)_{\mathcal{Z}},$$

and

$$\forall \mu \in \mathrm{L}^2([0,T],\mathcal{U}) \quad \big\langle \mathrm{D}_\nu \mathscr{J}_T(\zeta,\nu), \mu\big\rangle_{\mathrm{L}^2((0,T);\mathcal{U})} = \int_0^T \kappa^2(\nu(t),\mu(t))_{\mathcal{U}} + (q_T(t), B\mu(t))_{\mathcal{Z}} \;\mathrm{d}t.$$

At extremum, we obtain the Euler equation associated with the minimization

$$\forall(\zeta,\nu) \in \mathcal{V}_s \times \mathrm{L}^2((0,T);\mathcal{U}), \quad \langle \varPi_0^{-1}\bar{\zeta}_T, \zeta\rangle_{\mathcal{V}_s} + (\bar{q}(0),\zeta)_{\mathcal{Z}} + \int_0^T \kappa^2(\bar{\nu}_T(t),\nu(t))_{\mathcal{U}} + (\bar{q}_T(t), B\nu(t))_{\mathcal{Z}} \;\mathrm{d}t = 0, \tag{1.12}$$

where $\bar{q}_T$ is the adjoint variable associated with the optimal trajectory $\bar{z}_T = z_{|\bar{\zeta}_T,\bar{\nu}_T}$ and the available measurements y^δ. This leads to the so-called two-ends problem defining the optimal dynamics of the estimator:

$$\begin{cases} \dot{\bar{z}}_T + A\bar{z}_T = BQB^*\bar{q}_T, & \text{in } (0,T) \\ \dot{\bar{q}}_T - A^*\bar{q}_T = -C^*R(y^\delta - C\bar{z}(t)), & \text{in } (0,T) \\ \bar{z}_T(0) = \hat{z}_0 + \varPi_0\bar{q}_T(0), \\ \bar{q}_T(T) = 0. \end{cases} \tag{1.13}$$

where $Q = \kappa^{-2}\mathbb{1}_{\mathcal{U}}$ and $R = \gamma\mathbb{1}_{\mathcal{Y}}$. □

Remark 1.4. Note that here, the proof is compatible with the semigroup theory solutions whereas a slightly different approach based on the notion of variational solution could have been possible. In particular, the variational frameworks allows to introduce the Lagrangian associated with the dynamics constraint hence to illuminate the adjoint equation definition. We refer to the first chapters of [7] that adapts to observation theory the variational evolution equation control framework initially developed in [39].

Remark 1.5. In general (1.13) is solved using a gradient descent approach. Defining the gradient from Riesz' representation theorem in the space $\mathcal{V}_s \times \mathrm{L}^2((0,T);\mathcal{U})$, the gradient descent approach reads

$$\begin{cases} \zeta^{k+1} = \zeta^k - \rho^k \mathit{\Pi}_0 \nabla_\zeta \mathscr{J}_T(\zeta^k, \nu^k), & k \geq 0 \\ \nu^{k+1} = \nu^k - \rho^k \kappa^{-2} \nabla_\nu \mathscr{J}_T(\zeta^k, \nu^{k+1}), & k \geq 0 \end{cases}$$

from $(\zeta^0, \nu^0) = (0,0)$, can be proved to be convergent for an adequate relaxation parameter $\rho < 1$, small enough. This gradient descent consists in solving, from $(\zeta^0, \nu^0) = (0,0)$ and for $k \geq 0$, the weakly coupled system

$$\begin{cases} \dot{z}^{k+1} + Az^{k+1} = (1-\rho^k)B\nu^k + \rho^k BQB^* q_T^k, & \text{in } (0,T) \\ z^{k+1}(0) = (1-\rho^k)z^k(0) + \rho^k \mathit{\Pi}_0 q_T^k(0) \end{cases} \tag{1.14a}$$

and

$$\begin{cases} \dot{q}_T^{k+1} - A^* q_T^{k+1} = -\gamma C^*(y^\delta - Cz_T^{k+1}), & \text{in } (0,T) \\ q_T^{k+1}(T) = 0 \end{cases} \tag{1.14b}$$

Note that the existence of a solution of (1.13) can be understood as the limit of the well-posed dynamics (1.14a)–(1.14b).

1.5.2. Singular value decomposition

In this section, we want to give one example of possible choice of $\mathit{\Pi}_0$ among many others. In this respect, let us consider the compact operator $T = A^{-1} : \mathcal{Z} \to \mathcal{Z}$. We introduce $(e_n)_{n\in\mathbb{N}}$ and $(f_n)_{n\in\mathbb{N}}$ respectively the orthonormal basis associated with the diagonalization of T^*T and TT^* respectively. We denote $(\mu_n)_{n\in\mathbb{N}}$ the sequence of positive eigenvalues which decrease to 0.

We recall the following decomposition

$$Tz = \sum_{n\geq 0} \mu_n (z, f_n)_{\mathcal{Z}}\, e_n, \quad T^*z = \sum_{n\geq 0} \mu_n (z, e_n)_{\mathcal{Z}}\, f_n.$$

with $(e_m, e_n)_{\mathcal{Z}} = \delta_{mn}$, $(f_m, f_n)_{\mathcal{Z}} = \delta_{mn}$, and $Tf_n = \mu_n e_n, \quad T^*e_n = \mu_n f_n$. We can then define

$$(TT^*)^s z = \sum \mu_n^{2s} (z, e_n)_{\mathcal{Z}}\, e_n,$$

and given $s > 0$, we introduce

$$\mathcal{V}_s = \mathrm{Im}((TT^*)^s) = \Big\{ z \in \mathcal{Z} \,\Big|\, \sum_{n\geq 0} \frac{(z, e_n)_{\mathcal{Z}}^2}{\mu_n^{4s}} \leq +\infty \Big\}.$$

Note that we have in particular $\mathcal{V}_{\frac{1}{4}} = \mathcal{V}$ as we initially restrict our study to the case where $\mathcal{D}(A^{\frac{1}{2}}) = \mathcal{D}(A^{\frac{1}{2}*}) = \mathcal{V}$. We further assume that there exists $s_0 \geq \frac{1}{2}$ such that

$$\forall s \geq s_0, \quad \sum_{n\geq 0} \mu_n^{2s} < +\infty.$$

Then, we propose to define $\mathit{\Pi}_0$ as

$$\mathit{\Pi}_0 z = \sum_{n\geq 0} \alpha^2 \mu_n^{2s} (z, e_n)_{\mathcal{Z}}\, e_n. \tag{1.15}$$

If $s \geq 0$, $\mathit{\Pi}_0$ is compact. Moreover if $s \geq s_0$, $\mathit{\Pi}_0$ is a Hilbert-Schmidt operator (see the recalled definition in Sect. 2.1). Finally,

$$\langle \mathit{\Pi}_0^{-1} z, z\rangle_{\mathcal{V}_s} = \sum_{n\geq 0} \alpha^{-2}\mu_n^{-2s}(z, e_n)^2_{\mathcal{Z}} = \alpha^{-2}\|z\|^2_{\mathcal{V}_s},$$

hence ensuring that a_s is bounded and coercive in $\mathcal{V}_s$.

1.6. The Kalman sequential estimator

The principle of optimal sequential estimation [7] is to avoid solving (1.13) by decoupling the corresponding two-ends dynamics. In this respect, we will find a Cauchy problem formulation of the so-called *optimal sequential estimator* – also called *optimal observer* – defined by the following:

Definition 1.6 (The optimal sequential estimator)**.** For all time $t > 0$, considering the optimal trajectory $\bar{z}_t$ associated with the minimizer of $\mathscr{J}_t$, the optimal sequential estimator $\hat{z}$ is defined by

$$\forall t \geq 0, \quad \hat{z}(t) = \bar{z}_t(t). \tag{1.16}$$

We are going to prove that the optimal sequential estimator is in fact given by the estimator proposed by [32], and often called Kalman-Bucy estimator or simply Kalman estimator, here generalized to PDE formulations [7]. Yet, we need to introduce the so-called *Riccati operator* solution to a Riccati dynamics before characterizing the dynamics of $\hat{z}$.

1.6.1. Riccati dynamics

We introduce the spaces of linear auto-adjoint bounded operators

$$\mathcal{S}(\mathcal{Z}) = \Big\{ Q \in \mathcal{L}(\mathcal{Z}) \,\Big|\, Q = Q^* \Big\},$$

and the cone in $\mathcal{S}(\mathcal{Z})$

$$\mathcal{S}^+(\mathcal{Z}) = \Big\{ Q \in \mathcal{S}(\mathcal{Z}) \,\Big|\, \forall z \in \mathcal{Z},\ (z, Qz) \geq 0 \Big\}.$$

We then consider the following Riccati dynamics

$$\begin{cases} \dot{\mathit{\Pi}} + A\mathit{\Pi} + \mathit{\Pi}A^* + \mathit{\Pi}C^*RC\mathit{\Pi} - B^*QB = 0, \quad t > 0 \\ \mathit{\Pi}(0) = \mathit{\Pi}_0. \end{cases} \tag{1.17}$$

for which we seek a solution in $\mathrm{C}^0([0,T],\mathcal{S}^+(\mathcal{Z}))$, the set of all continuous mappings from $[0,T]$ to $\mathcal{S}^+(\mathcal{Z})$, endowed with the topology of pointwise convergence:

$$\lim_{n\to+\infty} P_n = P \quad \Leftrightarrow \quad \forall z\in\mathcal{Z}, \quad \lim_{n\to+\infty} P_n z = Pz.$$

As for the evolution equation (1.2), we face several types of solution of (1.17), listed in the next Definition 1.7. Before that, we need to introduce an operator over the set of bounded symmetric operator and its associated domain. Let us denote

$$\Upsilon : \left| \begin{aligned} \mathcal{S}(\mathcal{Z}) &\to \mathcal{S}(\mathcal{Z}) \\ Q &\mapsto AQ + QA^* \end{aligned} \right.$$

with, for any $Q\in\mathcal{S}(\mathcal{Z})$, the corresponding bilinear form

$$\upsilon_Q(z_1,z_2) = (Qz_1, A^*z_2) + (A^*z_1, Qz_2), \quad \forall (z_1,z_2)\in\mathcal{D}(A^*),$$

so that Υ is defined with the domain

$$\mathcal{D}(\Upsilon) = \{Q\in\mathcal{S}(\mathcal{Z}) \mid \upsilon_P \text{ is continuous in } \mathcal{Z}\times\mathcal{Z}\}.$$

Note that, going back to our definition of $\varPi_0$ from the singular value decomposition of A in Section 1.5.2, we have that $\varPi_0\in\mathcal{D}(\Upsilon)$ when $s\geq\frac{1}{2}$. Indeed, we easily find in this case

$$\begin{aligned} &(\varPi_0 z_1, A^*z_2) + (A^*z_1, \varPi_0 z_2) \\ &\qquad = \sum_{n\geq 0} \alpha^2 \mu_n^{2s} \big[(z_1,e_n)(A^*z_2,e_n) + (z_2,e_n)(A^*z_1,e_n)\big] \\ &\qquad = \sum_{n\geq 0} \alpha^2 \mu_n^{2s} \big[(z_1,e_n)(z_2,Ae_n) + (z_2,e_n)(z_1,Ae_n)\big] \\ &\qquad = \sum_{n\geq 0} \alpha^2 \mu_n^{2s-1} \big[(z_1,e_n)(z_2,f_n) + (z_2,e_n)(z_1,f_n)\big] \\ &\qquad \leq c_{\mathrm{st}} \sum_{n\geq 0} \big[(z_1,e_n)(z_2,f_n) + (z_2,e_n)(z_1,f_n)\big] \leq \tilde{c}_{\mathrm{st}} \|z_1\|_{\mathcal{Z}} \|z_2\|_{\mathcal{Z}}. \end{aligned}$$

Definition 1.7 (Notion of Riccati solution)**.** We list 3 different notions of solution to Problem (1.17):

(i) A strict solution is a function $\varPi\in\mathrm{C}^1([0,T];\mathcal{S}(\mathcal{Z}))$ solution to (1.17) where, for all $t\in[0,T]$, $\varPi(t)\in\mathcal{D}(\Upsilon)$ and $\Upsilon(P)\in\mathrm{C}^0([0,T],\mathcal{S}(\mathcal{Z}))$.

(ii) A mild solution to (1.17) is a function $\varPi\in\mathrm{C}^0([0,T],\mathcal{S}(\mathcal{Z}))$ that satisfies for all $z\in\mathcal{Z}$, $t\in[0,T]$,

$$\varPi(t)z = \varPhi(t)\varPi_0\varPhi^*(t) - \int_0^t \varPhi(t-s)\varPi(s)C^*RC\varPi(s)\varPhi^*(t-s)\ \mathrm{d}s + \int_0^t \varPhi(s)BQB^*\varPhi^*(s)\ \mathrm{d}s. \qquad (1.18)$$

(iii) A weak solution to (1.17) is a function $\varPi\in\mathrm{C}^0([0,T],\mathcal{S}(\mathcal{Z}))$ such that for all $(z_1,z_2)\in\mathcal{D}(A^*)$, $(\varPi(\cdot)z_1,z_2)_{\mathcal{Z}}$ is differentiable and verifies

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}(\varPi(t)z_1,z_2)_{\mathcal{Z}} + (\varPi(t)z_1,A^*z_2)_{\mathcal{Z}} + (\varPi(t)A^*z_1,z_2)_{\mathcal{Z}} \\ &\qquad\qquad + (C\varPi(t)z_1, RC\varPi(t)z_2)_{\mathcal{Y}} - (B^*z_1, QB^*z_2)_{\mathcal{U}} = 0, \quad t\in[0,T]. \qquad (1.19) \end{aligned}$$

Then we have the following existence results of the Riccati operator $\mathit{\Pi}$, also called *covariance operator* for its interpretation in the stochastic filtering framework [7, 19].

Theorem 1.8 (Existence of Riccati solution). *We list 2 cases of existence of a solution to Problem* (1.17)*:*

(i) Assuming that B and C are bounded and $\mathit{\Pi}_0 \in \mathcal{S}^+(\mathcal{Z})$, the Riccati dynamics (1.17) *admits one and only one weak solution $\mathit{\Pi} \in \mathrm{C}^0([0,T], \mathcal{S}^+(\mathcal{Z}))$, which is also a mild solution in the sense of* (1.18)*.*

(ii) Assuming moreover that $\mathit{\Pi}_0 \in \mathcal{D}(\Upsilon)$, then the Riccati dynamics (1.17) *admits one and only one strict solution $\mathit{\Pi} \in \mathrm{C}^1([0,T], \mathcal{S}^+(\mathcal{Z}))$.*

Proof. The existence and uniqueness of a mild solution is justified in IV-1 Theorem 2.2 of [9]. The fact that this solution is also a weak solution is given in IV-1 Proposition 2.1 of [9]. Finally, the existence of a strict solution is proved for variational operator in IV-1 Proposition 3.2 of [9]. □

Remark 1.9 (Time-dependent observation and control operators). Theorem 1.8 directly extends to cases where we have a time-dependent observation operator – time-dependent control operator resp. – as soon as $t \mapsto C(t)$ is continuous in time – $t \mapsto B(t)$ is continuous in time resp. –

We finally conclude this section by recalling the benefit of relying on variational semigroup to impose additional regularity properties for the Riccati operator $\mathit{\Pi}$, proved in IV-1 Theorem 3.3 of [9] using an initial result from [24].

Theorem 1.10. *Let $\mathit{\Pi} \in \mathrm{C}^1([0,T]; \mathcal{S}^+(\mathcal{Z}))$ be the strict solution to* (1.17) *initialized from* (1.15)*. Then for any $t \in (0,T)$ and $z \in \mathcal{D}(A^{*\frac{1}{2}})$ we have $\mathit{\Pi}(t)z \in \mathcal{D}(A^{\frac{1}{2}})$. Moreover if $\mathcal{D}(A^{\frac{1}{2}}) = \mathcal{D}(A^{*\frac{1}{2}}) = \mathcal{V}$, then $\mathit{\Pi}(t) \in \mathcal{L}(\mathcal{V}', \mathcal{V})$.*

1.6.2. Comparison principle

Fundamental properties of the Riccati operator come from comparison principles. We have already seen that for all $t \geq 0$, $\mathit{\Pi}(t) \in \mathcal{S}(\mathcal{Z})$. Moreover, Theorem 1.8 also gives $\mathit{\Pi}(t) \geq 0$, a property which can be easily understood from the recast dynamics

$$\dot{\mathit{\Pi}} + \Big[A + \frac{1}{2}\mathit{\Pi}(t)C^*RC\Big]\mathit{\Pi} + \mathit{\Pi}\Big[A^* + \frac{1}{2}\mathit{\Pi}(t)C^*RC\Big] - B^*QB = 0,$$

leading to the mild solution

$$\mathit{\Pi}(t) = \tilde{\mathit{\Phi}}(t,0)\mathit{\Pi}_0\tilde{\mathit{\Phi}}^*(t,0) + \int_0^t \tilde{\mathit{\Phi}}(t,s)BQB^*\tilde{\mathit{\Phi}}^*(t,s)\ \mathrm{d}s, \tag{1.20}$$

where $\tilde{\mathit{\Phi}}(s,t), 0 \leq s \leq t \leq T$ is the evolution operator associated with the perturbed operator $A + \frac{\gamma}{2}\mathit{\Pi}(t)C^*C$ – see IV-1 Theorem 2.1 and I-1 §3.5 of [9]. From (1.20) indeed, we directly infer that $\mathit{\Pi}(t) \geq 0$.

Then from (1.18), we infer that

$$\mathit{\Pi}(t) \leq \mathit{\Phi}(t)\mathit{\Pi}_0\mathit{\Phi}^*(t) + \int_0^t \mathit{\Phi}(s)BQB^*\mathit{\Phi}^*(s)\ \mathrm{d}s, \tag{1.21}$$

where the order relation in $\mathcal{S}^+(\mathcal{Z})$ is obviously given by

$$Q_1 \leq Q_2 \Leftrightarrow \forall z \in \mathcal{Z},\ (z, Q_1 z)_{\mathcal{Z}} \leq (z, Q_2 z)_{\mathcal{Z}}.$$

We now recall the classical comparison result of Riccati operator proved in IV-1 Propisiton 2.2 of [9]

Proposition 1.11. *Consider for $i = \{1, 2\}$, the two Riccati equations*

$$\begin{cases} \dot{\mathit{\Pi}}_i + A\mathit{\Pi}_i + \mathit{\Pi}_i A^* + \mathit{\Pi}_i C_i^* R_i C_i \mathit{\Pi}_i - B_i Q_i B_i = 0, \quad t > 0 \\ \mathit{\Pi}(0) = \mathit{\Pi}_{0,i} \end{cases}$$

with $\mathit{\Pi}_{0,1} \leq \mathit{\Pi}_{0,2}, \quad B_1 Q_1 B_1 \leq B_2 Q_2 B_2, \quad C_2 R_2 C_2 \leq C_1 R_1 C_1$.
Then we have for all $t \geq 0$, $\mathit{\Pi}_1(t) \leq \mathit{\Pi}_2(t)$.

By a direct use of Proposition 1.11, we obtain the following comparison

$$\forall t > 0, \quad \mathit{\Pi}(t) \geq \mathit{\Pi}_c(t),$$

where $\mathit{\Pi}_c$ − given our choice of $\mathit{\Pi}_0$ − is the strict solution of

$$\begin{cases} \dot{\mathit{\Pi}}_c + A\mathit{\Pi}_c + \mathit{\Pi}_c A^* + \mathit{\Pi}_c C^* R C \mathit{\Pi}_c = 0, \quad t > 0 \\ \mathit{\Pi}_c(0) = \mathit{\Pi}_0 \end{cases} \tag{1.22}$$

Let us now specify $\mathit{\Pi}_c$, which is given by the following decomposition.

