IOP Publishing Inverse Problems
Inverse Problems **36** (2020) 075006 (28pp) https://doi.org/10.1088/1361-6420/ab8bc5

Continuous limits for constrained ensemble Kalman filter

Michael Herty and Giuseppe Visconti

Institut für Geometrie und Praktische Mathematik (IGPM) RWTH Aachen University
Templergraben 55, 52062, Aachen, Germany

Received 18 December 2019, revised 6 April 2020
Accepted for publication 21 April 2020
Published 17 June 2020

Abstract

The ensemble Kalman filter method can be used as an iterative particle numerical scheme for state dynamics estimation and control-to-observable identification problems. In applications it may be required to enforce the solution to satisfy equality constraints on the control space. In this work we deal with this problem from a constrained optimization point of view, deriving corresponding optimality conditions. Continuous limits, in time and in the number of particles, allows us to study properties of the method. We illustrate the performance of the method by using test inverse problems from the literature.

Keywords: inverse problems, constrained optimization, nonlinear filtering methods, DAE systems, mean-field limit

(Some figures may appear in colour only in the online journal)

1. Introduction

We are concerned with the following abstract inverse problem or parameter identification problem

$$\mathbf{y} = \mathcal{G}(\mathbf{u}) + \boldsymbol{\eta} \tag{1}$$

where $\mathcal{G} : X \to Y$ is the (possible nonlinear) forward operator between the finite dimensional Hilbert spaces $X = \mathbb{R}^d$ and $Y = \mathbb{R}^K$, with $d, K \in \mathbb{N}$, $\mathbf{u} \in X$ is the control, $\mathbf{y} \in Y$ is the observation and $\boldsymbol{\eta}$ is observational noise. Typically, $\boldsymbol{\eta}$ is not explicitly known but only information on its distribution is available. Throughout the paper we assume that $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}^{-1})$, i.e. the observational noise is normally distributed with zero mean and given covariance matrix $\boldsymbol{\Gamma}^{-1} \in \mathbb{R}^{K\times K}$. Given noisy measurements or observations $\mathbf{y}$ and the known mathematical model $\mathcal{G}$, we are interested in finding the corresponding control $\mathbf{u}$. Inverse problems, in particular in view of a possible ill-posedness, have been discussed in vast amount of literature and we refer to [1] for an introduction and further references.

We are interested in particle methods for solving numerically (1). In particular, in the following we will investigate the ensemble Kalman filter (EnKF). While this method has already been introduced more than ten years ago [2], recent theoretical progress have been done both for the continuous time limit [3–7] and the mean-field limit on the number of the ensemble members [8–11].

Solving inverse problems or identification problems arising in realistic applications usually requires to consider constraints on the unknown control, as well as on the data. This has been demonstrated in the recent literature in several research fields, as for instance in weather forecasting [12], milling process [13] and process control [14]. Including constraints in Kalman filtering is usually done via optimization techniques, see e.g. [15]. Recently, algorithms for the ensemble Kalman filter for constrained problems have been studied in [16, 17]. In particular, in [16] the authors propose to solve the constrained compromise step for all the ensemble members that do not fulfill the constraints. However, we notice that no optimality conditions have been formally derived and that the solution of a large number of constrained optimization problems may be required. Instead, in [17] the authors focus on linear box-constraints and, via a projection-based method, derive a new variant of the EnKF and study the continuous time limit.

Here, we follow a similar approach to [16] by incorporating equality constraints in the compromise or update step of the unconstrained EnKF. Using the Lagrange theory for optimization problems, we derive first order necessary optimality conditions and discuss the link to game theory. The formulation of the optimality conditions shows that the unconstrained EnKF automatically satisfies linear equality constraints. Our analysis is therefore mainly focused to the case of nonlinear equality constraints. Under suitable scaling assumptions, we then compute the corresponding continuous time limit of the optimality conditions, which leads to a system of differential algebraic equations (DAEs). This reformulation allows us to perform an analysis of the method. A further continuous limit is analyzed in the mean-field limit, i.e. in the regime of infinitely many ensembles.

The main contributions can be summarized in the following points:

- We explicitly derive the ensemble Kalman update for inverse problems which automatically encapsulates and satisfies general nonlinear equality constraints. In this way, we avoid the possible solution of many constrained optimization problems for those ensemble members which do not satisfy the constraints.
- We compute and analyze the corresponding continuous limits, in time and in the number of the ensemble members, of the constrained ensemble Kalman update formula.

The paper is organized as follows. In section 2 we briefly review the derivation of the EnKF and then we compute first order necessary optimality conditions for the constrained optimization problem, assuming only equality constraints. In section 3 we study continuous limits, in time and in the number of ensembles, of the optimality conditions, providing an analysis for the resulting system of DAEs and mean-field equation. In section 4 we investigate the ability of the method to provide solution to an inverse problem with two types of equality constraints. Finally, we summarize the results in section 5. We report in appendix A detailed elaboration of computations in the proof of proposition 2.5. Appendix B is aimed to list the most used mathematical symbols throughout the paper.

2. Ensemble Kalman filter for constrained problems

A particular numerical method for solving (1) is the ensemble Kalman filter (EnKF), which has been originally introduced to estimate state variables, parameters, etc of stochastic dynamical

systems. The estimations are based on system dynamics and measurement data that are possibly perturbed by known noise. Therefore, in order to apply the EnKF to the inverse problem (1), this is usually rewritten as a partially observed and artificial dynamical system based on state augmentation, e.g. see [16, 18, 19].