Proposition 1.12. *The unique strict solution of* (1.22) *is given by*

$$\mathit{\Pi}_c(t) = \mathit{\Phi}(t)\Big(\mathit{\Pi}_0^{-1} + \int_0^t \mathit{\Phi}^*(s) C^* R C \mathit{\Phi}(s) \ \mathrm{d}s\Big)^{-1} \mathit{\Phi}^*(t). \tag{1.23}$$

Proof. On the one hand, we introduce the operator Λ solution of

$$\begin{cases} \dot{\Lambda} + \Lambda \mathit{\Phi}^* C^* R C \mathit{\Phi} \Lambda = 0, \quad t > 0 \\ \Lambda(0) = \mathit{\Pi}_0 \end{cases} \tag{1.24}$$

As $\mathit{\Pi}_0$ is a bounded operator, $\Lambda \in \mathrm{C}^1([0, T], \mathcal{S}^+(\mathcal{Z}))$ is the strict solution – see Theorem 1.8 and Remark 1.9 – of the Riccati equation (1.24). Moreover, as a mild solution, Λ satisfies

$$\Lambda : \mathbb{R} \ni t \mapsto \mathit{\Pi}_0 - \int_0^t \Lambda(s) \mathit{\Phi}^*(s) C^* R C \mathit{\Phi}(s) \Lambda(s) \ \mathrm{d}s \in \mathcal{S}^+(\mathcal{Z}),$$

On the other hand, we introduce

$$\forall t \in [0, T_\varepsilon], \quad U(t) = \mathit{\Pi}_0^{-1} + \int_0^t \mathit{\Phi}^*(s) C^* R C \mathit{\Phi}(s) \ \mathrm{d}s.$$

which is also C^1 in time since $s \mapsto \mathit{\Phi}^*(s) C^* R C \mathit{\Phi}(s) \in \mathrm{C}^0([0, T], \mathcal{S}^+(\mathcal{Z}))$. Moreover $U \geq \mathit{\Pi}_0^{-1}$ is invertible in $[0, T]$. By composing the derivatives, we find that

$$\frac{\mathrm{d}}{\mathrm{d}t}(U^{-1}) = -U^{-1} \dot{U} U^{-1} = -U^{-1} \mathit{\Phi}^* C^* R C \mathit{\Phi} U^{-1}, \quad \text{in } [0, T] \tag{1.25}$$

Namely $U^{-1} = \Lambda$ in $[0, T]$ by uniqueness of the strict solution of (1.24). In addition, we find that $\mathit{\Pi}_\Lambda = \mathit{\Phi}(t)\Lambda(t)\mathit{\Phi}^*(t)$ is solution of

$$\mathit{\Pi}_\Lambda(t) z = \mathit{\Phi}(t) \Lambda(t) \mathit{\Phi}^*(t) z$$

$$
\begin{aligned}
&= \mathit{\Phi}(t)\Big[\Lambda(0) - \int_0^t \Lambda(s)\mathit{\Phi}(s)C^* RC\mathit{\Phi}(s)\Lambda(s)\ \mathrm{d}s\Big]\mathit{\Phi}^*(t)z \\
&= \mathit{\Phi}(t)\Lambda(0)\mathit{\Phi}^*(t)z - \int_0^t \mathit{\Phi}(t-s)\mathit{\Pi}_\Lambda(s)C^* RC\mathit{\Pi}_\Lambda(s)\mathit{\Phi}^*(t-s)z\ \mathrm{d}s.
\end{aligned}
$$

This ensures, by uniqueness of the mild solution of (1.22), that $\mathit{\Pi}_\Lambda = \mathit{\Pi}_c$. Therefore $\mathit{\Pi}_c = \mathit{\Phi}\Lambda\mathit{\Phi}^* = \mathit{\Phi}U^{-1}\mathit{\Phi}^*$. □

To summarize, we have found in this section that, for all $t \geq 0$,

$$
0 \leq \mathit{\Phi}(t)\Big(\mathit{\Pi}_0^{-1} + \int_0^t \mathit{\Phi}^*(s)C^* RC\mathit{\Phi}(s)\ \mathrm{d}s\Big)^{-1}\mathit{\Phi}^*(t) \leq \mathit{\Pi}(t) \leq \mathit{\Phi}(t)\mathit{\Pi}_0\mathit{\Phi}^*(t) + \int_0^t \mathit{\Phi}(s)BQB^*\mathit{\Phi}^*(s)\ \mathrm{d}s, \tag{1.26}
$$

hence controlling the asymptotic behavior of $\mathit{\Pi}$.

1.6.3. Estimator dynamics

Once we have clarified the sense given to the covariance operator $\mathit{\Pi}$ solution to the Riccati dynamics (1.17), we are able to give to (1.16) a closed loop dynamics, as recall in the next theorems.

Theorem 1.13. *Let A be the generator of C^0-semigroup $\mathit{\Phi}$. Let C be a bounded operator and $\mathit{\Pi}$ a mild solution of the (1.17). There exists one and only one* mild *solution in $\mathrm{C}^0([0,T];\mathcal{Z})$ of the dynamics*

$$
\begin{cases}
\dot{\hat{z}} + A\hat{z} = \mathit{\Pi}C^* R(y^\delta - C\hat{z}), & \text{in } (0,T) \\
\hat{z}(0) = \hat{z}_0
\end{cases} \tag{1.27}
$$

in the sense that

$$
\hat{z}(t) = \mathit{\Phi}(t)\hat{z}_0 + \int_0^t \mathit{\Phi}(t-s)\mathit{\Pi}(s)C^*[y^\delta(s) - C\hat{z}(s)]\ \mathrm{d}s. \tag{1.28}
$$

Moreover, this solution is the unique weak *solution in the sense that, (1.) $z \in \mathrm{L}^2((0,T);\mathcal{Z})$, (2.) for all $q \in \mathcal{D}(A^*)$, $\langle q, \hat{z}(\cdot)\rangle$ belongs to $\mathrm{H}^1(0,T)$ and (3.) for almost all $t \in (0,T)$,*

$$
\forall q \in \mathcal{D}(A^*), \quad \frac{\mathrm{d}}{\mathrm{d}t}(\hat{z}, q) + (\hat{z}, A^* q) = \gamma(y^\delta - C\hat{z}, C\mathit{\Pi}(t)q)_\mathcal{Y}. \tag{1.29}
$$

Proof. Let us denote $\beta : t \mapsto \gamma\mathit{\Pi}(t)C^* y^\delta(t) \in \mathrm{L}^2((0,T),\mathcal{Z})$, and $G : t \mapsto \gamma\mathit{\Pi}(t)C^* C \in \mathrm{C}^0([0,T];\mathcal{L}(\mathcal{Z}))$. The dynamics (1.27) is a specific case of

$$
\begin{cases}
\dot{\hat{z}} + A\hat{z} = G(t)\hat{z}(t) + \beta(t), & t > 0 \\
\hat{z}(0) = \hat{z}_0
\end{cases} \tag{1.30}
$$

with $\hat{z}_0 \in \mathcal{Z}$ and $\beta \in \mathrm{L}^2((0,T);\mathcal{Z})$. From II-1 Propostion 3.4 of [9], Problem (1.30) admits a unique weak in $\mathrm{H}^1((0,T);\mathcal{D}(A^*)') \cap \mathrm{C}^0([0,T];\mathcal{Z})$ which coincides with the mild solution in the sense of (1.28).

□

Note that we can also deduce that $\hat{z}$ is also the unique *varitionnal* solution in the sense that, (1.) $z \in \mathrm{L}^2((0,T);\mathcal{V})$ and (2.) $\frac{\mathrm{d}z}{\mathrm{d}t} \in \mathrm{L}^2((0,T);\mathcal{V}')$ and (3.) for almost all $t \in (0,T)$,

$$
\forall w \in \mathcal{V}, \quad \Big\langle \frac{\mathrm{d}}{\mathrm{d}t}\hat{z} + A\hat{z}, w\Big\rangle_\mathcal{V} = \gamma(y^\delta - C\hat{z}, C\mathit{\Pi}(t)w)_\mathcal{Y} \tag{1.31}
$$

Theorem 1.14. *The Kalman observer defined by (1.16) is the unique solution of (1.27). Moreover assuming that $\Pi_0 \in \mathcal{D}(\Upsilon)$, we have the fundamental identity*

$$\forall t \in [0,T], \quad \bar{z}_T(t) = \hat{z}(t) + \Pi(t)\bar{q}_T(t). \tag{1.32}$$

Proof. From Theorem 1.3, we have the existence of a weak solution of the two-ends problem (1.13). From Theorem 1.8, we have the existence of a strict solution of the covariance operator $\Pi \in C^1([0,\infty[;\mathcal{S}^+(\mathcal{Z}))$. Additionally, a weak solution of the Kalman estimator $\hat{z}$ exists from Theorem 1.13.

Now let us introduce $\eta = \hat{z} - \bar{z}_T + \Pi\bar{q}_T$ and $v \in \mathcal{D}(A^*)$, and compute

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(\eta(t),v)_{\mathcal{Z}} = -\big(\hat{z}(t),A^*v\big)_{\mathcal{Z}} + \gamma\big(y^\delta(t) - C\hat{z}(t), C\Pi(t)v\big)_{\mathcal{Y}} \\ + \big(\bar{z}_T(t),A^*v\big)_{\mathcal{Z}} - \kappa^{-2}\big(B^*\bar{q}_T(t),B^*v\big)_{\mathcal{U}} \\ + \frac{\mathrm{d}}{\mathrm{d}t}\big(\Pi(t)\bar{q}_T(s),v\big)_{\mathcal{Z}}\Big|_{s=t} + \frac{\mathrm{d}}{\mathrm{d}t}\big(\bar{q}_T(t),\Pi(s)v\big)_{\mathcal{Z}}\Big|_{s=t}.\end{aligned} \tag{1.33}$$

Moreover, as Π is a weak solution of (1.17) such that for all $t \geq 0$ and $v \in \mathcal{D}(A^*)$, $\Pi(t)z \in \mathcal{D}(A^*)$, we have from (1.19)

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(\Pi(t)\bar{q}_T(s),v)\Big|_{s=t} + (\Pi(t)\bar{q}_T(t),A^*v)_{\mathcal{Z}} + (\bar{q}_T(t),A^*\Pi(t)v)_{\mathcal{Z}} \\ + \gamma(C\Pi(t)\bar{q}_T(t),C\Pi(t)v)_{\mathcal{Y}} - \kappa^{-2}(B^*\bar{q}_T(t),B^*v)_{\mathcal{U}} = 0,\end{aligned} \tag{1.34}$$

and, as $\bar{q}$ is a weak solution of (1.13),

$$\frac{\mathrm{d}}{\mathrm{d}t}(\bar{q}_T(t),\Pi(s)v)\Big|_{s=t} = (\bar{q}_T(t),A^*\Pi(t)v)_{\mathcal{Z}} - \gamma\big(y^\delta(t) - C\bar{z}_T(t),C\Pi(t)v\big)_{\mathcal{Y}}. \tag{1.35}$$

Gathering (1.33), (1.34) and (1.35), we finally obtain

$$\begin{cases}\dfrac{\mathrm{d}}{\mathrm{d}t}(\eta(t),v)_{\mathcal{Z}} + (\eta,[A^* + C^*RC\Pi(t)]v)_{\mathcal{Z}} = 0, \quad t \in [0,T] \\ \eta(0) = 0,\end{cases}$$

whence, by Theorem 1.13, $\eta = 0$ in $[0,T]$, which concludes the proof □

We directly deduce from (1.32) that

$$\forall T > 0, \quad \bar{z}_T(T) = \hat{z}(T).$$

As a consequence, the Kalman-Bucy estimator, solution of (1.27), is the optimal estimator in the sense of Definition 1.16.

Remark 1.15. The condition $\Pi_0 \in \mathcal{D}(\Upsilon)$ is a technical assumption facilitating the chain rule computations. It can be relaxed by using only mild solutions and Duhamel formulae. However, the proof of Theorem 1.14 is simplified in this case, and the next sections will be based on such regularity condition on Π_0.

2. Kernel representation of continuous-time infinite dimensional Riccati solutions

We are now going to show an additional regularity property of the Riccati solution which for $\varPi_0$ regularizing enough can be of Hilbert-Schmidt type as it is sometime highlighted for control problem – see for instance [20] and references therein – or [13–15] for observation problems. The benefit of such property will be the existence of an associated kernel, regular enough so that it will be numerically approximated and efficiently computed.

2.1. Hilbert-Schmidt Riccati solutions and Kernel representation

We denote by $\mathcal{J}_2(\mathcal{Z})$ the spaces of Hilbert-Schmidt operators, over the separable Hilbert space $\mathcal{Z}$. We recall that

$$\mathcal{J}_2(\mathcal{Z}) \subset \mathcal{K}(\mathcal{Z}) \subset \mathcal{L}(\mathcal{Z}),$$

where $\mathcal{K}(\mathcal{Z})$ is the space of compact operators. For any Hilbert basis $(e_n)_{n\geq 0}$ of $\mathcal{Z}$.

$$\|\varPi\|_2 = \sqrt{\operatorname{tr}(\varPi\varPi^*)} = \Big(\sum_{n\geq 0}(e_n, \varPi^2 e_n)_{\mathcal{Z}}\Big)^{\frac{1}{2}}.$$

We point out that this definition can be shown to be independent of the choice of the Hilbert basis. Moreover, if $\varPi \in \mathcal{J}_2(\mathcal{Z}) \cap \mathcal{S}^+(\mathcal{Z})$ then $\|\varPi\| \leq \|\varPi\|_2$. We have the following result taken from Theorem 3.6 of [14] – see also [13, 15] and the seminal work [50].

Theorem 2.1 ([14])**.** *Let us assume that (1) $\varPi_0 \in \mathcal{S}^+(\mathcal{Z}) \cap \mathcal{J}_2(\mathcal{Z})$, (2) $BB^* \in \mathcal{S}^+(\mathcal{Z}) \cap \mathcal{J}_2(\mathcal{Z})$ and (3) $C^*C \in \mathcal{S}^+(\mathcal{Z})$. Then, the Riccati dynamics* (1.17) *admits one and only one mild solution $\varPi$ in the sense of* (1.18) *and $\varPi \in C([0,T], \mathcal{S}^+(\mathcal{Z})) \cap C([0,T], \mathcal{J}_2(\mathcal{Z}))$.*

We now recall the following Kernel Theorems for Hilbert-Schmidt operators, see Theorem 12.6.2 and Theorem 12.7.2 of [4]

Theorem 2.2 (Kernel Theorem in L^2)**.** *An operator $\varPi$ from $\mathrm{L}^2(\Omega)$ to $\mathrm{L}^2(\Omega)$ is a Hilbert-Schmidt operator if and only if it is associated with a kernel $\pi \in \mathrm{L}^2(\Omega\times\Omega)$ such that*

$$\forall \varphi \in L^2(\Omega), \quad \forall x \in \Omega, \quad (\varPi\varphi)(x) = \int_\Omega \pi(x', x)\varphi(x') \ \mathrm{d}x'.$$

and

$$\|\varPi\|_2 = \|\pi\|_{L^2(\Omega\times\Omega)}. \tag{2.1}$$

Theorem 2.3 (Kernel Theorem in Sobolev Spaces)**.** *Let $(m,p) \in \mathbb{N}^2$. An operator $\varPi$ from $\mathrm{H}^m(\Omega)'$ to $\mathrm{H}^p(\Omega)$ is a Hilbert-Schmidt operator if and only if it is associated with a kernel $\pi \in \mathrm{H}^{m,p}(\Omega\times\Omega)$ such that*

$$\forall \psi \in H^m(\Omega)', \quad (\varPi\psi)(x) = \langle \psi, \pi(\cdot, x)\rangle_{H^m}.$$

Therefore in our framework, as we proved that $\varPi(t) \in \mathcal{L}(\mathcal{V}', \mathcal{V})$ for $t \in (0,T)$, we can expect an additional regularity for the kernel π associated with $\varPi$. This will be specified in the next section for the advection diffusion case.

2.2. Kernel representation of the Kalman estimator for an advection-diffusion problem

Considering our specific advection-diffusion example of dynamics defined in (1.7), we have the following representation theorem.

Theorem 2.4. *Let $\Pi_0 \in \mathcal{D}(\Upsilon) \cap \mathcal{J}_2(L^2(\Omega)) \cap \mathcal{L}(H^{-1}(\Omega), \mathrm{H}^1_0(\Omega))$. The Riccati dynamics (1.17) associated with the model (1.7) admits one and only one mild solution Π in the sense of (1.18) which is associated with a kernel $\pi \in \mathrm{C}^1([0,T]; \mathrm{L}^2(\Omega\times\Omega)) \cap \mathrm{L}^2([0,T]; \mathrm{H}^1_0(\Omega\times\Omega))$ such that for all $t \in (0,T)$, $(\Delta_x + \Delta'_x)\pi(t) \in \mathrm{L}^2(\Omega\times\Omega)$ and π is solution to the dynamics*

$$\begin{cases} \partial_t \pi(x',x,t) \\ \quad - v(x')\cdot\nabla_{x'}\pi(x',x,t) - v(x)\cdot\nabla_x\pi(x',x,t) \\ \quad - (\Delta_x + \Delta_{x'})\pi(x',x,t) \\ \quad = \kappa^{-2} f(x')f(x) \\ \qquad -\gamma \displaystyle\int_\omega \pi(t,x,x'')\pi(t,x'',x') \ \mathrm{d}x'', & (x',x,t) \in \Omega\times\Omega\times(0,T), \\ \pi(x',x,t) = 0 & (x',x,t), \in \partial\Omega\times\Omega\times(0,T), \\ \pi(x',x,t) = 0 & (x',x,t), \in \Omega\times\partial\Omega\times(0,T), \\ \pi(x',x,0) = \pi_0(x',x), & (x',x) \in \Omega\times\Omega. \end{cases} \tag{2.2}$$

where $\pi_0 \in \mathrm{H}^1(\Omega\times\Omega)$ is the kernel associated with the initial covariance operator Π_0.