Let us introduce the new variable $\mathbf{w} = \mathcal{G}(\mathbf{u})$, so that (1) can be reformulated as

$$\begin{aligned} \mathbf{w} &= \mathcal{G}(\mathbf{u}) \\ \mathbf{y} &= \mathbf{w} + \boldsymbol{\eta}. \end{aligned}$$

Taking $\mathbf{y}^{n+1} = \mathbf{y}$ and $\boldsymbol{\eta}^{n+1} = \boldsymbol{\eta}$ as the given data and the given noise, respectively, we obtain the following dynamical system:

$$\begin{aligned} \mathbf{u}^{n+1} &= \mathbf{u}^n \\ \mathbf{w}^{n+1} &= \mathcal{G}(\mathbf{u}^n) \\ \mathbf{y}^{n+1} &= \mathbf{w}^{n+1} + \boldsymbol{\eta}^{n+1}. \end{aligned} \tag{2}$$

Remark 2.1. We stress the fact that here n is an artificial time index, while in the case of a dynamic inverse problem physical time is included in $\mathcal{G}$. Moreover, in the following we keep the notation $\mathbf{y}^{n+1}$ even if we consider the free noise case where no perturbation in time is added to the initial noisy observation.

In the following we briefly recall the derivation of the EnKF. We use a compact formulation of the artificial dynamic process by defining $\mathbf{v} = (\mathbf{u}, \mathbf{w})^\intercal \in \mathbb{R}^{d+K}$, $\boldsymbol{\Xi}(\mathbf{v}) = (\mathbf{u}, \mathcal{G}(\mathbf{u}))^\intercal \in \mathbb{R}^{d+K}$ and the observational matrices $\mathbf{H} = [\mathbf{0}, \mathbf{I}] \in \mathbb{R}^{K\times(d+K)}$, $\mathbf{H}^\perp = [\mathbf{I}, \mathbf{0}] \in \mathbb{R}^{d\times(d+K)}$. We note that $\mathbf{H}\mathbf{v} = \mathbf{w}$ and $\mathbf{H}^\perp\mathbf{v} = \mathbf{u}$. Then, we rewrite (2) in the typical setting where the EnKF is applied for the solution of the inverse problem (1):

$$\begin{aligned} \mathbf{v}^{n+1} &= \boldsymbol{\Xi}(\mathbf{v}^n) \\ \mathbf{y}^{n+1} &= \mathbf{H}\mathbf{v}^{n+1} + \boldsymbol{\eta}^{n+1}. \end{aligned} \tag{3}$$

Let us introduce $\{\mathbf{v}^{j,n}\}_{j=1}^J$ the J particles (ensembles) at time n. The objective of the EnKF is to reach a compromise $\mathbf{v}^{j,n+1}$ between the background estimate $\hat{\mathbf{v}}^{j,n+1}$ of the model and additional information provided by data $\mathbf{y}^{n+1}$, for each ensemble member. The state of the particles at time $n+1$ is predicted using the dynamics model (3) to obtain

$$\hat{\mathbf{v}}^{j,n+1} = \boldsymbol{\Xi}(\mathbf{v}^{j,n}), \quad j = 1, \ldots, J. \tag{4}$$

Let $\mathbf{C}^{n+1} \in \mathbb{R}^{(d+K)\times(d+K)}$ be the empirical covariance matrix characterizing the uncertainties on predictions $\hat{\mathbf{v}}^{j,n+1}$. We have

$$\mathbf{C}^{n+1} = \frac{1}{J}\sum_{k=1}^{J} (\hat{\mathbf{v}}^{j,n+1} - \overline{\hat{\mathbf{v}}}^{n+1}) \otimes (\hat{\mathbf{v}}^{j,n+1} - \overline{\hat{\mathbf{v}}}^{n+1}), \quad \overline{\hat{\mathbf{v}}}^{n+1} = \frac{1}{J}\sum_{k=1}^{J} \hat{\mathbf{v}}^{j,n+1}.$$

It is also easy to check that

$$\mathbf{C}^{n+1} = \begin{bmatrix} \mathbf{C}_{\mathbf{uu}}^{n+1} & \mathbf{C}_{\mathbf{uw}}^{n+1} \\ \mathbf{C}_{\mathbf{uw}}^{n+1\intercal} & \mathbf{C}_{\mathbf{ww}}^{n+1} \end{bmatrix} \tag{5}$$

where, using the definition of Ξ,

$$\mathbf{C}_{\mathbf{uu}}^{n+1} = \frac{1}{J}\sum_{k=1}^{J}(\mathbf{u}^{j,n} - \overline{\mathbf{u}}^n)\otimes(\mathbf{u}^{j,n} - \overline{\mathbf{u}}^n), \quad \mathbf{C}_{\mathbf{uw}}^{n+1} = \frac{1}{J}\sum_{k=1}^{J}(\mathbf{u}^{j,n} - \overline{\mathbf{u}}^n)\otimes(\mathcal{G}(\mathbf{u}^{j,n}) - \overline{\mathcal{G}}^n)$$
$$\mathbf{C}_{\mathbf{ww}}^{n+1} = \frac{1}{J}\sum_{k=1}^{J}(\mathcal{G}(\mathbf{u}^{j,n}) - \overline{\mathcal{G}}^n)\otimes(\mathcal{G}(\mathbf{u}^{j,n}) - \overline{\mathcal{G}}^n)$$

are the empirical covariance matrices depending on the ensemble set $\{\mathbf{u}^{j,n}\}_{j=1}^J$ at iteration n and on $\{\mathcal{G}(\mathbf{u}^{j,n})\}_{j=1}^J$, i.e. the image of $\{\mathbf{u}^{j,n}\}_{j=1}^J$ at iteration n, and where we define by $\overline{\mathbf{u}}^n$ and $\overline{\mathcal{G}}^n$ the mean of $\{\mathbf{u}^{j,n}\}_{j=1}^J$ and $\{\mathcal{G}(\mathbf{u}^{j,n})\}_{j=1}^J$, namely

$$\overline{\mathbf{u}}^n = \frac{1}{J}\sum_{j=1}^{J}\mathbf{u}^{j,n}, \quad \overline{\mathcal{G}}^n = \frac{1}{J}\sum_{j=1}^{J}\mathcal{G}(\mathbf{u}^{j,n}).$$

Remark 2.2. The empirical covariance matrix of the ensembles is in fact computed by information at iteration step n. However, we keep the notation $\mathbf{C}_{\mathbf{uu}}^{n+1}$.

The compromise that is sought should have $\mathbf{v}^{j,n+1}$ close to $\hat{\mathbf{v}}^{j,n+1}$ and $\mathbf{H}\mathbf{v}^{j,n+1}$ close to $\mathbf{y}^{n+1}$. Then $\mathbf{v}^{j,n+1}$ is defined by the solution of the following minimization problem:

$$\mathbf{v}^{j,n+1} = \arg\min_{\mathbf{v}} \mathcal{J}^{j,n}(\mathbf{v}) := \frac{1}{2}\|\mathbf{y}^{n+1} - \mathbf{H}\mathbf{v}\|^2_{\mathbf{\Gamma}^{-1}} + \frac{1}{2}\|\mathbf{v} - \hat{\mathbf{v}}^{j,n+1}\|^2_{\mathbf{C}^{n+1}}. \tag{6}$$

Here, we recall that $\mathbf{\Gamma}^{-1} \in \mathbb{R}^{K\times K}$ is the covariance matrix characterizing the uncertainties on data $\mathbf{y}^{n+1}$ and that we assume $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Gamma}^{-1})$. Throughout the work, we also assume that $\mathbf{\Gamma}^{-1}$ is positive definite. We notice that the first term of $\mathcal{J}^{j,n}(\mathbf{v})$ corresponds to the least squares functional Φ given by

$$\Phi(\mathbf{u}, \mathbf{y}) := \frac{1}{2}\|\mathbf{\Gamma}^{\frac{1}{2}}(\mathbf{y} - \mathcal{G}(\mathbf{u}))\|^2. \tag{7}$$

Further, $\mathcal{J}^{j,n}(\mathbf{v})$ can be written as

$$\mathcal{J}^{j,n}(\mathbf{v}) = \frac{1}{2}\mathbf{v}^{\intercal}\left(\mathbf{H}^{\intercal}\mathbf{\Gamma}\mathbf{H} + \mathbf{C}^{n+1^{-1}}\right)\mathbf{v} - \left(\mathbf{C}^{n+1^{-\intercal}}\hat{\mathbf{v}}^{j,n+1} + \mathbf{H}^{\intercal}\mathbf{\Gamma}^{\intercal}\mathbf{y}^{n+1}\right)^{\intercal}\mathbf{v} + \overline{\mathcal{J}} \tag{8}$$

where $\overline{\mathcal{J}}$ collects all the terms independent of $\mathbf{v}$. The EnKF update formula is derived by imposing first order necessary condition $\nabla_{\mathbf{v}}\mathcal{J}^{j,n}(\mathbf{v}) = \mathbf{0}$. For an extensive discussion, we refer e.g. to [16, 19]. In particular, we point-out here that the derivation of the EnKF update formula through (8) requires that the empirical covariance $\mathbf{C}^{n+1}$ is positive definite. This is a strong assumption which cannot be guaranteed in a general fashion and it is usually overcame by introducing a shifting of $\mathbf{C}^{n+1}$ as $\mathbf{C}_{\varepsilon}^{n+1} = \mathbf{C}^{n+1} + \varepsilon\mathbf{I}$, where $\mathbf{I} \in \mathbb{R}^{(d+K)\times(d+K)}$ is the identity matrix, and study the limit of $\varepsilon \to 0$. We refer to [16]. For the derivation of the EnKF update formula for constrained problems we will follow this approach, see proposition 2.5 in the following section.

We recall here that in order to help the reader, we report a list of the main mathematical symbols in appendix B.

2.1. Equality constraints on the control space

We consider imposing equality constraints in the space of the control and formulate those as

$$\mathcal{A}(\mathbf{u}) = \mathbf{0}_{\mathbb{R}^m} \tag{9}$$

where $\mathcal{A}$ is the vector valued differentiable operator $\mathcal{A} : \mathbf{u} \in \mathbb{R}^d \mapsto \mathcal{A}(\mathbf{u}) \in \mathbb{R}^m$, containing the $m \leqslant d$ constraint values.

Remark 2.3. Throughout the paper, we use the subscript (ℓ), $\ell = 1, \ldots, N$, to denote the ℓth component of vectors and vector-valued functions on $\mathbb{R}^N$.

Let $\mathbf{J}_{\mathcal{A}} = \left[\nabla_{\mathbf{u}}\mathcal{A}_{(1)}, \ldots, \nabla_{\mathbf{u}}\mathcal{A}_{(m)}\right]^{\intercal}$ be the $m \times d$ Jacobian matrix of the operator $\mathcal{A}$. In order to satisfy constraint qualification for equality constraints, we require that if $\mathbf{u}^*$ is a feasible point then $\mathbf{J}_{\mathcal{A}}(\mathbf{u}^*)$ has full rank, i.e. m. In other words, $\mathbf{u}^*$ is a regular point of the constraint.