Proof. We introduce the sequence of eigenvalues $(\lambda_n)_{n\in\mathbb{N}}$ of $(-\Delta)^{-1}$, which is decreasing to 0. The corresponding eigenvectors $(u_n)_{n\in\mathbb{N}}$ define an orthonormal basis of $L^2(\Omega)$,

From Theorem 2.1, The Riccati dynamics (1.17) admits a unique strict solution $\Pi \in \mathrm{C}^1([0,T]; \mathcal{S}^+(\mathcal{Z}))$ with also $\Pi \in \mathrm{C}^0([0,T]; \mathcal{J}_2(\mathcal{Z}))$. Therefore, we have for all $t \in (0,T)$, $\sum_{n\geq 0}(u_n, \Pi^2(t)u_n) < +\infty$ which implies from Theorem 2.2 that $\Pi(t)$ admits a kernel representation $\pi(t) \in \mathrm{L}^2(\Omega\times\Omega)$. Moreover, from Theorem 1.10, we also have that for all $t \in (0,T)$, $\Pi(t) \in \mathcal{L}(\mathcal{V}', \mathcal{V})$. Let us define for all n, $h_n = \frac{1}{\sqrt{\lambda_n}} u_n$ and $g_n = \sqrt{\lambda_n} u_n$. We know that $(h_n)_{n\in\mathbb{N}}$ is a Hilbert basis of $\mathcal{V}' = \mathrm{H}^{-1}(\Omega)$ and $(g_n)_{n\in\mathbb{N}}$ is a Hilbert basis of $\mathcal{V} = \mathrm{H}^1_0(\Omega)$. We thus have

$$\sum_{n\geq 0}(g_n, \Pi^2(t)h_n) = \sum_{n\geq 0}(u_n, \Pi^2(t)u_n) < +\infty,$$

which implies that $\Pi(t)$ is a Hilbert-Schmidt operator from $\mathrm{H}^{-1}(\Omega)$ to $\mathrm{H}^1_0(\Omega)$. Therefore this time, from Theorem 2.3, $\Pi(t)$ admits a kernel representation $\pi(t) \in \mathrm{H}^1(\Omega\times\Omega)$. In particular, π admits a trace at the boundary $\partial(\Omega\times\Omega)$

Let us now characterize more specifically this kernel π. First, we have for all $t \in [0,T]$, $\Pi(t) \in \mathcal{S}^+(\mathcal{Z})$, hence

$$\forall (x,x',t) \in \Omega\times\Omega\times[0,T], \quad \pi(x',x,t) = \pi(x,x',t).$$

Second, we recall that, from Theorem 1.10, for $t > 0$ and $z \in \mathcal{D}(A^{*\frac{1}{2}}) = \mathrm{H}^1_0(\Omega)$, $P(t)z \in \mathcal{D}(A^{\frac{1}{2}}) = \mathrm{H}^1_0(\Omega)$. Therefore, for all $t \in (0,T)$, and by density of $\mathrm{H}^1_0(\Omega)$ in $\mathrm{L}^2(\Omega)$

$$\forall \varphi \in \mathrm{H}^1_0(\Omega), \forall (x,t) \in \partial\Omega\times[0,T], \quad \int_\Omega \pi(x',x,t)\varphi(x') \ \mathrm{d}x' = 0,$$
$$\Rightarrow \forall (x',x,t) \in \Omega\times\partial\Omega\times[0,T] \quad \pi(x',x,t) = 0,$$

and by symmetry

$$\forall (x', x, t) \in \partial\Omega \times \Omega \times [0, T], \quad \pi(x', x, t) = 0.$$

Third, let consider $(z_1, z_2) \in \mathcal{D}(\Omega) \times \mathcal{D}(\Omega)$, using Fubini and the boundary property of π,

$$\begin{aligned}\langle \Delta_x \pi, z_1 z_2 \rangle_{\mathcal{D}(\Omega\times\Omega)} &= -\int_\Omega \int_\Omega \nabla_x \pi(x', x, t) z_1(x') \nabla z_2(x) \ \mathrm{d}x' \ \mathrm{d}x \\ &= \int_\Omega \int_\Omega \pi(x', x, t) z_1(x') \Delta z_2(x) \ \mathrm{d}x' \ \mathrm{d}x = (\mathit{\Pi}(t) z_1, \Delta z_2)_{\mathcal{Z}} \,.\end{aligned}$$

We also find

$$\begin{aligned}\langle b \cdot \nabla_x \pi, z_1 z_2 \rangle_{\mathcal{D}(\Omega\times\Omega)} &= -\int_\Omega \int_\Omega \pi(x', x, t) z_1(x') \nabla_x \cdot (b(x) z_2(x)) \ \mathrm{d}x' \ \mathrm{d}x \\ &= -\int_\Omega \int_\Omega \pi(x', x, t) z_1(x') b(x) \cdot \nabla z_2(x) \ \mathrm{d}x' \ \mathrm{d}x \\ &= (\mathit{\Pi}(t) z_1, - b \cdot \nabla z_2)_{\mathcal{Z}} \,.\end{aligned}$$

giving

$$\big\langle b \cdot \nabla_x \pi + \Delta_x \pi, z_1 z_2 \big\rangle_{\mathcal{D}(\Omega\times\Omega)} = (\mathit{\Pi}(t) z_1, A^* z_2)_{\mathcal{Z}} \,.$$

Identically, we have

$$\big\langle b \cdot \nabla_{x'} \pi + \Delta_{x'} \pi, z_1 z_2 \big\rangle_{\mathcal{D}(\Omega\times\Omega)} = (A^* z_1, \mathit{\Pi}(t) z_2)_{\mathcal{Z}} \,.$$

Now, we recall that for all $t \in (0, T)$, $\mathit{\Pi}(t) \in \mathcal{D}(\Upsilon)$. Therefore, there exists a constant c_{st} such that

$$|\langle b \cdot \nabla_x \pi + b \cdot \nabla_{x'} \pi + \Delta_x \pi + \Delta_{x'} \pi, z_1 z_2| \leq |\, (\mathit{\Pi}(t) z_1, A^* z_2)_{\mathcal{Z}} + (A^* z_1, \mathit{\Pi}(t) z_2)_{\mathcal{Z}} \,| \leq c_{\mathrm{st}} \|z_1\|_{\mathrm{L}^2(\Omega)} \|z_2\|_{\mathrm{L}^2(\Omega)}$$

As $\pi(t) \in \mathrm{H}^1(\Omega \times \Omega)$, this implies that there exists a constant c_{st} such that

$$|\langle \Delta_x \pi + \Delta_{x'} \pi, z_1 z_2| \leq c_{\mathrm{st}} \|z_1\|_{\mathrm{L}^2(\Omega)} \|z_2\|_{\mathrm{L}^2(\Omega)}.$$

By density of $\mathcal{D}(\Omega) \otimes \mathcal{D}(\Omega)$ in $\mathcal{D}(\Omega \times \Omega)$, we conclude that $\Delta_x \pi(t) + \Delta_{x'} \pi(t) \in \mathrm{L}^2(\Omega \times \Omega)$.

Additionally, we have

$$(C\mathit{\Pi}(t) z_1, C\mathit{\Pi}(t) z_2)_{\mathcal{Y}} = \int_\Omega \int_\Omega \int_\omega \pi(t, x, x'') \pi(t, x'', x') z_1(x') z_2(x) \ \mathrm{d}x'' \ \mathrm{d}x' \ \mathrm{d}x,$$

and also

$$(B^* z_1, Q B^* z_2)_{\mathcal{U}} = \int_\Omega \int_\Omega \kappa^{-2} f(x') f(x) z_1(x') z_2(x) \ \mathrm{d}x' \ \mathrm{d}x.$$

Moreover $\mathit{\Pi} \in \mathrm{C}^1([0, T]; \mathcal{S}^+(\mathcal{Z}))$ implies that $\partial_t \pi \in \mathrm{L}^2(\Omega \times \Omega)$, with

$$\langle \partial_t \pi, z_1 z_2 \rangle_{\mathcal{D}(\Omega\times\Omega)} = \int_\Omega \int_\Omega \partial_t \pi(t, x, x'') z_1(x') z_2(x) \ \ \mathrm{d}x'' \ \mathrm{d}x' = \langle \dot{\mathit{\Pi}}(t) z_1, z_2 \rangle$$

As Π is a weak solution of Riccati in the sense of (1.19), and $\mathcal{D}(\Omega) \otimes \mathcal{D}(\Omega)$ is dense into $\mathcal{D}(\Omega \times \Omega)$, we have for all $\psi \in \mathcal{D}(\Omega \times \Omega)$

$$\int_\Omega \int_\Omega \Big[\partial_t \pi(x', x, t) - b(x') \cdot \nabla_{x'} \pi(x', x, t) - b(x) \cdot \nabla_x \pi(x', x, t) \\ - (\Delta_x + \Delta_{x'})\pi(x', x, t)\gamma \int_\omega \pi(t, x, x'')\pi(t, x'', x') \; \mathrm{d}x'' + \kappa^{-2} f(x') f(x)\Big] \psi(x, x') \; \mathrm{d}x \; \mathrm{d}x' = 0.$$

ending the justification of (2.2).

□

Remark 2.5. Such non-linear integro-differential equation associated with the Riccati kernel was already mentioned in Chapter 3, Section 5 of [39], but without justifying the necessary regularity conditions allowing to write (2.2). In [50], the regularity question was fully treated, however with different arguments and a slightly different Riccati equation.

From the kernel representation of the covariance operator, we can now specify the weak form and strong form of the optimal estimator in the context of an advection-diffusion problem. We directly find from the weak solution associated with (1.27) that

$$\big(\partial_t \hat{z}, w\big)_{\mathrm{L}^2(\Omega)} - \big(b \cdot \nabla \hat{z}, w\big)_{\mathrm{L}^2(\Omega)} - \big(\nabla \hat{z}, \nabla w\big)_{\mathrm{L}^2(\Omega)} = \int_\omega \int_\Omega \gamma \pi(x', x, t) w(x', t) \big(y^\delta(x, t) - \hat{z}(x, t)\big) \; \mathrm{d}x' \; \mathrm{d}x. \quad (2.3)$$

Then using the symmetry of Π – and of its associated kernel π – we recall that

$$\begin{aligned} \big(C\Pi(t)w, (y^\delta - C\hat{z})\big)_{\mathcal{Y}} &= \int_\omega \int_\Omega \pi(x', x, t) w(x', t) \big(y^\delta(x, t) - \hat{z}(x, t)\big) \; \mathrm{d}x' \; \mathrm{d}x \\ &= \int_\Omega \int_\omega \pi(x', x, t) \big(y^\delta(x, t) - \hat{z}(x, t)\big) w(x', t) \; \mathrm{d}x' \; \mathrm{d}x \\ &= \big(w, \Pi(t) C^* (y^\delta - C\hat{z})\big)_{\mathcal{Z}}, \end{aligned}$$

which directly gives the strong form of the Kalman estimator for our advection-diffusion example:

$$\begin{cases} \partial_t \hat{z}(x, t) - v(x) \cdot \nabla \hat{z}(x, t) - \Delta \hat{z}(x, t) = \gamma \displaystyle\int_\omega \gamma \pi(x', x, t) \big(y^\delta(x', t) - \hat{z}(x', t)\big) \; \mathrm{d}x', & x \in \Omega, t > 0, \\ \hat{z}(x, t) = 0, & x \in \partial\Omega, t > 0 \\ \hat{z}(x, 0) = \hat{z}_0(x), & x \in \Omega \end{cases} \quad (2.4)$$

Here, we recognize a nonlinear integro-differential equation that we proved to define a well-posed problem.

Remark 2.6. For clarity, we have limited the kernel representation calculations to the initial advection-diffusion example. However, an analogous representation could be performed for many other types of parabolic equations, provided that Π is a Hilbert-Schmidt operator and, furthermore, $\Pi \in \mathcal{L}(\mathcal{V}', \mathcal{V})$. This last property is justified by Theorem 1.10, unfortunately – to our knowledge – under the restrictive condition that $\mathcal{D}(A^{\frac{1}{2}}) = \mathcal{D}(A^{*\frac{1}{2}}) = \mathcal{V}$.

3. Numerical analysis

3.1. Discretization of the direct model

We consider a spatial discretization of the model (1.1) using Ritz-Galerkin method, for instance a finite element method. To be more specific with our example of advection-diffusion equation (1.7), we consider a

Lagrange finite element discretization in a finite dimensional space $\mathcal{V}^h \simeq \mathbb{R}^{N_\mathcal{Z}}$. The orthogonal projection from $\mathcal{Z}$ to $\mathcal{Z}^h$ is denoted by P^h. We now introduce the operator A^h defined by

$$\forall (u^h, v^h) \in \mathcal{Z}^h, \quad (A^h u^h, v^h)_\mathcal{Z} = a(u^h, v^h) = \langle A u^h, v^h \rangle_\mathcal{V}.$$

We proceed identically with the model error and the observations. For the sake of simplicity, we restrict the present article to the case where only the state needs to be discretized. Our discretize observation operator is hence given by

$$C^h = C P^{h*} \in \mathcal{L}(\mathcal{Y}, \mathcal{Z}^h)$$

For the model error, the same reasoning applies but as in our specific example we have $\dim(\mathcal{U})$ finite hence, there is no need to introduce a specific notation for its spatial discretization. We just introduce $B^h = P^h B \in \mathcal{Z}^{h\prime}$. Therefore, our model dynamics (1.1) is spatially discretized in $\mathcal{Z}^h \subset \mathcal{Z}$ as

$$\begin{cases} \dot{z}^h + A^h z^h = B^h \nu, & \text{in } (0,T), \\ z^h(0) = P^h \hat{z}_0 + \zeta^h, \end{cases} \tag{3.1}$$

Of note, in a very general configuration, a full numerical analysis should also take into account the observation sampling and discretization, and the model error discretization.

We now proceed to the full space-time discretization using a backward Euler time scheme. We consider the following uniform discretization of the time interval $[0,T]$: $t_k = k\Delta t$ for $k = 0, 1, \cdots, n$, where $T = n\Delta t$. We have

$$\begin{cases} \dfrac{z_{k+1}^{h,\tau} - z_k^{h,\tau}}{\tau} + A^h z_{k+1}^{h,\tau} = B^h \nu_{k+1}^\tau, & 0 \leq k \leq n-1, \\ z_0^h = P^h \hat{z}_0 + \zeta^h, \end{cases} \tag{3.2}$$

We denote

$$\Phi_n^{h,\tau} = (\mathbb{1}_{\mathcal{Z}^h} + \tau A^h)^{-n}$$

which is clearly invertible as A^h is maximal monotone. Moreover, we recall the following estimate ([25], Thm. 2.7)

$$\forall z_0 \in \mathcal{Z}, \quad \|\Phi(t_n) z_0 - \Phi_n^{h,\tau} z_0\|_\mathcal{Z} \leq c_{\mathrm{st}} \frac{h^2 + \tau}{t_n} \|z_0\|_\mathcal{Z}, \quad n \geq 0. \tag{3.3}$$

Finally, we also introduce $\hat{z}_0^h = P^h \hat{z}_0$ and

$$B^{h,\tau} = \tau(\mathbb{1} + \tau A^h)^{-1} B^h = \tau \Phi_1^{h,\tau} B^h,$$

such that the discrete-time dynamics (3.2) rewrites

$$\begin{cases} z_{k+1}^{h,\tau} = \Phi_1^{h,\tau} z_k^{h,\tau} + B^{h,\tau} \nu_{k+1}^\tau, & 0 \leq k \leq n-1, \\ z_0^h = \hat{z}_0^h + \zeta^h, \end{cases} \tag{3.4}$$

3.2. Discretization of the estimator

Using a quadrature rule to approximate the integrals in (1.8), we obtain the following time-discretized criterion:

$$\begin{aligned}\mathscr{J}_n^{h,\tau}(\zeta^h,(\nu_k^\tau)_{1\leq k\leq n}) = \frac{1}{2}\big(\zeta^h,(P^h\varPi_0P^{h*})^{-1}\zeta^h\big)_{\mathcal{Z}}+\frac{1}{2}\sum_{k=1}^{n}\kappa^2\tau\|\nu_k^\tau\|^2_{\mathcal{U}}& \\ +\frac{1}{2}\sum_{k=0}^{n-1}\tau\gamma\|y_k^\delta-C^h z^{h,\tau}_{k|\zeta^h,(\nu_k^\tau)_{0\leq k\leq n}}\|^2_{\mathcal{Y}}.&\end{aligned} \tag{3.5}$$

We denote by

$$z^{h,\tau}_{k|n}=z^{h,\tau}_{k|\zeta,(\nu_k^\tau)_{1\leq k\leq n}},\qquad z^{h,\tau}_{|n}=(z^{h,\tau}_{k|n})_{1\leq k\leq n},\qquad \nu^\tau_{|n}=(\nu_k^\tau)_{1\leq k\leq n},$$

and by $(\bar{\zeta}^{h,\tau}_{|n},\bar{\nu}^{h,\tau}_{|n})$ the minimizer of the full-discretized criterion $\mathscr{J}_n^{h,\tau}$. This minimizer generates a trajectory $\bar{z}^{h,\tau}_{|n}$. As in the continuous case, we can easily prove the following theorem

Theorem 3.1. *The minimizer* $(\bar{\zeta}^{h,\tau}_{|n},\bar{\nu}^{h,\tau}_{|n})$ *of* $\mathscr{J}_n^{h,\tau}$ *satisfies*

$$\bar{\zeta}^{h,\tau}_{|n}=\varPi_0^h\bar{q}^{h,\tau}_{0|n},\qquad \bar{\nu}_{k|n}=Q^\tau B^{h,\tau*}\bar{q}^{h,\tau}_{k|n},\ k=1,\cdots,n,$$

where $Q^\tau=\kappa^{-2}\tau^{-1}\mathbb{1}_{\mathcal{U}}$, $R^{h,\tau}=\gamma\tau\mathbb{1}_{\mathcal{Y}^h}$, $\varPi_0^h=P^h\varPi_0P^{h*}$ *and*

$$\begin{cases}\bar{z}^{h,\tau}_{k+1|n}=\varPhi_1^{h,\tau}\bar{z}^{h,\tau}_{k|n}+B^{h,\tau}Q^\tau B^{h,\tau*}\bar{q}^{h,\tau}_{k|n}, & 0\leq k\leq n-1, \qquad (3.6a)\\ \bar{z}^{h,\tau}_{0|n}=\hat{z}_0^h+\varPi_0^h\bar{q}^{h,\tau}_{0|n}, & \qquad (3.6b)\\ \bar{q}^{h,\tau}_{k|n}=\varPhi_1^{h,\tau*}\bar{q}^{h,\tau}_{k+1|n}+C^{h*}R^{h,\tau}\big(y_k^\delta-C^h\bar{z}^{h,\tau}_{k|n}\big), & 0\leq k\leq n-1, \qquad (3.6c)\\ \bar{q}^{h,\tau}_{n|n}=0. & \qquad (3.6d)\end{cases}$$

Proof. We recall that we are in the case of the minimization of strictly convex quadratic functional in finite dimension under a linear discrete-time dynamics constraint, hence we have one and only one minimizer. We then introduce the following Lagrangian

$$\begin{aligned}\mathscr{L}_n^{h,\tau}(z^{h,\tau}_{|n},q^{h,\tau}_{|n},\nu^\tau_{|n}) = \frac{1}{2}\big(z^{h,\tau}_{0|n},(\varPi_0^h)^{-1}z^{h,\tau}_{0|n}\big)_{\mathcal{Z}}+\frac{1}{2}\sum_{k=1}^{n}\kappa^2\tau\|\nu_k^\tau\|^2_{\mathcal{U}}& \\ +\frac{1}{2}\sum_{k=0}^{n-1}\tau\gamma\|y_k^\delta-C^h z^{h,\tau}_{k|n}\|^2_{\mathcal{Y}}& \\ +\sum_{k=0}^{n-1}\big(q^{h,\tau}_{k+1|n},z^{h,\tau}_{k+1}-\varPhi_1^{h,\tau}z^{h,\tau}_k-B^{h,\tau}\nu^\tau_{k+1}\big)_{\mathcal{Z}}&\end{aligned} \tag{3.7}$$

and minimizing $\mathscr{J}_n^{h,\tau}$ in the finite dimensional Hilbert space $\mathcal{Z}^h\times\mathcal{U}^n$ under the constraint of the discrete-time dynamics (3.4), is equivalent to find the saddle point of $\mathscr{L}_n^{h,\tau}$. Derivating $\mathscr{L}_n^{h,\tau}$, we obtain in particular for $0 < k < n$,

$$\forall \xi \in \mathcal{Z}^h, \quad \big\langle \mathrm{D}_{z^{h,\tau}_{k|n}} \mathscr{L}^{h,\tau}_n(z^{h,\tau}_{|n}, q^{h,\tau}_{|n}, \nu^{\tau}_{|n}), \xi \big\rangle_{\mathcal{Z}^h} = -\tau\gamma\big(y^\delta_k - C^h z^{h,\tau}_{k|n}, C^h \xi\big)_{\mathcal{Y}} + \big(q^{h,\tau}_{k|n}, \xi\big)_{\mathcal{Z}} - \big(q^{h,\tau}_{k+1|n}, \Phi^{h,\tau}_1 \xi\big)_{\mathcal{Z}}.$$