Theorem 2.4 (*see* [20, 21]). *If $\mathbf{u}^*$ is a feasible point and $\mathbf{J}_{\mathcal{A}}(\mathbf{u}^*)$ is a full rank matrix, the constraint qualification LICQ holds at $\mathbf{u}^*$. Hence, the Lagrange multipliers theorem gives necessary optimality condition.*

In order to formulate the constrained optimization problem associated to the minimization of (8), we define also the extension of the operator $\mathcal{A}$ to the space $\mathbb{R}^{d+K}$ as $\tilde{\mathcal{A}} : \mathbf{v} \in \mathbb{R}^{d+K} \mapsto \tilde{\mathcal{A}}(\mathbf{v}) = \mathcal{A}(\mathbf{H}^{\perp}\mathbf{v}) \in \mathbb{R}^m$. Then observe that the Jacobian $\mathbf{J}_{\tilde{\mathcal{A}}} \in \mathbb{R}^{m\times(d+K)}$ of the operator $\tilde{\mathcal{A}}$ is $\mathbf{J}_{\tilde{\mathcal{A}}} = [\mathbf{J}_{\mathcal{A}}, \mathbf{0}]$ with $\mathbf{0} \in \mathbb{R}^{m\times K}$ since $\partial_{\mathbf{v}_{(\ell)}}\mathcal{A}_{(k)} = 0$, for $\ell = d+1, \ldots, d+K$ and $k = 1, \ldots, m$. Moreover, if $\mathbf{u}^*$ is feasible and $\mathbf{J}_{\mathcal{A}}(\mathbf{u}^*)$ has rank m, then $\mathbf{J}_{\tilde{\mathcal{A}}}(\mathbf{u}^*)$ has also rank m.

The optimization step becomes then

$$\min_{\mathbf{v}} \mathcal{J}^{j,n}(\mathbf{v}), \quad \text{subject to } \tilde{\mathcal{A}}(\mathbf{v}) = \mathbf{0}_{\mathbb{R}^m}, \quad j = 1, \ldots, J. \tag{10}$$

It is clear that (10) requires to solve J constrained optimization problems sequentially, at each iteration step n. We point out that the objective functional (8) is convex.

Proposition 2.5. *Let $\mathbf{v}^{j,n+1}$ be an optimal solution to the constrained optimization problem (10) for a given j and n, satisfying the constraint qualification for the differentiable equality constraint defined by $\tilde{\mathcal{A}}$. Then $\mathbf{u}^{j,n+1} = \mathbf{H}^{\perp}\mathbf{v}^{j,n+1}$ satisfies the first order necessary optimality conditions*

$$\begin{aligned}\mathbf{u}^{j,n+1} =& \mathbf{u}^{j,n} + \mathbf{C}_{\mathbf{uw}}^{n+1}(\mathbf{C}_{\mathbf{ww}}^{n+1} + \mathbf{\Gamma}^{-1})^{-1}(\mathbf{y}^{n+1} - \mathcal{G}(\mathbf{u}^{j,n})) + \mathbf{C}_{\mathbf{uw}}^{n+1}(\mathbf{C}_{\mathbf{ww}}^{n+1} + \mathbf{\Gamma}^{-1})^{-1} \\ &\times \mathbf{C}_{\mathbf{uw}}^{n+1\intercal}\mathbf{J}_{\mathcal{A}}^{\intercal}(\mathbf{u}^{j,n+1})\boldsymbol{\lambda}^{j,n+1} - \mathbf{C}_{\mathbf{uu}}^{n+1}\mathbf{J}_{\mathcal{A}}^{\intercal}(\mathbf{u}^{j,n+1})\boldsymbol{\lambda}^{j,n+1} \\ \mathcal{A}(\mathbf{u}^{j,n+1}) =& \mathbf{0}_{\mathbb{R}^m}.\end{aligned} \tag{11}$$

Proof. We apply Lagrange multiplier technique to the objective $\mathcal{J}_{\varepsilon}^{j,n}(\mathbf{v})$ defined by (8) with shifting of the empirical covariance $\mathbf{C}^{n+1} \mapsto \mathbf{C}_{\varepsilon}^{n+1} := \mathbf{C}^{n+1} + \varepsilon\mathbf{I}$, where $\mathbf{I} \in \mathbb{R}^{(d+k)\times(d+K)}$ is the identity matrix and $\varepsilon > 0$. Then we study the limit $\varepsilon \to 0$. We observe that, thanks to the shifted matrix, $\mathbf{H}^{\intercal}\mathbf{\Gamma}\mathbf{H} + \mathbf{C}_{\varepsilon}^{n+1^{-1}}$ is strictly positive definite. Consequently, the objective $\mathcal{J}_{\varepsilon}^{j,n}$ is strongly convex and there is existence and uniqueness of the solution to the minimization problem.

Detailed computations required in the following proof are given in appendix A.

The constrained minimum of $\mathcal{J}_{\varepsilon}^{j,n}(\mathbf{v})$ corresponds to a stationary point of the Lagrangian function

$$\mathfrak{L}(\mathbf{v}, \boldsymbol{\lambda}) = \mathcal{J}_{\varepsilon}^{j,n}(\mathbf{v}) + \boldsymbol{\lambda}^{\intercal}\tilde{\mathcal{A}}(\mathbf{v})$$

where the vector $\boldsymbol{\lambda} \in \mathbb{R}^m$ contains the Lagrange multipliers. Variations of the Lagrangian function $\mathfrak{L}$ with respect to $\mathbf{v}$ and $\boldsymbol{\lambda}$ are

$$\nabla_{\mathbf{v}}\mathfrak{L}(\mathbf{v},\boldsymbol{\lambda}) = \left(\mathbf{H}^\intercal\boldsymbol{\Gamma}\mathbf{H} + \mathbf{C}_\varepsilon^{n+1^{-1}}\right)\mathbf{v} - \left(\mathbf{H}^\intercal\boldsymbol{\Gamma}^\intercal\mathbf{y}^{n+1} + \mathbf{C}_\varepsilon^{n+1^{-1}}\hat{\mathbf{v}}^{j,n+1}\right) + (\boldsymbol{\lambda}^\intercal\mathbf{J}_{\tilde{\mathcal{A}}}(\mathbf{v}))^\intercal$$
$$\nabla_{\boldsymbol{\lambda}}\mathfrak{L}(\mathbf{v},\boldsymbol{\lambda}) = \tilde{\mathcal{A}}(\mathbf{v}).$$

If $(\mathbf{v}^{j,n+1},\boldsymbol{\lambda}^{j,n+1}) \in \mathbb{R}^{d+K}\times\mathbb{R}^m$ is an optimal solution then it satisfies $\nabla_{\mathbf{v}}\mathfrak{L}(\mathbf{v}^{j,n+1},\boldsymbol{\lambda}^{j,n+1}) = \mathbf{0}_{\mathbb{R}^{d+K}}$ and $\nabla_{\boldsymbol{\lambda}}\mathfrak{L}(\mathbf{v}^{j,n+1},\boldsymbol{\lambda}^{j,n+1}) = \mathbf{0}_{\mathbb{R}^m}$, that are the first order necessary optimality conditions.

Due to the assumption of a regular point, the system of optimality conditions is determined. In particular, imposing $\nabla_{\mathbf{v}}\mathfrak{L}(\mathbf{v}^{j,n+1},\boldsymbol{\lambda}^{j,n+1}) = \mathbf{0}_{\mathbb{R}^{d+K}}$, it is possible to determine

$$\begin{aligned}\mathbf{v}^{j,n+1} &= \left(\mathbf{H}^\intercal\boldsymbol{\Gamma}\mathbf{H} + \mathbf{C}_\varepsilon^{n+1^{-1}}\right)^{-1}\left(\mathbf{H}^\intercal\boldsymbol{\Gamma}\mathbf{y}^{n+1} + \mathbf{C}_\varepsilon^{n+1^{-1}}\hat{\mathbf{v}}^{j,n+1}\right)\\ &\quad - \left(\mathbf{H}^\intercal\boldsymbol{\Gamma}\mathbf{H} + \mathbf{C}_\varepsilon^{n+1^{-1}}\right)^{-1}\mathbf{J}_{\tilde{\mathcal{A}}}^\intercal(\mathbf{v}^{j,n+1})\boldsymbol{\lambda}^{j,n+1}.\end{aligned} \tag{12}$$

Now we apply the Woodbury matrix identity to the term multiplying $\hat{\mathbf{v}}^{j,n+1}$ and to the last term in (12), cf appendix A. Instead, for the term multiplying $\mathbf{y}^{n+1}$ in (12) we observe that

$$\left(\mathbf{H}^\intercal\boldsymbol{\Gamma}\mathbf{H} + \mathbf{C}_\varepsilon^{n+1^{-1}}\right)^{-1}\mathbf{H}^\intercal\boldsymbol{\Gamma} = \mathbf{C}_\varepsilon^{n+1}\mathbf{H}^\intercal\left(\mathbf{H}\mathbf{C}_\varepsilon^{n+1}\mathbf{H}^\intercal + \boldsymbol{\Gamma}^{-1}\right)^{-1}.$$

Then we get, cf appendix A,

$$\begin{aligned}\mathbf{v}^{j,n+1} &= \hat{\mathbf{v}}^{j,n+1} + \mathbf{C}_\varepsilon^{n+1}\mathbf{H}^\intercal\left(\mathbf{H}\mathbf{C}_\varepsilon^{n+1}\mathbf{H}^\intercal + \boldsymbol{\Gamma}^{-1}\right)^{-1}(\mathbf{y}^{n+1} - \mathbf{H}\hat{\mathbf{v}}^{j,n+1})\\ &\quad - \left(\mathbf{C}_\varepsilon^{n+1} - \mathbf{C}_\varepsilon^{n+1}\mathbf{H}^\intercal\left(\mathbf{H}\mathbf{C}_\varepsilon^{n+1}\mathbf{H}^\intercal + \boldsymbol{\Gamma}^{-1}\right)^{-1}\mathbf{H}\mathbf{C}_\varepsilon^{n+1}\right)\mathbf{J}_{\tilde{\mathcal{A}}}^\intercal(\mathbf{v}^{j,n+1})\boldsymbol{\lambda}^{j,n+1}.\end{aligned} \tag{13}$$