When applied to the stationary point $(\bar z^{h,\tau}_{|n}, \bar q^{h,\tau}_{|n}, \bar \nu^{\tau}_{|n})$, we get (3.6c) for $0 < k < n-1$, while for $k = n$,

$$\forall \xi \in \mathcal{Z}^h, \quad \big\langle \mathrm{D}_{z^{h,\tau}_{n|n}} \mathscr{L}^{h,\tau}_n(z^{h,\tau}_{|n}, q^{h,\tau}_{|n}, \nu^{\tau}_{|n}), \xi \big\rangle_{\mathcal{Z}^h} = \big(q^{h,\tau}_{n|n}, \xi\big)_{\mathcal{Z}},$$

leads to (3.6d). And for $k = 0$,

$$\forall \xi \in \mathcal{Z}^h, \quad \big\langle \mathrm{D}_{z^{h,\tau}_{0|n}} \mathscr{L}^{h,\tau}_n(z^{h,\tau}_{|n}, q^{h,\tau}_{|n}, \nu^{\tau}_{|n}), \xi \big\rangle = \big(z^{h,\tau}_{0|n}, (\Pi^h_0)^{-1}\xi\big)^2_{\mathcal{Z}} - \tau\gamma\big(y^{\delta,h}_0 - C^h z^{h,\tau}_{0|n}, C^h \xi\big)_{\mathcal{Y}} - \big(q^{h,\tau}_{1|n}, \Phi^{h,\tau}_1 \xi\big)_{\mathcal{Z}},$$

gives (3.6b) for the stationary point as soon as we extend the Lagrange multiplier definition up to $\bar q^{h,\tau}_{0|n}$ in (3.6c). Finally, we have

$$\forall \mu \in \mathcal{U}, \quad \big\langle \mathrm{D}_{\nu^{h,\tau}_{k|n}} \mathscr{L}^{h,\tau}_n(z^{h,\tau}_{|n}, q^{h,\tau}_{|n}, \nu^{\tau}_{|n}), \mu \big\rangle_{\mathcal{U}} = \big(\nu^{h,\tau}_{k|n}, \mu\big)_{\mathcal{U}} - \big(B^{h,\tau *} q^{h,\tau}_{k+1|n}, \mu\big)_{\mathcal{U}},$$

which coupled to

$$\forall \lambda \in \mathcal{Z}^h, \quad \big\langle \mathrm{D}_{q^{h,\tau}_{k+1|n}} \mathscr{L}^{h,\tau}_n(z^{h,\tau}_{|n}, q^{h,\tau}_{|n}, \nu^{\tau}_{|n}), \lambda \big\rangle_{\mathcal{Z}} = \big(z^{h,\tau}_{k+1} - \Phi^{h,\tau}_1 z^{h,\tau}_k - B^{h,\tau} \nu^\tau_{k+1}, \lambda\big)_{\mathcal{Z}},$$

gives (3.6a) at the stationary point. □

The analogous version of Theorems 1.13–1.14 in the fully discrete case is stated as follows:

Theorem 3.2. *We have the following identity:*

$$\bar z^{h,\tau}_{k|n} = \hat z^{h,\tau-}_k + \Pi^{h,\tau-}_k \bar q^{h,\tau}_{k|n}, \quad \forall 0 \le k \le n+1, \tag{3.8}$$

where $(\hat z^{h,\tau-}_k, \Pi^{h,k-}_k)_{k\ge 0}$ *are defined by*

$$\begin{cases} \hat z^{h,\tau-}_0 = \hat z^h_0, & & \text{(3.9a)} \\ \hat z^{h,\tau+}_n = \hat z^{h,\tau-}_n + \Pi^{h,\tau+}_n C^{h*} R^{h,\tau} \Big(y^\delta_n - C^h \hat z^{h,\tau-}_n\Big), & n \ge 0, & \text{(3.9b)} \\ \hat z^{h,\tau-}_{n+1} = \Phi^{h,\tau}_1 \hat z^{h,\tau+}_n, & n \ge 0, & \text{(3.9c)} \end{cases}$$

and

$$\begin{cases} \Pi^{h,\tau-}_0 = \Pi^h_0, & & \text{(3.10a)} \\ \Pi^{h,\tau+}_n = \big[(\Pi^{h,\tau-}_n)^{-1} + C^{h*} R^{h,\tau} C^h\big]^{-1}, & n \ge 0, & \text{(3.10b)} \\ \Pi^{h,\tau-}_{n+1} = \Phi^{h,\tau}_1 \Pi^{h,\tau+}_n \Phi^{h,\tau*}_1 + B^{h,\tau} Q^\tau B^{h,\tau*}, & n \ge 0. & \text{(3.10c)} \end{cases}$$

We recall that $\Phi^{h,\tau}_1$ is invertible, hence $\Pi^{h,\tau-}_{n+1}$ is positive definite if $\Pi^{h,\tau+}_n$ is. This recursively ensures that $\Pi^{h,\tau-}_n$ is well defined and positive definite for all n.

The discrete-time dynamics (3.9) defines the discrete-time optimal sequential estimator as we clearly follow the definition (3.14) from the identity (3.8). The equations of (3.9) constitute what is called the *prediction/correction algorithm* for the discrete-time Kalman estimator as initially developed by R.E. Kalman in his seminal paper [31]: (3.9) consists in the step of *prediction* of the observer and filter (3.10b), which is followed (or preceded) by the correction step given by (3.9b). This *prediction/correction algorithm* can be seen as a splitting algorithm of the continuous-time dynamics (1.27). Indeed, we find that $\hat{z}_{n+1}^{h,\tau+}$ follows the one-step time scheme.

$$\hat{z}_{n+1}^{h,\tau+} = \varPhi_1^{h,\tau} \hat{z}_n^{h,\tau+} + \varPi_{n+1}^{h,\tau+} C^{h*} R^{h,\tau} \Big(y_{n+1}^{\delta,h} - C^h \varPhi_1^{h,\tau} \hat{z}_n^{h,\tau+} \Big), \tag{3.11}$$

which is a consistent time scheme of order 1 in τ of (1.27), as soon as the discrete covariance operator converges to the continuous covariance operator as it will be proved in Section 3.3.

We also want to point out that a simple computation gives for all $n > 0$

$$\begin{aligned} G_n^{h,\tau} &= \varPi_n^{h,\tau+} C^{h*} R_n^{h,\tau} \\ &= \varPi_n^{h,\tau+} C^{h*} R^h (C^h \varPi_n^{h,\tau-} C^{h*} + W^{h,\tau})(C^h \varPi_n^{h,\tau-} C^{h*} + W^{h,\tau})^{-1} \\ &= \varPi_n^{h,\tau+} (C^{h*} R^{h,\tau} C^h + (\varPi^{h,\tau-})^{-1}) \varPi^{h,\tau-} C^{h*} (C^h \varPi_n^{h,\tau-} C^{h*} + W^{h,\tau})^{-1} \\ &= \varPi_n^{h,\tau-} C^{h*} \left(C^h \varPi_n^{h,\tau-} C^{h*} + W_n^{h,\tau} \right)^{-1}, \end{aligned} \tag{3.12}$$

An alternative formula of the Kalman gain. Finally, note that from the Sherman-Morrison-Woodbury identity, we also have an alternative formula for $\varPi_{n+1}^{h,\tau-}$ where the inversions are performed in the observation space, namely for all $n > 0$

$$\varPi_{n+1}^{h,\tau+} = \varPi_{n+1}^{h,\tau-} - \varPi_{n+1}^{h,\tau-} C^{h*} \left[C^h \varPi_{n+1}^{h,\tau-} C^{h*} + W^{h,\tau} \right]^{-1} C^{h,\tau} \varPi_{n+1}^{h,\tau-}. \tag{3.13}$$

Proof of Theorem 3.2. We proceed by induction:

- We have from (3.6b),

  $$\bar{z}_{0|n}^{h,\tau} = \hat{z}_0^h + \varPi_0^h \bar{q}_{0|n}^{h,\tau} = \hat{z}_0^{h,\tau-} + \varPi_0^h \bar{q}_{0|n}^{h,\tau},$$

  which is exactly the formula (3.8) for $k = 0$.
- Let us assume that the formula (3.8) is satisfied for all $i \in \{0, 1, \dots, k\}$. Since

  $$\bar{z}_{k+1|n}^{h,\tau} = \varPhi_1^{h,\tau} \bar{z}_{k|n}^{h,\tau} + B^{h,\tau} Q^\tau B^{h,\tau*} \bar{q}_{k|n}^{h,\tau},$$

  we deduce

  $$\bar{z}_{k+1|n}^{h,\tau} = \varPhi_1^{h,\tau} \left(\hat{z}_k^{h,\tau-} + \varPi_k^{h,\tau-} \bar{q}_{k|n}^{h,\tau} \right) + B^{h,\tau} Q^\tau B^{h,\tau*} \bar{q}_{k|n}^{h,\tau}.$$

  Taking into account the equations giving $\hat{z}_k^{h,\tau+}$ and $\varPi_k^{h,\tau+}$ from $\hat{z}_k^{h,\tau-}$ and $\varPi_k^{h,\tau-}$ respectively, we obtain

  $$\begin{aligned} \bar{z}_{k+1|n}^{h,\tau} = \varPhi_1^{h,\tau} & \left(\hat{z}_k^{h,\tau+} - G_k^{h,\tau} \big(y_k^\delta - C \hat{z}_k^{h,\tau-} \big) \right) \\ & + \varPhi_1^{h,\tau} \left(\varPi_k^{h,\tau+} + G_k^{h,\tau} C^h \varPi_k^{h,\tau-} \right) \bar{q}_{k|n+1}^{h,\tau} + B^{h,\tau} Q^\tau B^{h,\tau*} \bar{q}_{k+1|n}^{h,\tau} \\ = \varPhi_1^{h,\tau} & \hat{z}_k^+ - \varPhi_1^{h,\tau} G_k^{h,\tau} \left(y_k^\delta - C^h \big(\hat{z}_k^{h,\tau-} + \varPi_k^{h,\tau-} \bar{q}_{k|n}^{h,\tau} \big) \right) \\ & \qquad + \varPhi_1^{h,\tau} \varPi_k^{h,\tau+} \bar{q}_{k|n}^{h,\tau} + B^{h,\tau} Q^\tau B^{h,\tau*} \bar{q}_{k+1|n}^{h,\tau}. \end{aligned}$$

We deduce from the hypothesis on k,

$$\bar{z}^{h,\tau}_{k+1|n} = \varPhi^{h,\tau}_1 \hat{z}^{h,\tau+}_k - \varPhi^{h,\tau}_1 G^{h,\tau}_k \left(y^\delta_k - C^h \bar{z}^{h,\tau}_{k|n} \right) + \varPhi^{h,\tau}_1 \varPi^{h,\tau+}_k \bar{q}^{h,\tau}_{k|n} + B^{h,\tau} Q^\tau B^{h,\tau*} \bar{q}^{h,\tau}_{k+1|n}.$$

The equation (3.10b) giving $\hat{z}^{h,\tau-}_{k+1}$ from $\hat{z}^{h,\tau+}_k$ implies that

$$\bar{z}^{h,\tau}_{k+1|n} = \hat{z}^{h,\tau-}_{k+1} - \varPhi^{h,\tau}_1 G^{h,\tau}_k \left(y^\delta_k - C^{h,\tau} \bar{z}^{h,\tau}_{k|n} \right) + \varPhi^{h,\tau}_1 \varPi^{h,\tau+}_k \bar{q}^{h,\tau}_{k|n} + B^{h,\tau} Q^\tau B^{h,\tau*} \bar{q}^{h,\tau}_{k+1|n}.$$

Replacing $\bar{q}^{h,\tau}_{k|n}$ using (3.6c), we obtain

$$\begin{aligned}
\bar{z}^{h,\tau}_{k+1|n} &= \hat{z}^{h,\tau-}_{k+1} - \varPhi^{h,\tau}_1 G^{h,\tau}_k \left(y^\delta_k - C^h \bar{z}^{h,\tau}_{k|n} \right) \\
&\quad + \varPhi^{h,\tau}_1 \varPi^{h,\tau+}_k \left(\varPhi^{h,\tau*}_1 \bar{q}^{h,\tau}_{k+1|n} + C^{h*} R^{h,\tau} \left(y^\delta_k - C^{h,\tau} \bar{z}^{h,\tau}_{k|n} \right) \right) + B^{h,\tau} Q^\tau B^{h,\tau*} \bar{q}^{h,\tau}_{k+1|n} \\
&= \hat{z}^{h,\tau-}_{k+1} - \varPhi^{h,\tau}_1 \left(G^{h,\tau}_k - \varPi^{h,\tau+}_k C^{h*} R^{h,\tau} \right) \left(y^\delta_k - C^h \bar{z}^{h,\tau}_{k|n} \right) \\
&\qquad + \left(\varPhi^{h,\tau}_1 \varPi^{h,\tau+}_k \varPhi^{h,\tau*}_1 + B^{h,\tau} Q^\tau B^{h,\tau*} \right) \bar{q}^{h,\tau}_{k+1|n}.
\end{aligned}$$

Using (3.10c) defining $\varPi^{h,\tau-}_{k+1}$ from $\varPi^{h,\tau+}_k$, we deduce that

$$\bar{z}^{h,\tau}_{k+1|n} = \hat{z}^{h,\tau-}_{k+1} + \varPi^{h,\tau-}_{k+1} \bar{q}^{h,\tau}_{k+1|n} - \varPhi^{h,\tau}_1 \left(G^{h,\tau}_k - \varPi^{h,\tau+}_k C^{h*} R^{h,\tau}_k \right) \left(y^\delta_k - C^h \bar{z}^{h,\tau}_{k|n} \right).$$

Therefore, we have proved that $\bar{z}^{h,\tau}_{k+1|n} = \hat{z}^{h,\tau-}_{k+1} + \varPi^{h,\tau-}_{k+1} \bar{q}^{h,\tau}_{k+1|n}$, namely the formula (3.8) is satisfied at step $k+1$, which ends the proof by induction.

□

From Theorem 3.2, we see that the discrete-time Kalman filter is exactly the discrete-time optimal sequential estimator in the sense of the following definition.

Definition 3.3. The discrete-time optimal sequential estimator is defined by

$$\forall n \in \mathbb{N}, \quad \hat{z}^{h,\tau}_n = \bar{z}^{h,\tau}_{n|n}. \tag{3.14}$$

This definition then allows to justify the discrete-time Kalman filter optimality in a purely deterministic framework, far from the original stochastic setting envisioned by R.E. Kalman in his seminal work [31].

3.3. Space-time convergence analysis

From the prediction-correction splitting time scheme (3.10) we can deduce a *one-step* recursive formula for the discretized covariance operator $\varPi^{h,\tau-}_n$. Indeed by combining (3.13) and (3.12), we have for all $n \in \mathbb{N}$

$$\varPi^{h,\tau-}_{n+1} = \varPhi^{h,\tau}_1 \varPi^{h,\tau-}_n \varPhi^{h,\tau*}_1 + B^{h,\tau} Q^\tau B^{h,\tau*} - \varPi^{h,\tau+}_n C^{h*} R^{h,\tau} C^h \varPi^{h,\tau-}_n, \tag{3.15}$$

which ensures that the discretized covariance operator $\varPi^{h,\tau+}_n$ satisfied a discrete-time version of (1.18) with for all $n \in \mathbb{N}$

$$\varPi^{h,\tau-}_n = \varPhi^{h,\tau}_n \varPi^{h,\tau}_0 \varPhi^{h,\tau*}_n + \sum_{k=0}^{n-1} \varPhi^{h,\tau}_k B^{h,\tau} Q^\tau B^{h,\tau*} \varPhi^{h,\tau*}_k - \sum_{k=0}^{n-1} \varPhi^{h,\tau}_{n-k} \varPi^{h,\tau+}_k C^{h*} R^{h,\tau} C^h \varPi^{h,\tau-}_k \varPhi^{h,\tau*}_{n-k}. \tag{3.16}$$

In fact, (3.16) mixes $\Pi_n^{h,\tau+}$ and $\Pi_n^{h,\tau-}$. However, using (3.9c) or (3.13), we will prove – see the next proposition proof – that

$$\forall n > 0, \quad \|\Pi_k^{h,\tau+} - \Pi_k^{h,\tau-}\| = O(h).$$

Moreover, a variant of (3.16), namely

$$\Pi_n^{h,\tau-} = \Phi_n^{h,\tau} \Pi_0^{h,\tau} \Phi_n^{h,\tau*} + \sum_{k=0}^{n-1} \Phi_k^{h,\tau} B^{h,\tau} Q^\tau B^{h,\tau*} \Phi_k^{h,\tau*} \\ - \sum_{k=0}^{n-1} \Phi_{n-k}^{h,\tau} \Pi_k^{h,\tau+} C^{h*} [R^{h,\tau} C^h \Pi_n^{h,\tau-} C^{h*} R^{h,\tau} + R^{h,\tau}] C^h \Pi_k^{h,\tau+} \Phi_{n-k}^{h,\tau*}, \quad (3.17)$$

recursively ensures that $\Pi_n^{h,\tau-}$ satisfies

$$\forall n > 0, \quad 0 < \Pi_n^{h,\tau+} \leq \|\Pi_0^{h,\tau}\| + T\|B^{h,\tau}\| \quad (3.18)$$

and identically from (3.9c), $\Pi_n^{h,\tau+} > 0$

We can now propose a full space-time numerical analysis of the Riccati solution, in the spirit of what was done in [26] for a space discretization only.