If we multiply the previous equation by the observation matrix $\mathbf{H}^\perp$ and compute the limit $\varepsilon \to 0$, we obtain the update formula for the control, cf appendix A:

$$\begin{aligned}\mathbf{u}^{j,n+1} &= \mathbf{u}^{j,n} + \mathbf{C}_{\mathbf{uw}}^{n+1}(\mathbf{C}_{\mathbf{ww}}^{n+1} + \boldsymbol{\Gamma}^{-1})^{-1}(\mathbf{y}^{n+1} - \mathcal{G}(\mathbf{u}^{j,n}))\\ &\quad + \mathbf{C}_{\mathbf{uw}}^{n+1}(\mathbf{C}_{\mathbf{ww}}^{n+1} + \boldsymbol{\Gamma}^{-1})^{-1}\mathbf{C}_{\mathbf{uw}}^{n+1^\intercal}\mathbf{J}_{\mathcal{A}}^\intercal(\mathbf{u}^{j,n+1})\boldsymbol{\lambda}^{j,n+1} - \mathbf{C}_{\mathbf{uu}}^{n+1}\mathbf{J}_{\mathcal{A}}^\intercal(\mathbf{u}^{j,n+1})\boldsymbol{\lambda}^{j,n+1},\end{aligned}$$

coupled to $\mathcal{A}\left(\mathbf{u}^{j,n+1}\right) = \mathbf{0}_{\mathbb{R}^m}$, which provide a set of necessary optimality conditions for the constrained optimization problem (10). □

Observe that control update (11) is the classical ensemble Kalman filter update formula, given by the first two terms in the right-hand side, perturbed by the last two terms due to the constraint (9). Moreover, (11) depends also on the multipliers and their values must be determined as part of the solution.

Corollary 2.6. *Consider the assumptions of proposition 2.5. If the feasible set is convex, then the optimization problem (10) is convex, and this implies that the necessary optimality conditions (11) are also sufficient. In particular, this holds true for affine equality constraints.*

2.1.1. A game theory viewpoint and existence results. Proposition 2.5 is stated for a fixed ensemble member j at a fixed iteration n, by assuming that the other ensembles play a role of parameters in the optimization problem. This is closely related to the fact that problem (10) can be seen from a game theoretic point of view. In fact, if we interpret each ensemble member $j \in \{1, \ldots, J\}$ as a player which chooses a control variable $\mathbf{v}^{j,n+1}$ in its set of feasible controls given by the constraint $\tilde{\mathcal{A}}\left(\mathbf{v}^{j,n+1}\right) = \mathbf{0}$ while seeking to minimize its payoff function $\mathcal{J}^{j,n}$ which depends, via the covariance matrix $\mathbf{C}^{n+1}$, on all other players' controls at iteration n, i.e. $\mathbf{v}^{-j,n}$, cf remark 2.2. Here, we indicate $\mathbf{v}^{-j,n+1} = \{\mathbf{v}^{k,n+1} : k = 1, \ldots, J, k \neq j\}$. Observe that each player has no knowledge on the strategy adopted by other players. Moreover, we notice that we are in presence of a repeated game.

In a non-cooperative and simultaneous game, we recall that $\{\mathbf{v}^{j,n+1} : j = 1, \ldots, J\}$ represents a Nash equilibrium if for every $\mathbf{v}$ satisfying the constraint one has $\mathcal{J}^{j,n}\left(\mathbf{v}, \mathbf{v}^{-j,n+1}\right) \geqslant \mathcal{J}^{j,n}\left(\mathbf{v}^{j,n+1}, \mathbf{v}^{-j,n+1}\right)$, for $j = 1, \ldots, J$.

In general, a Nash equilibrium may not exist and need not to be unique. Conditions for the existence of Nash equilibrium are recalled in the following theorem.

Theorem 2.7 (*Nikaidô and Isoda* [22]). *Let $\Gamma = \{X_i, f_i\}_{i=1}^N$ be a game with nonempty, compact and convex strategy sets $X_i \subset \mathbb{R}^{n_i}$ and continuous payoff functions $f_i : X_i \to \mathbb{R}$, which are quasiconvex in x^i for every fixed x^{-i} for all $i = 1, \ldots, N$. Then Γ has (at least) one Nash equilibrium.*

We recall that quasiconvexity is implied by convexity. Then the Nikaidô and Isoda theorem is satisfied for problem (10) if the feasible set is convex. In fact, the payoff function (8) is convex since $\mathbf{H}^\intercal \mathbf{\Gamma} \mathbf{H} + \mathbf{C}^{n+1^{-1}}$ is positive semidefinite, and consequently $\mathcal{J}^{j,n}$ is quasiconvex for each $j = 1, \ldots, J$. In particular, there exists at least one equilibria for affine equality constraints. For further references, we refer to [23–25].

2.2. The case of linear equality constraints

Let us assume that the constraint $\mathcal{A}$ on the control space is linear. Then we can write $\mathcal{A}(\mathbf{u}) = \mathbf{A}\mathbf{u}$, with $\mathbf{A} \in \mathbb{R}^{m \times d}$. In this case we need to assume that $\mathbf{A} \in \mathbb{R}^{m \times d}$ is a full row rank matrix in order to have satisfied the condition of regular point. The linearity of the equality constraint guarantees an expression to evaluate the multipliers explicitly by substituting the update formula given in (11) into the constraint (9). The following result holds true in the case of linear equality constraints.

Proposition 2.8. *Let $\mathcal{A} : \mathbb{R}^d \to \mathbb{R}^m$ be a linear operator and $\{\mathbf{u}^{j,0}\}_{j=1}^J$ be a set of initial ensembles such that $\mathcal{A}\left(\mathbf{u}^{j,0}\right) = \mathbf{0}_{\mathbb{R}^m}$. Let $\{\mathbf{u}^{j,n+1}\}_{j=1}^J$ be the ensemble set computed via the unconstrained ensemble Kalman filter update. Then $\mathcal{A}\left(\mathbf{u}^{j,n+1}\right) = \mathbf{0}_{\mathbb{R}^m}$, $\forall j = 1, \ldots, J$.*

Proof. We prove the statement by induction on n. Notice that $\mathbf{J}_{\mathcal{A}} = \mathbf{A}^\intercal$ since the constraint is linear. It is easy to check that if $\{\mathbf{u}^{j,n}\}_{j=1}^J$ satisfy the constraint then $\mathbf{C}_{\mathbf{uw}}^{n+1^\intercal}\mathbf{A}^\intercal = \mathbf{0}_{\mathbb{R}^{K \times m}}$ and $\mathbf{C}_{\mathbf{uu}}^{n+1^\intercal}\mathbf{A}^\intercal = \mathbf{0}_{\mathbb{R}^{d \times m}}$. Then the update formula given in (11) reduces to

$$\mathbf{u}^{j,n+1} = \mathbf{u}^{j,n} + \mathbf{C}_{\mathbf{uw}}^{n+1}\left(\mathbf{C}_{\mathbf{ww}}^{n+1} + \mathbf{\Gamma}^{-1}\right)^{-1}\left(\mathbf{y}^{n+1} - \mathcal{G}(\mathbf{u}^{j,n})\right)$$

which represents the unconstrained ensemble Kalman filter formula. Substituting into the constraint we have $\mathbf{A}\mathbf{u}^{j,n+1} = \mathbf{A}\mathbf{u}^{j,n} = \mathbf{0}_{\mathbb{R}^m}$ since also $\mathbf{A}\mathbf{C}_{\mathbf{uw}}^{n+1} = \mathbf{0}_{\mathbb{R}^{m \times K}}$. □

We point-out that the previous result is a direct consequence of the well-known subspace property of the classical unconstrained EnKF formula, which states that the ensemble iterates

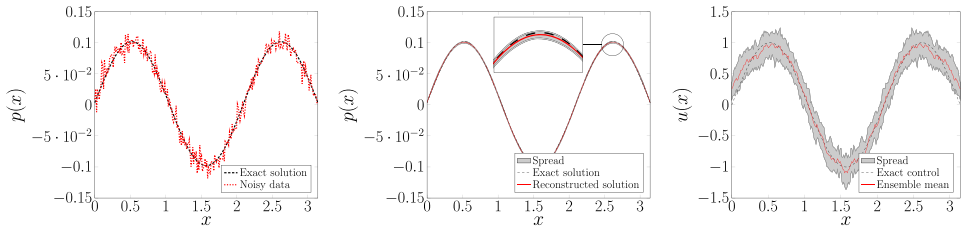

Figure 1. Application of the unconstrained ensemble Kalman filter based on the discrete formulation, assuming that the control function satisfies a linear equality constraint given as symmetry with respect to $x = \frac{\pi}{2}$.

are a linear combination of the initial ensemble members, namely they stay in the subspace spanned by the initial ensemble [19].

We provide an experimental evidence of the result in proposition 2.8. We consider the inverse problem of finding the force function of a linear elliptic equation in one spatial dimension assuming that noisy observation of the solution to the problem are available. This is a standard problem in the mathematical literature on the EnKF for inverse problems, e.g. see [6, 16, 19].

The problem is prescribed by the following one dimensional elliptic PDE

$$-\frac{\mathrm{d}^2}{\mathrm{d}x^2}p(x) + p(x) = u(x), \quad x \in [0, \pi] \tag{14}$$

endowed with boundary conditions $p(0) = p(\pi) = 0$. We assign a continuous exact control $u(x)$, being symmetric with respect to $x = \frac{\pi}{2}$, namely $u(x) = u(-x + \pi)$. Introducing a uniform mesh consisting of $d = K = 2^8$ equidistant points on the interval $[0, \pi]$, we let $\mathbf{u}^\dagger \in \mathbb{R}^d$ be the vector of the evaluations of the control function $u(x)$ on the mesh. Noisy observations $\mathbf{y} \in \mathbb{R}^K$ are simulated as

$$\mathbf{y} = \mathbf{p} + \boldsymbol{\eta} = \mathbf{G}\mathbf{u}^\dagger + \boldsymbol{\eta},$$

where $\mathbf{G} \in \mathbb{R}^{K\times d}$ is the finite difference discretization of the continuous linear operator defining the elliptic PDE (14). For simplicity we assume that $\boldsymbol{\eta}$ is a Gaussian white noise, more precisely $\boldsymbol{\eta} \sim \mathcal{N}(0, \gamma^2\mathbf{I})$ with $\gamma \in \mathbb{R}^+$ and $\mathbf{I} \in \mathbb{R}^{d\times d}$ is the identity matrix. We are interested in recovering the control $\mathbf{u}^\dagger \in \mathbb{R}^d$ from the noisy observations $\mathbf{y} \in \mathbb{R}^K$ only.

Let us consider $u(x) = \sin(3x), \forall x \in [0, \pi]$. In figure 1 we show the solution to this problem provided by the unconstrained ensemble Kalman filter. The initial ensemble set $\{\mathbf{u}^{j,0}\}_{j=1}^J$ is artificially built in order to satisfy symmetry with respect to $x = \frac{\pi}{2}$, after being sampled from a Brownian bridge, e.g. as in [6]. This implies that $m = \frac{d}{2} = 2^7$ constraints are taken into account. We solve the inverse problem by updating the ensemble members with

$$\mathbf{u}^{j,n+1} = \mathbf{u}^{j,n} + \mathbf{C}_{\mathbf{uw}}^{n+1}\left(\mathbf{C}_{\mathbf{ww}}^{n+1} + \boldsymbol{\Gamma}^{-1}\right)^{-1}\left(\mathbf{y} - G(\mathbf{u}^{j,n})\right)$$

for $J = 100$ and a noise level $\gamma = 0.01$. The filter method converges, meeting the discrepancy principle in very few iterations, and allowing each ensemble member and the corresponding mean to satisfy the constraint.