Theorem 3.4. *Let Π be the solution of Problem (1.18) in $\mathrm{C}^1([0,T], \mathcal{J}_2(\mathcal{Z}))$, and $(\Pi_n^{h,\tau-})_{n\in\mathbb{N}} \in (\mathcal{J}_2(\mathcal{Z}^h))^{\mathbb{N}}$ the solution of (3.16). We have*

$$\lim_{h,\tau\to 0} \sup_{k\in[0,n]} \|\Pi_k^{h,\tau-} - P^h \Pi(t_k) P^{h*}\|_2 = 0.$$

Proof. Our proof is inspired by Theorem 3 of [26], with the additional treatment of the time discretization. We first compute

$$\|\Pi_n^{h,\tau-} - P^h \Pi(t_n) P^{h*}\|_2 = \Bigg\| \Phi_n^{h,\tau} \Pi_0^{h,\tau} \Phi_n^{h,\tau*} + \sum_{k=0}^{n-1} \Phi_k^{h,\tau} B^{h,\tau} Q^\tau B^{h,\tau*} \Phi_k^{h,\tau*} \\ - \sum_{k=0}^{n-1} \Phi_{n-k}^{h,\tau} \Pi_k^{h,\tau+} C^{h*} R^{h,\tau} C^h \Pi_k^{h,\tau-} \Phi_{n-k}^{h,\tau*} \\ - P^h \Phi(t_n) \Pi_0 \Phi^*(t_n) P^{h*} - \int_0^{t_n} P^h \Phi(s) B Q B^* \Phi^*(s) P^{h*} \ \mathrm{d}s \\ + \int_0^{t_n} P^h \Phi(t_n - s) \Pi(s) C^* R C \Pi(s) \Phi^*(t_n - s) P^{h*} \ \mathrm{d}s \Bigg\|_2.$$

Therefore, we have

$$\|\Pi_n^{h,\tau-} - P^h \Pi(t_n) P^{h*}\|_2 \leq e_{\mathrm{init}}^h + e_{\mathrm{quad}}^\tau + \sum_{k=0}^{n-1} [e_{k,\mathrm{obs}}^{h,\tau} + e_{k,\mathrm{err}}^{h,\tau}]$$

where

$$e_{\mathrm{init}}^h = \big\|\Phi_n^{h,\tau} \Pi_0^{h,\tau} \Phi_n^{h,\tau*} - P^h \Phi(t_n) \Pi_0 \Phi^*(t_n) P^{h*}\big\|_2,$$

and

$$\begin{aligned}e^{\tau}_{\text{quad}} = \Bigg\| \int_0^{t_n} P^h \varPhi(t_n - s)\varPi(s) C^* R C \varPi(s) \varPhi^*(t_n - s) P^{h*} \;\, \mathrm{d}s - \int_0^{t_n} P^h \varPhi(s) B Q B^* \varPhi^*(s) P^{h*} \;\, \mathrm{d}s \\ - \sum_{k=0}^{n} \tau P^h \varPhi(t_n - t_k) \varPi(t_k) C^* R C \varPi(t_k) \varPhi^*(t_n - t_k) P^{h*} - \sum_{k=1}^{n} \tau P^h \varPhi(t_k) B Q B^* \varPhi^*(t_k) P^{h*} \Bigg\|_2,\end{aligned}$$

are combined with a summation of terms of the form

$$e^{h,\tau}_{k,\text{obs}} = \big\| \varPhi^{h,\tau}_{n-k} \varPi^{h,\tau+}_k C^{h*} R^{h,\tau} C^h \varPi^{h,\tau-}_k \varPhi^{h,\tau*}_{n-k} - \tau P^h \varPhi(t_n - t_k) \varPi(t_k) C^* R C \varPi(t_k) \varPhi^*(t_n - t_k) P^{h*} \big\|_2,$$

and

$$e^{h,\tau}_{k,\text{err}} = \big\| \varPhi^{h,\tau}_k B^{h,\tau} Q^{\tau} B^{h,\tau*} \varPhi^{h,\tau*}_k - \tau P^h \varPhi(t_{k+1}) B Q B^* \varPhi^*(t_{k+1}) P^{h*} \big\|_2.$$

We now proceed to the estimation of each term. Using Lemma 3 of [26], we first have

$$\lim_{\substack{h\to 0\\ \tau \to 0}} e^{h,\tau}_{\text{init}} = \lim_{\substack{h\to 0\\ \tau \to 0}} \big\| \varPhi^{h,\tau}_n [P^h \varPi_0 P^{h*}] \varPhi^{h,\tau*}_n - P^h[\varPhi(t_n) \varPi_0 \varPhi^*(t_n)] P^{h*} \big\|_2 = 0.$$

Identically, we show from $B^{h,\tau} = \tau \varPhi^{h,\tau}_1 P^h B$ and $Q^{\tau} = \tau^{-1}\kappa^{-2} = \tau^{-1} Q$ that

$$\lim_{\substack{h\to 0\\ \tau \to 0}} \tau^{-1} e^{h,\tau}_{k,\text{err}} = \lim_{\substack{h\to 0\\ \tau \to 0}} \big\| \varPhi^{h,\tau}_{k+1} [P^h B Q B^* P^{h*}] \varPhi^{h,\tau*}_{k+1} - P^h [\varPhi(t_{k+1}) B Q B^* \varPhi^*(t_{k+1})] P^{h*} \big\|_2 = 0.$$

The term associated with the observation operator is more intricate. First we decompose

$$\begin{aligned}e^{h,\tau}_{k,\text{obs}} \leq \big\| \varPhi^{h,\tau}_{n-k} \varPi^{h,\tau+}_k C^{h*} R^{\tau,h} C^h \varPi^{h,\tau+}_k \varPhi^{h,\tau*}_{n-k} \\ - \tau P^h \varPhi(t_n - t_k) \varPi(t_k) C^* R C \varPi(t_k) \varPhi^*(t_n - t_k) P^{h*} \big\|_2 \\ + \big\| \varPhi^{h,\tau}_{n-k} \varPi^{h,\tau+}_k C^{h*} \big\{ R^{h,\tau} C^h \varPi^{h,\tau-}_n C^{h*} R^{h,\tau} \big\} C^h \varPi^{h,\tau+}_k \varPhi^{h,\tau*}_{n-k} \big\|_2,\end{aligned}$$

Then, we remark that

$$\| R^{h,\tau} C^h \varPi^{h,\tau-}_n C^{h*} R^{h,\tau} \| = O(\tau^2).$$

We obtain

$$\begin{aligned}\lim_{\substack{h\to 0\\ \tau \to 0}} \tau^{-1} e^{h,\tau}_{k,\text{obs}} = \lim_{\substack{h\to 0\\ \tau \to 0}} \big\| \varPhi^{h,\tau}_{n-k} \varPi^{h,\tau-}_k P^{h*} C^* R C P^h \varPi^{h,\tau-}_k \varPhi^{h,\tau*}_{n-k} \\ - P^h \varPhi(t_n - t_k) \varPi(t_k) C^* R C \varPi(t_k) \varPhi^*(t_n - t_k) P^{h*} \big\|_2.\end{aligned}$$

Following the exact same computation as in Proof of Theorem 3 of [26] we obtain

$$\begin{aligned}\big\|\Phi^{h,\tau}_{n-k}\Pi^{h,\tau-}_{k}&P^h C^* R C P^{h*}\Pi^{h,\tau-}_{k}\Phi^{h,\tau*}_{n-k}\\&\quad - P^h\Phi(t_n-t_k)\Pi(t_k)C^*RC\Pi(t_k)\Phi^*(t_n-t_k)P^{h*}\big\|_2\\&\leq c^1_{\mathrm{st}}\big\|\Pi^{h,\tau-}_k - P^h\Pi(t_k)P^h\big\|_2 + c^2_{\mathrm{st}}\big\|P^{h*}P^h\Pi(t_k)-\Pi(t_k)\big\|_2\\&\quad + c^3_{\mathrm{st}}\big\|\Phi^{h,\tau}_{n-k}[P^h\Pi(t_k)C^*RC\Pi(t_k)P^h]\Phi^{h,\tau*}_{n-k}\\&\qquad\qquad - P^h[\Phi(t_n-t_k)\Pi(t_k)C^*RC\Pi(t_k)\Phi(t_n-t_k)]P^h\big\|_2\end{aligned}$$

Again, the last term tends to 0 from Lemma 3 of [26], whereas the second term goes to 0 with h.

Finally, let us analyze the error e^τ_{quad} introduced by replacing the integral form with a quadrature rule. We introduce

$$\Gamma_t : [0,t]\ni s\mapsto \Phi(t-s)\Pi(s)C^*RC\Pi(s)\Phi^*(t-s),$$

which belongs to $\mathrm{C}^0([0,t],\mathcal{S}^+)$. Indeed, it is the mild solution of

$$\begin{cases}\dot\Gamma_t(s) + A\Gamma(s) + \Gamma A^*(s) + E_t(s) = 0, & s\in(0,t)\\ \Gamma_t(0) = \Phi(t)\Pi_0 C^* S C\Pi_0\Phi^*(t)\end{cases}$$

with

$$E_t(s) = \Phi(t-s)\Xi(t)C^*SC\Pi(t)\Phi(t-s) + \Phi(t-s)\Pi(t)C^*SC\Xi(t)\Phi(t-s)$$

and $\Xi = \dot\Pi\in\mathrm{C}^0([0,T],\mathcal{S}^+)$ since in our case $\Pi\in\mathrm{C}^1([0,T],\mathcal{S}^+)$. Furthermore, taking $(z_1,z_2)\in\mathcal{Z}$, we have

$$\begin{aligned}\upsilon_{\Gamma(0)}(z_1,z_2) &= (\Gamma(0)z_1, A^*z_2)_{\mathcal{Z}} + (A^*z_1,\Gamma(0)z_2)_{\mathcal{Z}}\\&= (C\Pi_0\Phi^*(t)z_1, C\Pi_0\Phi^*(t)A^*z_2)_{\mathcal{Y}} + (C\Pi_0\Phi^*(t)A^*z_1, C\Pi_0\Phi^*(t)z_2)_{\mathcal{Y}}\\&= (C\Pi_0\Phi^*(t)z_1, C\Pi_0A^*\Phi^*(t)z_2)_{\mathcal{Y}} + (C\Pi_0A^*\Phi^*(t)z_1, C\Pi_0\Phi^*(t)z_2)_{\mathcal{Y}}\end{aligned}$$

We have chosen $\Pi_0\in\mathcal{D}(\Upsilon)$, hence υ_{Π_0} is bounded and

$$\big|\upsilon_{\Gamma(0)}(z_1,z_2)\big| \leq \|C\|^2\|\upsilon_{\Pi_0}\|\|\Phi(t)\|^2\|z_1\|_{\mathcal{Z}}\|z_2\|_{\mathcal{Z}},$$

hence, $\Gamma(0)\in\mathcal{D}(\Upsilon)$ which implies that $\Gamma_t\in\mathrm{C}^1([0,t],\mathcal{S}^+)$. From Peano's Kernel Theorem [22],

$$\left\|\int_0^{t_n}\Gamma_{t_n}(s)\ \mathrm{d}s - \sum_{k=1}^{n}\tau\Gamma_{t_n}(t_{k+1})\right\| \leq \frac{\tau^2}{2}\sup_{s\in[0,t_n]}\|\dot\Gamma_{t_n}(s)\|.$$

Therefore, we find that

$$\lim_{\substack{h\to 0\\ \tau\to 0}} e^\tau_{\mathrm{quad}} = 0.$$

Combining all the estimations, we finally get that there exists $\epsilon^{h,\tau}$ satisfying $\lim_{h,\tau\to 0}\epsilon^{h,\tau}=0$ and a constant $c_{\rm st}$ such that

$$\|\mathit{\Pi}_n^{h,\tau-}-P^h\mathit{\Pi}(t_n)P^{h*}\|_2\leq\epsilon^{h,\tau}+c_{\rm st}\sum_{k=1}^{n}\tau\|\mathit{\Pi}_k^{h,\tau+}-P^h\mathit{\Pi}(t_k)P^{h*}\|_2,$$

which yields the theorem by Gronwall's inequality. □

Note that convergence analysis of Riccati problems has been widely studied, in particular when considering space discretization [34], even with more general unbounded observation and control operators. Our result, in the wave of [14, 26], gives convergence in the class of Hilbert-Schmidt operators. This will validate our choice of numerical tools, aka $\mathcal{H}$-matrices, adapted to such operators with kernel. Moreover, our choice of time discretization corresponds to the discrete-time Kalman filter hence fills the gap between the deterministic approach and the Kalman filter as it is understood in a stochastic framework, when the time-sampling is studied, see for instance the recent analysis in [1].

4. H-matrix based algorithm

4.1. Matrix-based algorithm

We consider a spatial discretization using $\mathbb{P}_k$ Lagrange finite elements and denotes by $(\varphi_i)_{1\leq i\leq {\rm N}_z}$ the associated basis functions where ${\rm N}_z=\dim(\mathcal{V}^h)$. Typically for a given $v^h\in\mathcal{V}^h$, we associate the vector of degrees of freedom in ${\rm v}^h\in\mathbb{R}^{{\rm N}_z}$ such that

$$v^h\sim{\rm v}^h=\begin{pmatrix}{\rm v}_1\\ \vdots\\ {\rm v}_{{\rm N}_z}\end{pmatrix}\Leftrightarrow v^h(x)=\sum_{i=1}^{{\rm N}_z}{\rm v}_i^h\varphi_i(x),\quad x\in\Omega.$$

Note that here, we consider only the degrees of freedom associated with an actual unknown of the problem, namely after elimination of the Dirichlet conditions for instance. We then define for all $\mathcal{V}^h\times\mathcal{V}^h\ni(v^h,u^h)\sim({\rm v}^h,{\rm u}^h)\in\mathbb{R}^{{\rm N}_z}\times\mathbb{R}^{{\rm N}_z}$,

$${\rm v}^{h\intercal}{\rm M}^h{\rm u}^h=(v^h,u^h)_{\mathcal{Z}},\quad {\rm v}^{h\intercal}{\rm K}^h{\rm u}=(v^h,u^h)_{\mathcal{V}},$$

The discretization of the bilinear form reads similarly

$${\rm v}^{h\intercal}{\rm A}^h{\rm u}^h=a(v^h,u^h),\quad \mathcal{V}^h\times\mathcal{V}^h\ni(v^h,u^h)\sim({\rm v}^h,{\rm u}^h)\in\mathbb{R}^{{\rm N}_z}\times\mathbb{R}^{{\rm N}_z}.$$

We further introduce the invertible matrix

$$\Phi_1^{h,\tau}=({\rm M}^h+\tau{\rm A}^h)^{-1}{\rm M}^h,$$

and

$${\rm B}^{h,\tau}=\tau({\rm M}^h+\tau{\rm A}^h)^{-1}{\rm M}^h{\rm B}^h,$$

where with $\dim(\mathcal{U})<\infty$,

$$\mathcal{V}^h\ni v^h\sim{\rm v}^h\in\mathbb{R}^{{\rm N}_z},\quad {\rm v}^{h\intercal}{\rm M}{\rm B}^h\nu=(v^h,P^hB\nu)_{\mathcal{Z}}.$$

Moreover, the operator Q^τ is already finite dimensional leading to $\mathrm{Q}^\tau = \kappa^{-2}\tau^{-1}\mathbb{1}$.

Finally for the observations, as we assume that they are available at each discretization node, hence C^h is simply the matrix selecting the observed nodes, such that

$$\mathrm{y}^{h,\intercal}\mathrm{M}^h_{\mathrm{obs}}\mathrm{C}^h\mathrm{z}^h = (y_h, C^h z_h)_{\mathcal{Y}}, \quad \mathcal{V}^h \times \mathcal{Y}^h \ni (z^h, y^h) \sim (\mathrm{z}^h, \mathrm{y}^h) \in \mathbb{R}^{\mathrm{N}_z} \times \mathbb{R}^{\mathrm{N}_y}$$

with

$$\mathrm{y}_1^\intercal \mathrm{M}^h_{\mathrm{obs}}\mathrm{y}_2 = (y_1^h, y_2^h)_{\mathcal{Y}}, \quad \mathcal{Y}^h \times \mathcal{Y}^h \ni (y_1^h, y_2^h) \sim (\mathrm{y}_1^h, \mathrm{y}_2^h) \in \mathbb{R}^{\mathrm{N}_y} \times \mathbb{R}^{\mathrm{N}_y}$$

Then, C^{h*} is discretized by $(\mathrm{M}^h)^{-1}\mathrm{C}^{h\intercal}\mathrm{M}^h_{\mathrm{obs}}$ and $R^{h,\tau}$ by $\mathrm{R}^{h,\tau} = \gamma\tau\mathrm{M}^h_{\mathrm{obs}}$.

We now proceed to the matrix-based formulation of the discrete-time Kalman filter following the initial algorithm introduced by [32]. It starts with the *initialization*-step:

$$\hat{\mathrm{z}}_0^{h,\tau-} \sim P_h(\hat{z}_0), \tag{4.1a}$$

$$\Pi_0^{h,\tau-} = \mathrm{M}^h(\mathrm{A}^{h\intercal}\mathrm{A}^h)^{-s}\mathrm{M}^h, \tag{4.1b}$$

and then alternates with a *correction*-step:

$$\hat{\mathrm{z}}^+_{n+1} = \hat{\mathrm{z}}^-_{n+1} + \Pi^{h,\tau+}_{n+1}\mathrm{C}^{h\intercal}\mathrm{R}^{h,\tau}\Big(\mathrm{y}^\delta_{n+1} - \mathrm{C}\hat{\mathrm{z}}^-_{n+1}\Big), \qquad n \geq 0. \tag{4.2a}$$

$$\Pi^{h,\tau+}_{n+1} = \big[(\Pi^{h,\tau-}_{n+1})^{-1} + \mathrm{C}^{h\intercal}\mathrm{R}^{h,\tau}\mathrm{C}^h\big]^{-1}, \qquad n \geq 0. \tag{4.2b}$$

followed by a *prediction*-step:

$$\hat{\mathrm{z}}^{h,\tau-}_{n+1} = \Phi_1^{h,\tau}\hat{\mathrm{z}}_n^{h,\tau+}, \qquad n \geq 0, \tag{4.3a}$$

$$\Pi^{h,\tau-}_{n+1} = \Phi_1^{h,\tau}\Pi_n^+\Phi_1^{h,\tau\intercal} + \mathrm{B}^{h,\tau}\mathrm{Q}^\tau\mathrm{B}^{h,\tau\intercal}, \qquad n \geq 0, \tag{4.3b}$$

Note that the above algorithm is well defined because $\Phi_1^{h,\tau}$ is here an invertible matrix. This implies, indeed, that $\Pi^{h,\tau-}_{n+1}$ and $\Pi^{h,\tau+}_{n+1}$ are invertible for all n.

We deduce by mixing the prediction step (4.3) and the correction step (4.2), a one-step time scheme for the covariance matrix

$$\begin{aligned}\Pi^{h,\tau+}_{n+1} = {}& \Phi_1^{h,\tau}\Pi_n^{h,\tau+}\Phi_1^{h,\tau\intercal} + \mathrm{B}^{h,\tau}\mathrm{Q}^\tau\mathrm{B}^{h,\tau\intercal} \\ & - \Phi_1^{h,\tau}\Pi_n^{h,\tau+}\mathrm{C}^{h,\intercal}\left(\mathrm{C}^h\Pi_n^{h,\tau+}\mathrm{C}^{h\intercal} + (\mathrm{R}^{h,\tau})^{-1}\right)^{-1}\mathrm{C}^h\Pi_n^{h,\tau+}\Phi_1^{h,\tau\intercal}, \quad n \geq 0,\end{aligned} \tag{4.4}$$

where we have used the Sherman–Morrison–Woodbury formula [48].

Let us now compute, with the obtained matrices and

$$\Phi_1^{h,\tau} = (\mathbb{1} + \tau(\mathrm{M}^h)^{-1}\mathrm{A}^h) = \mathbb{1} - \tau(\mathrm{M}^h)^{-1}\mathrm{A}^h + \mathcal{O}(\tau^2),$$

the Taylor expansion

$$\begin{aligned}\frac{\Pi^{h,\tau+}_{n+1}\mathrm{M}^h - \Pi^{h,\tau+}_n\mathrm{M}^h}{\tau} = {}& - (\mathrm{M}^h)^{-1}\mathrm{A}^h\left(\Pi_n^{h,\tau+}\mathrm{M}^h\right) - \left(\Pi_n^{h,\tau+}\mathrm{M}^h\right)(\mathrm{M}^h)^{-1}\mathrm{A}^{h\intercal} \\ & + \mathrm{B}^{h,\tau}\mathrm{Q}^\tau\mathrm{B}^\intercal\mathrm{M}^h \\ & - \left(\Pi_n^{h,\tau+}\mathrm{M}^h\right)(\mathrm{M}^h)^{-1}\mathrm{C}^{h\intercal}\mathrm{M}^h_{\mathrm{obs}}\mathrm{C}^h(\mathrm{M}^h)^{-1}\left(\Pi_n^{h,\tau+}\mathrm{M}^h\right)^\intercal + \mathcal{O}(\tau^2),\end{aligned}$$

which is consistent with the Riccati dynamics of $\mathit{\Pi}(t_n)$. This confirms that the M^h-symmetrical matrix $\Pi_n^{h,\tau+}\mathrm{M}^h$ is the matrix representative of $\mathit{\Pi}_n^{h,\tau+}$. The same result holds for $\Pi_n^{h,\tau-}\mathrm{M}^h$ is the matrix representative of $\mathit{\Pi}_n^{h,\tau-}$. We deduce that $\Pi_n^{h,\tau-}$ and $\Pi_n^{h,\tau+}$ are in fact the degrees of freedom representative of the kernel $\pi^{h,\tau-}(x,x') \in \mathrm{H}^1(\Omega\times\Omega)$ and $\pi^{h,\tau+}(x,x') \in \mathrm{H}^1(\Omega\times\Omega)$ which converges in $\mathrm{L}^2(\Omega\times\Omega)$ to the kernel $\pi(x,x',t_n)$. In other words, using $\mathbb{P}_k$ finite elements, we have

$$\forall (x_i,x_j) \in \mathcal{T}_h, \quad \Pi_{n,ij}^{h,\tau-} = \pi^{h,\tau-}(x_i,x_j) = \Pi_{n,ji}^{h,\tau-},$$

and identically

$$\forall (x_i,x_j) \in \mathcal{T}_h, \quad \Pi_{n,ij}^{h,\tau+} = \pi^{h,\tau+}(x_i,x_j) = \Pi_{n,ji}^{h,\tau-}.$$

4.2. $\mathcal{H}$-matrix representation

Hierarchical matrices have been first introduced recently (see *e.g.* [10, 11, 29, 30]) in the context of Partial Differential Equations in order to compress the matrices that typically come from their discretization when using the Boundary Element Method. The main intuitive idea behind this is that the matrices discretize operators acting on spatial functions. The operators under consideration are typically obtained by discretizing convolution operators with Green functions or kernel operators. In the construction of $\mathcal{H}$-matrices, the spatial unknowns are recursively scattered into smaller and smaller boxes in a dyadic tree. The interactions between groups of unknowns of different boxes at different levels correspond to additional diagonal blocks in the matrix and can be approximated using a low-rank representation. It turns out that this approximation can be very accurate for regular kernels and leads to a compressed approximation to the original block. Namely, if one considers a block B of size $m\times n$ in the original matrix, and approximate it to the desired accuracy by a matrix $\tilde{B}$ of rank r, one can compute r pairs of vectors $(u_i,v_i)_{1\le i\le r}$ such that

$$B \simeq \tilde{B} = \sum_{i=1}^{r} u_i v_i^{\intercal}.$$

Since, $u_i \in \mathbb{R}^m$ and $v_i \in \mathbb{R}^n$, storing $\tilde{B}$ requires only $r(m+n)$ data which is much lower than mn if the rank r is much smaller than $\min(m,n)$.

Remarkably, $\mathcal{H}$-matrices not only provide a way to compress matrices in a problem but allow the user to perform algebraic operations, such as additions, multiplications or matrix inversions. Those operations are by now classical and we refer to [10, 30] for their practical implementations which lead to a complete algebra. A native MATLAB version, openHmx, developed by M. Aussal[1] is also available on an open source basis inside the Gypsilab software [2].

We have proved that the covariance operator of the Kalman filter is associated with a kernel of certain regularity, see Theorems 2.3 and 2.4. We have also shown a comparison principle (1.26). Since $\mathit{\Phi}$ is the semigroup associated with the heat-like equation, our comparison principle ensures that the operator has lower and upper bounds represented by low-rank operators generated by the first modes of A and the directions introduced by B. All these theoretical elements justify that the Kalman filter algorithm defined by (4.1)–(4.3)–(4.2) can be well approximated by an $\mathcal{H}$-matrix representation once the following conditions are satisfied:

C1 – the covariance initialization (4.1b) can be proved to be well approximated by an $\mathcal{H}$-matrix;

C2 – the $\mathcal{H}$-matrix linear algebra can recursively detect the $\mathcal{H}$-matrix evolution compression when performing (4.2b) and (4.3b).

For the first condition C1 we know that FEM -based matrices as in (4.1b) are well represented by the $\mathcal{H}$-matrix formalism, with a controlled tolerance [6]. We have introduced a method to obtain a hierarchical representation

[1] The software can be dowloaded at https://github.com/matthieuaussal/gypsilab/tree/master/openHmx

of a sparse operator and its inverses. We consider a spatial partition of the degrees of freedom with a binary tree leading to the classical 2x2 block partition. For a sparse matrix, during the subdivision steps, far interactions lead to empty leaves (without data) and close interactions lead to sparse leaves. Only the sparse leaves are subdivided recursively to the deepest part of the tree. This choice allows us to limit the size of the full leaves that appear during the inversion of the sparse leaves to optimize the hierarchical storage of the inverse matrix. For example, in Figure 4, a hierarchical sparse matrix is similar to a renumbering with hierarchical permutations, and the inverse is well compressed. Only the diagonal leaves are full (red) and the rank of the extra-diagonal leaves is weak (blue), regardless of the size of the block.

To satisfy the second condition C2, we develop an iterative strategy to follow the evolution of the $\mathcal{H}$-matrix representation through the prediction and the correction to preserve the hierarchical structure over time. In particular, the correction requires special treatment. Considering Π^-_{n+1} an $\mathcal{H}$-matrix representing the covariance matrix at the unknowns degrees of freedom, and C a sparse matrix, representing the reduction matrix from the unknowns degrees of freedom to the measurement degrees of freedom, Kalman filtering must first calculate the reduction product $\mathrm{C}\Pi^-_{n+1}\mathrm{C}^\intercal$. This operation is performed in the hierarchical domain, converting the sparse restriction matrix into an $\mathcal{H}$-matrix. However, if the number of measurements is much smaller than the number of unknowns, the standard algebra may no longer be efficient enough to move from one space to another. This is because the transition matrices C become rectangular, which has a large impact on memory and computational cost, as the low-rank representation loses effectiveness. For example, using a low-rank representation $P = XY^\intercal$ with $P \in \mathcal{M}^{mn}$, $X \in \mathcal{M}^{mr}$ and $Y \in \mathcal{M}^{nr}$, the case $m \ll n$ leads to $m \approx r$, which is clearly inefficient. To circumvent this limitation in this particular case, we develop a special $\mathcal{H}$-matrix builder, with a recursive construction directly from the result of $\mathrm{C}\Pi^-_{n+1}\mathrm{C}^\intercal$, instead of successive algebraic operations. Finally, the $\mathcal{H}$ matrix algebra is used at each step of the algorithm, constantly transitioning from state space to measurement space and vice versa. This allows us to maintain a maximum compressibility rate of the operators throughout the process, enabling the computation of large systems. We note that our approach has an analogy with the much more comprehensive $\mathcal{H}$-matrix study [27], but the structure of our matrix equation differs as we consider the $\mathcal{H}$-matrix of the discrete-time Kalman filter.

5. Numerical illustrations

5.1. The 1D heat example

We consider a 1D heat equation on the domain $\Omega = (0, 1)$ with homogeneous Dirichlet conditions. The state space is therefore $\mathcal{Z} = \mathrm{L}^2(0, 1)$ whereas $\mathcal{V} = \mathrm{H}^1_0(0, 1)$. The observation domain is $\omega = (0.3, 0.6)$, namely $\mathcal{Y} = \mathrm{L}^2(0.3, 0.6)$. Therefore C is the restriction operator to ω and C^* extend a function defined in ω by 0 in $\Omega\backslash\omega$. Additionally, we define $B \in \mathcal{L}(\mathbb{R}, \mathcal{Z})$ such that for all $x \in [0, L]$ and $\nu \in \mathbb{R}$, $(B\nu)(x) = \nu\mathbb{1}_\Omega(x)$. We then have

$$B^* z = \int_\Omega z \, \mathrm{d}x.$$

Concerning the initial condition of the covariance operator, we can choose in this 1D case

$$\forall (u, v) \in \mathcal{V}, \quad a_s(u, v) = (\nabla u, \nabla v)_{\mathrm{L}^2(\Omega)(0,1)}.$$

Indeed here, the eigenvalue associated with the Laplacian on $(0, 1)$ with Dirichlet boundary conditions are given by $\lambda_i = \pi^2 i^2$ and $u_i = x \mapsto \sin(i\pi x)$ is a Hilbert basis of $L^2(0, 1)$. Therefore, by choosing $\mathit{\Pi}_0 = -\Delta_0^{-1}$, we have

$$\mathit{\Pi}_0 u_i = \frac{1}{\pi^2 i^2} \text{ and } \sum_{i \geq 1} \frac{1}{\pi^2 i^2} < \infty,$$

hence, $\mathit{\Pi}_0 \in \mathcal{J}_2$ and obviously $\mathit{\Pi}_0 \in \mathcal{D}(\Upsilon)$.

```
% Estimation parameters
covInit  = 1;
covObs   = 1e-2;
covError = 1e-2;


% Dynamics
[L,U] = lu(M+dt*K);
PHI   = (U \ (L \ M));

% Kalmann operators
P     = covInit * M*inv(K)*M;
Bdt   = dt * (MpdtK \ B);
Q     = covError/dt;
BQBt  = Bdt*Q*Bdt';
W     = inv(C*M*C').*covObs/dt;

% Implicit scheme
for t = 1:Nt+1
    % Correction
    Mz   = inv(C*P*C' + W);
    yhat = yhat + P*(C'*(Mz*(zh(:,t)-C*yhat)));
    P    = P - P*C'*Mz*C*P;

  % Prediction
    yhat = MpdtK \ (M*yhat);
    P    = PHI*P*PHI' + BQBt;
end
```

```
% Estimation parameters
covInit  = 1;
covObs   = 1e-2;
covError = 1e-2;
tol      = 1e-6;

% Dynamics;
Mh      = hmx(Xunk,Xunk,M,tol);
Kh      = hmx(Xunk,Xunk,K,tol);
[Lh,Uh] = lu(Mh+dt*Kh);
PHIh    = (Uh \ (Lh \ Mh));

% Kalmann operators
Ph      = covInit * Mh*inv(Kh)*Mh;
Bdt     = dt * (MpdtK\B);
Q       = covError/dt;
BQBth   = hmx(Xunk,Xunk,Bdt*Q,Bdt',tol);
Wh      = inv(hmx(Xmes,Xmes,C*M*C',tol) ).*(covObs/dt);

% Implicit scheme
for t = 1:Nt+1
    % Correction
    CPCth  = hmxBuilderPrj(Xmes,Xmes,C,Ph,C',tol);
    Mzh    = inv(CPCth + Wh);
    CtPzCh = hmxBuilderPrj(Xunk,Xunk,C',Mzh,C,tol);
    yhat   = yhat + Ph*(C'*(Mz*(zh(:,t) - C*yhat)));
    Ph     = Ph - Ph * CtPzCh * Ph;

    % Prediction
    yhat = MpdtK \ (M*yhat);
    Ph   = PHIh * Ph * PHIh' + BQBth;
end
```

Table 1. Matlab Code listing for a simple Kalman filter implementation and its corresponding $\mathcal{H}$-matrix version in Gypsilab.

From a numerical point of view, we discretize the problem with $\mathbb{P}_1$ finite element and choose grids from $N = 10^3$ to $N = 10^4$ elements. We have $\mathrm{N}_z = N - 2$ and the initial covariance $\Pi_0^{h,\tau+} = \mathrm{M}^h(\mathrm{K}^h)^{-1}\mathrm{M}^h$ can be stored in a full format. Then, $\Pi_n^{h,\tau-}$ and $\Pi_n^{h,\tau-}$ can be computed in full format, following the algorithm built on (4.2)–(4.3), and implemented as in the script presented in Table 1-(top) – see the illustrative time-steps of the kernel evolution presented in Figure 1.

This allows to proceed to the numerical verification of Theorem 3.4. In this respect we consider a fine discretization $h_1 = 10^{-4}$, $\tau_1 = 10^{-3}$ and coarser discretizations $10^{-4} \leq h_2 \leq 10^{-3}$, $10^{-2} < \tau_2 < 10^{-3}$ with $\tau_2/\tau_1 \in \mathbb{N}$. We expect to compute for $t_n = n\tau_2 = k(\tau_2/\tau_1)\tau_1$ with $0 \leq n \leq T/\tau_2$,

$$\|P_{h_1}^{h_2}\mathit{\Pi}_n^{h_1,\tau_1,-}P_{h_1}^{h_2*} - \mathit{\Pi}_n^{h_2,\tau_2,+}\|_2 = \Bigg(\sum_{1\leq i\leq N_y}\Big(e_i^{h_2}, \big[P_{h_1}^{h_2}\mathit{\Pi}_n^{h_1,\tau_1,+}P_{h_1}^{h_2*} - \mathit{\Pi}_n^{h_2,\tau_2,-}\big]^2 e_i^{h_2}\Big)_{\mathrm{L}^2(0,1)}\Bigg)^{\frac{1}{2}}$$

where $P_{h_1}^{h_2} \in \mathcal{L}(\mathcal{Z}^{h_1}, \mathcal{Z}^{h_2})$ is the orthogonal projection from $\mathcal{Z}^{h_1}$ to $\mathcal{Z}^{h_2}$ and $(e_i^{h_2})_{1\leq i\leq N_y}$ is the finite element basis of $\mathcal{Z}^{h_2}$. Therefore, we get using matrices

$$\|P_{h_1}^{h_2}\mathit{\Pi}_n^{h_1,\tau_1,-}P_{h_1}^{h_2*} - \mathit{\Pi}_n^{h_2,\tau_2,+}\|_2 = \Bigg(\sum_{1\leq i\leq N_y}\mathrm{e}_i^{h_2\intercal}\mathrm{M}^{h_2}\big[\mathrm{P}_{h_1}^{h_2}\Pi_n^{h_1,\tau_1,-}\mathrm{M}^{h_1}\mathrm{P}_{h_1}^{h_2\intercal}\mathrm{M}^{h_2} - \Pi_n^{h_2,\tau_2,-}\mathrm{M}^{h_2}\big]\mathrm{e}_i^{h_2}\Bigg)^{\frac{1}{2}}$$

with $\mathrm{P}_{h_1}^{h_2} = (\mathrm{M}^{h_2})^{-1}\mathrm{I}_{h_1}^{h_2}$ – while $\mathrm{I}_{h_1}^{h_2} \in \mathbb{R}^{N_2\times N_1}$ is the interpolation matrix from the grid of step h_1 to the grid of step h_2 – and its corresponding adjoint $\mathrm{P}_{h_1}^{h_2\intercal}.\mathrm{M}^{h_2}$. We obtain

$$\|P_{h_1}^{h_2}\mathit{\Pi}_n^{h_1,\tau_1,-}P_{h_1}^{h_2*} - \mathit{\Pi}_n^{h_2,\tau_2,-}\|_2 = \|\mathrm{I}_{h_1}^{h_2}\mathrm{M}^{h_1}\Pi_n^{h_1,\tau_1,-}\mathrm{M}^{h_1}\mathrm{I}_{h_1}^{h_2\intercal} - \mathrm{M}^{h_2}\Pi_n^{h_2,\tau_2,-}\mathrm{M}^{h_2}\|_F \tag{5.1}$$

where $\|\cdot\|_F$ denotes the Frobenius matrix norm. This last identity allows us to investigate the convergence of the Kalman filter algorithm in dense format when h and τ tend to 0. The convergence results are presented in Figure 2 (*Top*), hence illustrating Theorem 3.4.

In the same Figure 2 (Bottom), we also numerical illustrate the approximation by an $\mathcal{H}$-matrix $\tilde{\Pi}_n^{h,\tau,\epsilon,-}$ of the covariance operator $\Pi_n^{h,\tau,-}$ as the tolerance ϵ tends to 0. At final time $T = 1$ and $\epsilon = 10^-6$, the $\mathcal{H}$-matrix $\tilde{\Pi}_n^{h,\tau,\epsilon,-}$ is only represented by a low rank matrix of $d = 6$ vectors $(w_k)_{1\leq k\leq d}$ plotted in Figure 3. Namely we have

$$\tilde{\Pi}_{N_T}^{h,\tau,\epsilon,-} = \sum_{k=1}^{d} w_k^\intercal w_k \simeq \Pi_{N_T}^{h,\tau,-}.$$

One existing numerical question is whether a low rank strategy based on Riccati operator reduction [35, 46] could have given the same type of results directly without relying on the $\mathcal{H}$-matrix machinery. We typically could have thought on directly projecting the initial covariance matrix on

$$\mathcal{E}_m = \operatorname{span}(e_0, e_1, \ldots, e_m),$$

where $(u_i)_{1\leq i\leq m}$ are the first m eigenvectors of the Laplacian with Dirichlet conditions and $e_0 = (\mathbb{1} - \tau A)^{-1}w \in \mathcal{D}(A) \in \mathcal{V}$ is defined from the vector $w = 1_\Omega$ used to construct B^*. Indeed in Figure 3 we recognize e_0 as very close to the first vector w_1 in the low rank representation of $\tilde{\Pi}_{N_T}^{h,\tau,\epsilon,-}$. This is natural with respect to the Duhamel-like formula (1.18) and the comparison principle in Section 1.6.2.

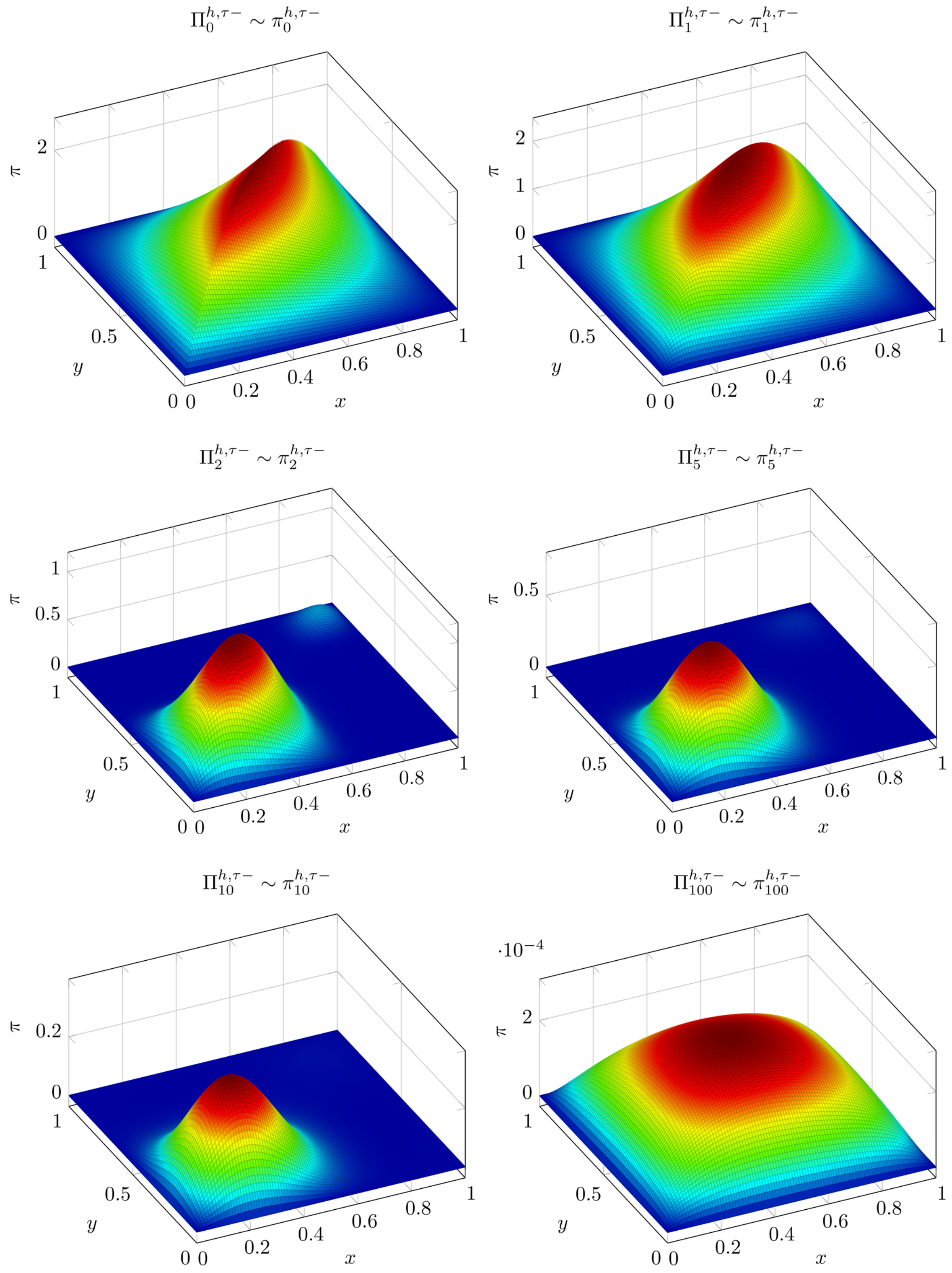
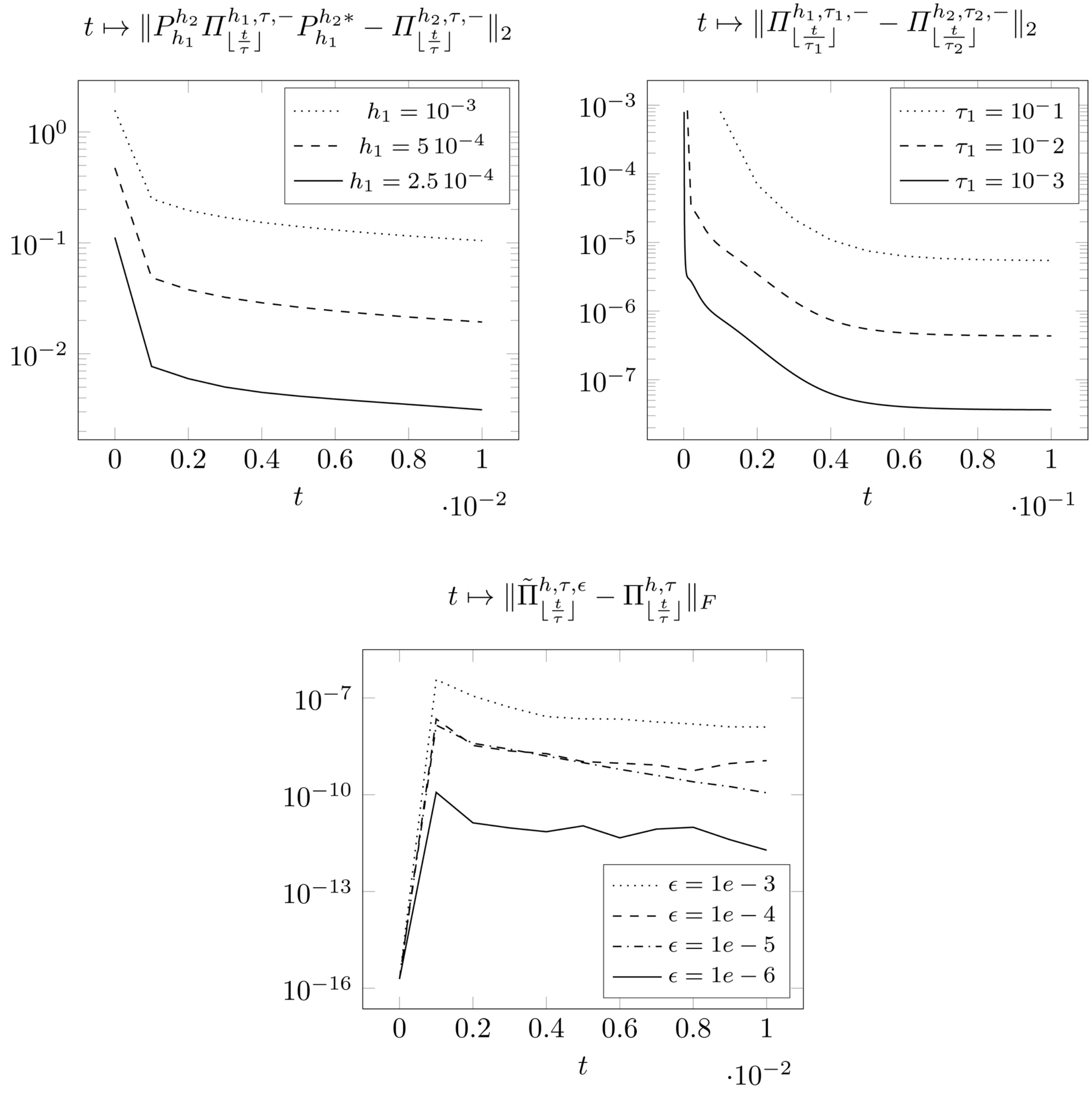


FIGURE 1. Kernel evolution $\pi(x, x', t)$ at different time step $t = 0, \tau, 2\tau, 5\tau, 10\tau, 100\tau$, approximated with $\pi_n^{h,\tau,-} \sim \Pi_n^{h,\tau,-}$, with $n \in \{0, 1, 2, 5, 10, 100\}$.



$t \mapsto \|P_{h_1}^{h_2} \mathit{\Pi}_{\lfloor \frac{t}{\tau} \rfloor}^{h_1,\tau,-} P_{h_1}^{h_2 *} - \mathit{\Pi}_{\lfloor \frac{t}{\tau} \rfloor}^{h_2,\tau,-}\|_2$

$t \mapsto \|\mathit{\Pi}_{\lfloor \frac{t}{\tau_1} \rfloor}^{h_1,\tau_1,-} - \mathit{\Pi}_{\lfloor \frac{t}{\tau_2} \rfloor}^{h_2,\tau_2,-}\|_2$

$t \mapsto \|\tilde{\Pi}_{\lfloor \frac{t}{\tau} \rfloor}^{h,\tau,\epsilon} - \Pi_{\lfloor \frac{t}{\tau} \rfloor}^{h,\tau}\|_F$

FIGURE 2. Convergence results with respect to the space-discretization, time-discretization and the $\mathcal{H}$-matrix representation precision.

To investigate numerically this question, we can compute a sort of distance between the vector spaces $\mathcal{E}_m$ and the space $\mathcal{W}_6 = \operatorname{span}(w_1, \cdots, w_6)$ as m increases. In this respect, we use the quantity introduced in [47]

$$\theta_m = \min\left(\sup_{\substack{v \in \mathcal{W}_6 \\ \|v\|=1}} d(v, \mathcal{E}_m), \sup_{\substack{v \in \mathcal{E}_m \\ \|v\|_{\mathcal{Z}}=1}} d(v, \mathcal{W}_6) \right).$$

$$\sup_{\substack{v\in\mathcal{E}_m\\ \|v\|_{\mathcal{Z}}=1}} d(v,\mathcal{W}_d) = \sup_{v\in\mathcal{E}_m}\inf_{w\in\mathcal{W}_d}\frac{\|v-w\|_{\mathcal{Z}}^2}{\|v\|_{\mathcal{Z}}^2}.$$

Computing the distance, can be done numerically based on the following computation. We introduce the Grammian matrices

$$\Lambda_d = ((w_i,w_j)_{\mathcal{Z}})_{1\leq i,j\leq d},\quad \Lambda_m = ((u_i,u_j)_{\mathcal{Z}})_{0\leq i,j\leq m},\quad \Lambda_{n,m} = ((w_i,u_j)_{\mathcal{Z}})_{\substack{1\leq i\leq d\\ 0\leq j\leq m}}.$$

As we have the following decomposition

$$\forall w\in\mathcal{W}_d,\quad w = \sum_{1\leq i\leq d}\mathrm{w}_i w_i,\ \text{ with } \mathrm{w} = \begin{pmatrix}\mathrm{w}_1\\ \vdots\\ \mathrm{w}_d\end{pmatrix}\in\mathbb{R}^d,$$

we can compute

$$d(v,\mathcal{W}_d) = \inf_{\mathrm{w}\in\mathbb{R}^d}\frac{\mathrm{w}^\intercal\Lambda_d\mathrm{w} + \mathrm{v}^\intercal\Lambda_m\mathrm{v} - 2\mathrm{w}^\intercal\Lambda_{d,m}\mathrm{v}}{\mathrm{v}^\intercal\Lambda_d\mathrm{v}}.$$

As the minimum is obtained in $\mathrm{w} = \Lambda_d^{-1}\Lambda_{d,m}\mathrm{v}$, we get

$$\sup_{v\in\mathcal{E}_m} d(v,\mathcal{W}_d) = \sup_{\mathrm{v}\in\mathbb{R}^m}\left\{R(\Lambda_m - \Lambda_{d,m}'\Lambda_d^{-1}\Lambda_{d,m},\Lambda_m,\mathrm{v}) = \frac{\mathrm{v}^\intercal(\Lambda_m - \Lambda_{d,m}'\Lambda_d^{-1}\Lambda_{d,m})\mathrm{v}}{\mathrm{v}^\intercal\Lambda_m\mathrm{v}}\right\},$$

where we recognize the supremum of the Rayleigh quotient with real symmetric matrices, hence corresponding to the largest eigenvalue of the eigenvalue problem

$$(\Lambda_m - \Lambda_{d,m}^\intercal\Lambda_d^{-1}\Lambda_{d,m})\mathrm{v} = \lambda\Lambda_m\mathrm{v}.$$

Therefore by solving two eigenvalue problems, we easily compute numerically

$$\theta_m = \min\Big(\max_{\mathrm{v}\in\mathbb{R}^m} R(\Lambda_m - \Lambda_{d,m}'\Lambda_d^{-1}\Lambda_{d,m},\Lambda_m,\mathrm{v}),\ \max_{\mathrm{w}\in\mathbb{R}^m} R(\Lambda_d - \Lambda_{d,m}\Lambda_m^{-1}\Lambda_{d,m}^\intercal,\Lambda_d,\mathrm{w})\Big),$$

giving a space-similarity index between 0 and 1, as plotted in Figure 3 (right). We see here that the first 3 vectors could have been envisioned *a priori* in the decomposition of $\tilde{\Pi}_N^{h,\tau,\epsilon,-}$, whereas, we need much more eigenvectors to represent the additional 3 vectors. Our $\mathcal{H}$-matrix formulation offers therefore a real new point of view with respect to reduced-based strategy with *a priori* reduction of the covariance operator.

5.2. The 2D heat example

We now proceed to a 2D case $\Omega = (0,1)^2$ with an observation domain $\omega = (0.3,0.6)^2$. The eigenvalue associated with the Laplacian on $(0,1)^2$ with Dirichlet boundary conditions are given by $\lambda_{i,j} = \pi^2(i^2+j^2)$ and $u_i = x\mapsto\sin(i\pi x)\sin(j\pi x)$. We choose $\mathit{\Pi}_0 = \Delta_0^{-1}\in\mathcal{L}(\mathrm{L}^2(\Omega),\mathrm{H}^2(\Omega))$. As $\mathrm{H}^2(\Omega)\subset\mathrm{L}^\infty(\Omega)$, we know that

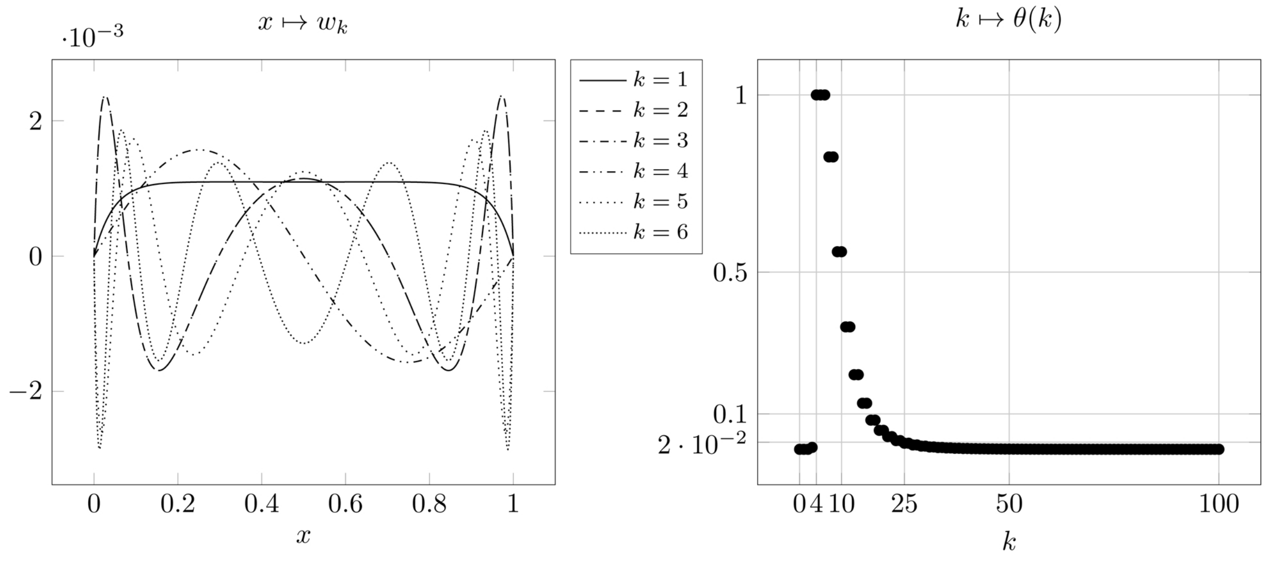


FIGURE 3. (*Left*) final normalized base of vector generating the low rank approximation $\Pi_T^{h,\tau,\epsilon}$. (*Right*) Dissimilarity index between subspace $\mathcal{W}^{h,k}$ and $\mathcal{E}^{h,k}$.

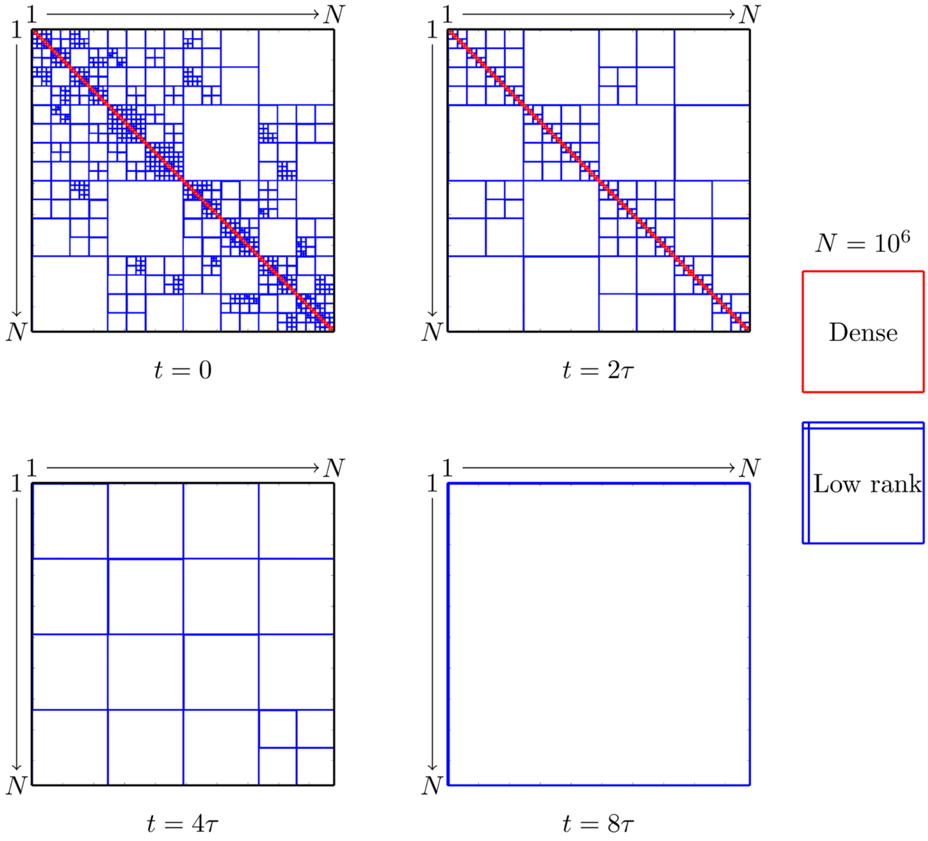
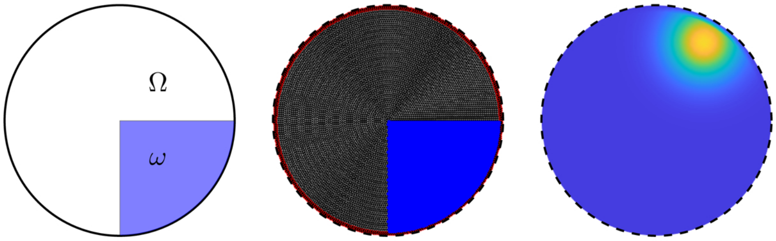
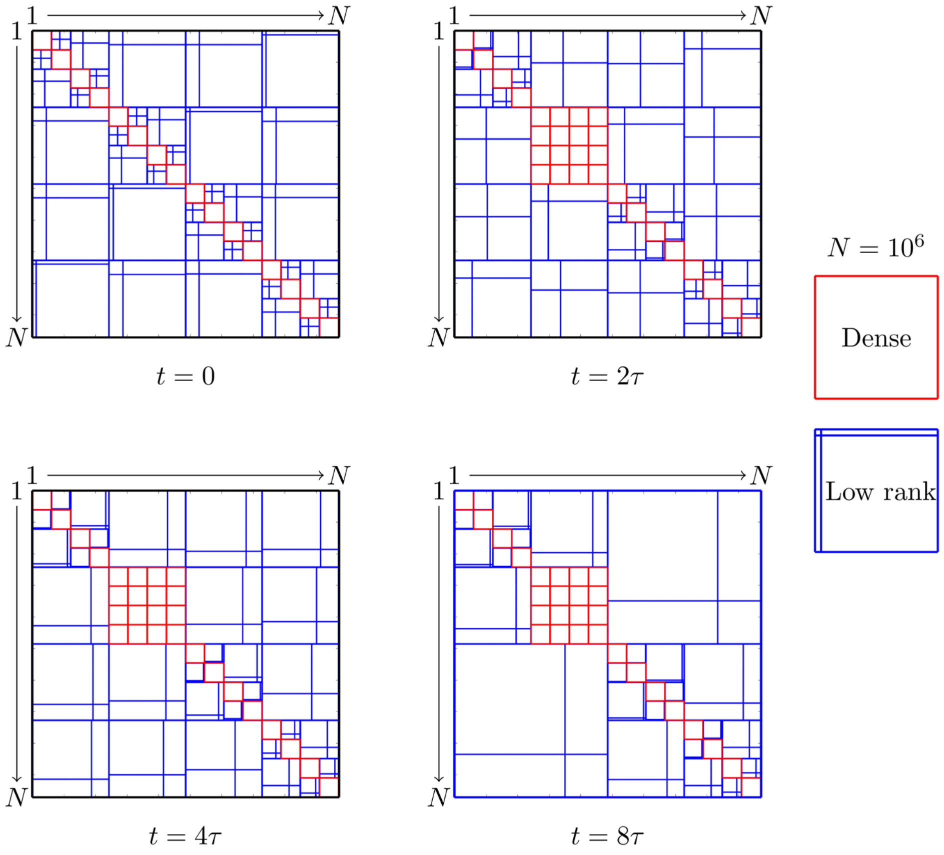


FIGURE 4. $\mathcal{H}$-matrix evolution for the heat equation.



(a) geometry (b) mesh (c) Initial solution

FIGURE 5. Geometry, mesh and initial condition of the advection-diffusion problem.



$t = 0$ $t = 2\tau$ $t = 4\tau$ $t = 8\tau$

FIGURE 6. $\mathcal{H}$-matrix evolution for the advection-diffusion equation.

$\mathit{\Pi}_0 \in \mathcal{J}_2(\mathcal{Z})$ – see Section 1.8.4 of [3]. This is also confirmed by

$$\sum_{(i,j)\in\mathbb{N}^2}\frac{1}{\lambda_{i,j}^2} = \sum_{k\in\mathbb{N}^2}\sum_{\substack{(i,j)\in\mathbb{N}^2\\ i^2+j^2=k^2}}\frac{1}{r^4} = \sum_{k\in\mathbb{N}^2}\frac{r_2^2(k^2)}{r^4} < +\infty,$$

where $r_2^2(k^2) = \sharp\{(i,j)\in\mathbb{N}^2 \,|\, i^2+j^2=k^2\} < k^2$.

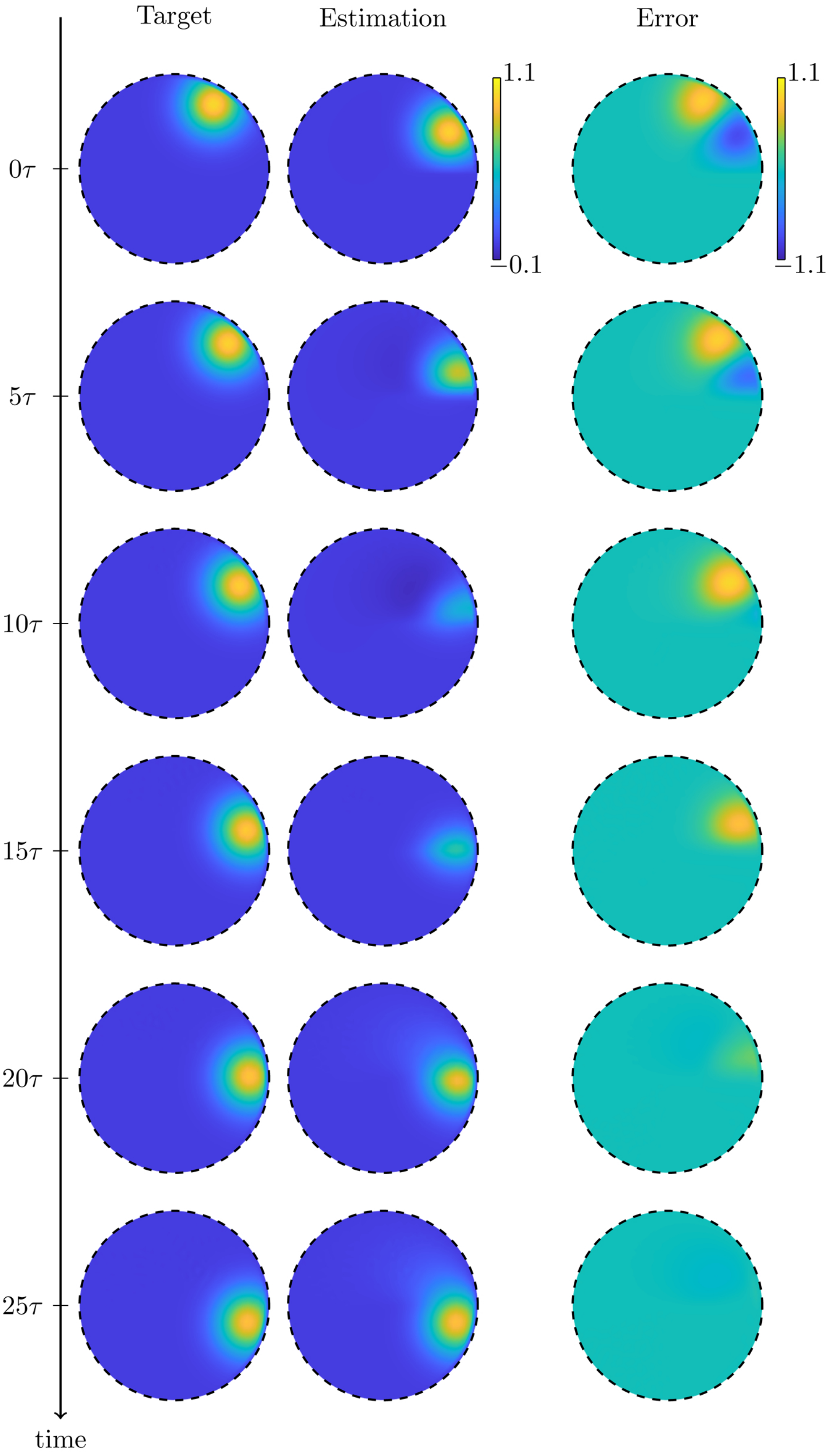


FIGURE 7. Direct solution and corresponding state estimator for the advection-diffusion problem: first time steps.

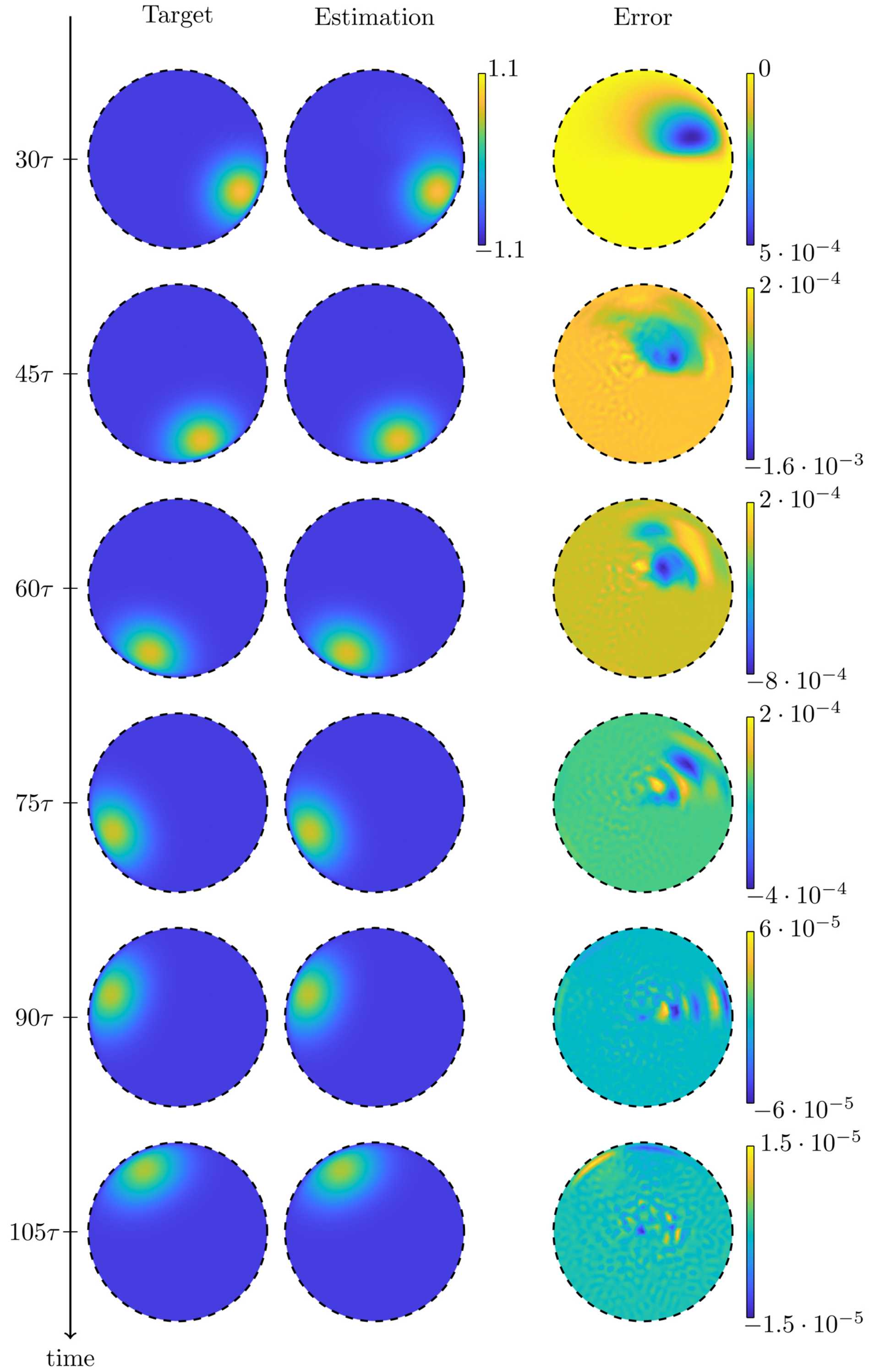


FIGURE 8. Direct solution and corresponding state estimator for the advection-diffusion problem: larger time steps.

We discretize the square with $h = 10^{-3}$ in each direction, leading to 10^6 degrees of freedom for the full discretized problem. Storing the dense matrix representation of $\Pi_n^{h,\tau,+}$ and $\Pi_n^{h,\tau,-}$ require at each iteration n. However, with our $\mathcal{H}$-matrix representation, we circumvent the curse of dimensionality and the memory cost diminishes through time as the covariances become closer and closer to a low rank operator. This is illustrated in Figure 4, where we plot the $\mathcal{H}$-matrix representation through time.

5.3. An advection diffusion example

Finally, we propose a final illustration of our approach through the more involved advection-diffusion example in a circular geometry see Figure 5. The advection field is orthoradial to be compatible with the boundary condition and the target initial solution is a Gaussian function positioned north-east with observation on the south-east quadrant. The $\mathcal{H}$-matrix evolution is pictured in Figure 6 while the estimation results are presented in Figure 7-8, illustrating the algorithm performance on this more advanced case.

References

[1] A. Aalto, Convergence of discrete-time Kalman filter estimate to continuous time estimate. *Int. J. Control* **89** (2016) 668–679.

[2] F. Alouges and M. Aussal, FEM and BEM simulations with the gypsilab framework. *SMAI J. Comput. Math.* **4** (2018) 297–318.

[3] W. Arendt, Heat kernels, Technical report, ISEM course, 2005–2006.

[4] J.P. Aubin, Applied Functional Analysis, 2nd ed. Wiley (2000).

[5] U. Baur, P. Benner and L. Feng, Model order reduction for linear and nonlinear systems: a system-theoretic perspective. *Arch. Comput. Methods Eng.* **21** (2014) 331–358.

[6] W. Bebendorf and M. Hackbusch, Existence of H-matrix approximants to the inverse FE-matrix of elliptic operators with L^∞-coefficients. *Numer. Math.* **95** (2003) 1–28.

[7] A. Bensoussan, Filtrage optimal des systèmes linéaires. Dunod (1971).

[8] A. Bensoussan, Estimation and Control of Dynamical Systems. Interdisciplinary Applied Mathematics. Springer, Cham (2018).

[9] A. Bensoussan, M.C. Delfour, G. Da Prato and S.K. Mitter, Representation and Control of Infinite Dimensional Systems, second edition. Birkhauser Verlag, Boston (2007).

[10] S. Börm, volume 14 *Efficient numerical methods for non-local operators: H2-matrix compression, algorithms and analysis*. European Mathematical Society (2010).

[11] S. Börm, L. Grasedyck and W. Hackbusch, Introduction to hierarchical matrices with applications. *Eng. Anal. Boundary Elem.* **27** (2003) 405–422.

[12] E. Burman and L. Oksanen, Data assimilation for the heat equation using stabilized finite element methods. *Numer. Math.* **139** (2018) 505–528.

[13] J.A. Burns, E.M. Cliff and C.N. Rautenberg, A distributed parameter control approach to optimal filtering and smoothing with mobile sensor networks, In *17th Mediterranean Conference on Control and Automation* (2009), pp. 181–186.

[14] J.A. Burns and C.N. Rautenberg, Solutions and approximations to the Riccati integral equation with values in a space of compact operators. *SIAM J. Control Optim.* **53** (2015) 2846–2877.

[15] J.A. Burns and C.N. Rautenberg, The infinite-dimensional optimal filtering problem with mobile and stationary sensor networks. *Numer. Funct. Anal. Optim.* **36** (2015) 181–224.

[16] D. Chapelle, M. Fragu, V. Mallet and P. Moireau, Fundamental principles of data assimilation underlying the Verdandi library: applications to biophysical model personalization within euHeart. *Med. Biol. Eng. Comput.* (2012).

[17] D. Chapelle, A. Gariah, P. Moireau and J. Sainte-Marie, A Galerkin strategy with proper orthogonal decomposition for parameter-dependent problems - analysis, assessments and applications to parameter estimation. *ESAIM: Math. Model. Numer. Anal.* **47** (2013) 1821–1843.

[18] G. Chavent, Nonlinear Least Squares for Inverse Problems. Springer (2010).

[19] R.F. Curtain, A survey of infinite-dimensional filtering. *SIAM Rev.* **17** (1975) 395–411.

[20] R.F. Curtain, K. Mikkola and A. Sasane, The Hilbert-Schmidt property of feedback operators. *J. Math. Anal. Appl.* **329** (2007) 1145–1160.

[21] R.F. Curtain and H. Zwart, An introduction to infinite-dimensional linear systems theory. Vol. 21 of *Texts in Applied Mathematics*. Springer-Verlag, New York (1995).

[22] S. De Marchi and M. Vianello, Peano's kernel theorem for vector-valued functions and some applications. *Numer. Funct. Anal. Optim.* **17** (1996) 57–64.

[23] H.W. Engl, M. Hanke and A. Neubauer, Regularization of inverse problems. Vol. 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht (1996).

[24] F. Flandoli, On the semigroup approach to stochastic evolution equations. *Stoch. Anal. Appl.* **10** (1992) 181–203.

[25] H. Fujita, N. Saito and T. Suzuki, Operator Theory and Numerical Methods, Studies in Mathematics and Its Applications. North Holland (2001).

[26] A. Germani, L. Jetto and M Piccioni, Galerkin approximation for optimal linear filtering of infinite-dimensional linear systems. *SIAM J. Control Optim.* **26** (1988) 1287–1305.

[27] L. Grasedyck, W. Hackbusch and B.N. Khoromskij, Solution of large scale algebraic matrix Riccati equations by use of hierarchical matrices. *Comput. Arch. Sci. Comput.* **70** (2003) 121–165.

[28] S. Guerrero and G. Lebeau, Singular optimal control for a transport-diffusion equation. *Commun. Partial Differ. Equ.* **32** (2007) 1813–1836.

[29] W. Hackbusch, A sparse matrix arithmetic based on H-matrices. I. Introduction to H-matrices. *Comput. Arch. Sci. Comput.* **62** (1999) 89–108.

[30] W. Hackbusch, Hierarchical Matrices: Algorithms and Analysis, Springer Publishing Company, Incorporated, 1st edition (2015).

[31] R.E. Kalman, A new approach to linear filtering and prediction problems. *J. Basic Eng.* **82** (1960) 35–45.

[32] R.E. Kalman and R.S. Bucy, New results in linear filtering and prediction theory. *J. Basic Eng.* **83** (1961) 95–108.

[33] T. Kato, Fractional powers of dissipative operators. *J. Math. Soc. Jpn.* **13** (1961) 246–274.

[34] I. Lasiecka and R. Triggiani, Differential and Algebraic Riccati Equations with Application to Boundary/Point Control Problems: Continuous Theory and Approximation Theory. Lecture Notes in Control and Information Sciences. Springer, Berlin, Heidelberg (1991).

[35] C. Le Bris and P. Rouchon, Low-rank numerical approximations for high-dimensional Lindblad equations. *Phys. Rev. A* **87** (2013) 022125.

[36] F.-X. Le Dimet, Optimal control for data assimilation in meteorology, In *Control theory of distributed parameter systems and applications (Shanghai, 1990).* Vol. 159 of *Lecture Notes in Control and Inform. Sci.* Springer, Berlin (1991), pp. 51–60.

[37] F.-X. Le Dimet and O. Talagrand, Variational algorithms for analysis and assimilation of meteorological observation: theoretical aspects. *Tellus* **38** (1986) 97–110.

[38] J.Y. Li, S. Ambikasaran, E.F. Darve and P.K. Kitanidis, A Kalman filter powered by H2-matrices for quasi-continuous data assimilation problems. *Water Resour. Res.* **50** (2014) 3734–3749.

[39] J.-L. Lions, *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Avant propos de P. Lelong. Dunod, Paris (1968).

[40] J.-L. Lions, Exact controllability, stabilization and perturbations for distributed systems. *SIAM Rev.* **30** (1988) 1–68.

[41] J.-L Lions, Espaces d'interpolation et domaines de puissances fractionnaires d'opérateurs. *J. Math. Soc. Jpn.* **14** (1962) 233–241.

[42] Y. Maday, A.T. Patera, J.D. Penn and M. Yano, A parameterized-background data-weak approach to variational data assimilation: formulation, analysis, and application to acoustics. *Int. J. Numer. Methods Eng.* **102** (2014) 933–965.

[43] A. Nassiopoulos and F. Bourquin, Fast three-dimensional temperature reconstruction. *Comput. Methods Appl. Mech. Eng.* **199** (2010) 3169–3178.

[44] S. Pagani, A. Manzoni and A. Quarteroni, Efficient state/parameter estimation in nonlinear unsteady PDEs by a reduced basis ensemble Kalman filter. *SIAM/ASA J. Uncert. Quantif.* **5** (2017) 890–921.

[45] A. Pazy, Semigroups of linear operators and applications to partial differential equations. Vol. 44 of *Applied Mathematical Sciences.* Springer-Verlag, New York (1983).

[46] D.T. Pham, J. Verron and M.C. Roubaud, A singular evolutive extended Kalman filter for data assimilation in oceanography. *J. Mar. Syst.* **16** (1998) 323–340.

[47] S. Sellam and A. Forcioli, Introduction de la notion d'écart entre sous-espaces vectoriels en analyse de données. *RAIRO: Oper. Res.* **14** (1980).

[48] D. Simon, Optimal State Estimation: Kalman, H^∞, and Nonlinear Approaches. Wiley-Interscience (2006).

[49] H. Tanabe, Equations of evolution. Vol. 6 of *Monographs and Studies in Mathematics*, Pitman (Advanced Publishing Program), Boston, Mass.-London (1979).

[50] R. Temam, Sur l'équation de Riccati associée à des opérateurs non bornés, en dimension infinie. *J. Funct. Anal.* **7** (1971) 85–115.