Proposition 2.8 guarantees that the constrained optimization problem (10) is solved by the unconstrained ensemble Kalman filter when the constraint $\mathcal{A}$ is linear, independently on the linearity of the model $\mathcal{G}$. As consequence, analysis and continuous limits of the ensemble

Kalman filter with linear equality constraints coincide with the recent results [4, 6, 7, 10, 11] for unconstrained problems, and in the following we will focus on nonlinear constraints only.

3. Continuous limits of the constrained ensemble Kalman filter

In this section we investigate continuous limits of the constrained version of the ensemble Kalman filter, when nonlinear equality constraints are taken into account. These limits allow us to reformulate the fully discrete filter into the framework of ordinary and partial differential equations, thus making the method computationally simpler and amenable to an analysis of its properties.

3.1. Continuous time limit

In order to compute the continuous time limit equation, we make the following assumptions, see e.g. [6, 11].

Scaling assumption 1. The covariance matrix $\mathbf{\Gamma}^{-1}$, accounting for uncertainties due to data, is scaled by a scalar parameter Δt.

Scaling assumption 2. The multipliers $\boldsymbol{\lambda}^{j,n+1}$ are scaled as $\Delta t\boldsymbol{\lambda}^{j,n+1}$.

We notice that the second assumption is in fact not restrictive since multipliers are unique up to a multiplicative constant.

Proposition 3.1. *Under the scaling assumptions 1 and 2, the constrained ensemble Kalman filter (11) formally converges to the semi-explicit system of differential algebraic equations (DAEs)*

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{dt}}\mathbf{u}^j &= \mathbf{C}_{\mathbf{uw}}\mathbf{\Gamma}\left(\mathbf{y}-\mathcal{G}(\mathbf{u}^j)\right) - \mathbf{C}_{\mathbf{uu}}\mathbf{J}_{\mathcal{A}}^{\intercal}\left(\mathbf{u}^j\right)\boldsymbol{\lambda}^j \\ \mathbf{0}_{\mathbb{R}^m} &= \mathcal{A}\left(\mathbf{u}^j(t)\right), \end{aligned} \tag{15}$$

in the limit $\Delta t \to 0^+$.

Proof. Using the scaling assumptions, the solution (11) to the constrained optimization problem can be reformulated as

$$\begin{aligned} \mathbf{u}^{j,n+1} = \mathbf{u}^{j,n} &+ \Delta t\mathbf{C}_{\mathbf{uw}}^{n+1}(\Delta t\mathbf{C}_{\mathbf{ww}}^{n+1}+\mathbf{\Gamma}^{-1})^{-1}\left(\mathbf{y}^{n+1}-\mathcal{G}(\mathbf{u}^{j,n})\right) \\ &- \Delta t\mathbf{C}_{\mathbf{uu}}^{n+1}\mathbf{J}_{\mathcal{A}}^{\intercal}\left(\mathbf{u}^{j,n+1}\right)\boldsymbol{\lambda}^{j,n+1} + \mathcal{R}\left(\mathbf{u}^{j,n+1},\boldsymbol{\lambda}^{j,n+1}\right) \end{aligned} \tag{16}$$

where $\mathcal{R}$ is a term of order $\mathcal{O}(\Delta t^2)$, in fact

$$\mathcal{R}(\mathbf{u}^{j,n+1},\boldsymbol{\lambda}^{j,n+1}) = \Delta t^2\mathbf{C}_{\mathbf{uw}}^{n+1}(\Delta t\mathbf{C}_{\mathbf{ww}}^{n+1}+\mathbf{\Gamma}^{-1})^{-1}\mathbf{C}_{\mathbf{uw}}^{n+1\intercal}\mathbf{J}_{\mathcal{A}}^{\intercal}(\mathbf{u}^{j,n+1})\boldsymbol{\lambda}^{j,n+1}.$$

Now we interpret the parameter Δt as an artificial time step for the iteration, i.e. we take $\Delta t \sim N_t^{-1}$ where N_t is the maximum number of iterations. Assume then $\mathbf{u}^{j,n} \approx \mathbf{u}^j(n\Delta t)$ and

$\boldsymbol{\lambda}^{j,n} \approx \boldsymbol{\lambda}^j(n\Delta t)$ for $n \geqslant 0$ and $j = 1, \ldots, J$. Computing the limit $\Delta t \to 0^+$, (16) is a first order implicit–explicit approximation of the following system of ordinary differential equations (ODEs):

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{u}^j = \mathbf{C}_{\mathbf{uw}}\boldsymbol{\Gamma}(\mathbf{y} - \mathcal{G}(\mathbf{u}^j)) - \mathbf{C}_{\mathbf{uu}}\mathbf{J}_{\mathcal{A}}^{\intercal}(\mathbf{u}^j)\boldsymbol{\lambda}^j, \quad j = 1, \ldots, J. \tag{17}$$

Imposing that the constraint needs to be satisfied at each time $t > 0$, we obtain the system of DAEs (15) endowed with initial conditions $\mathbf{u}^{j,0} = \mathbf{u}^j(0) \in \mathbb{R}^d$ such that $\mathcal{A}(\mathbf{u}^{j,0}) = \mathbf{0}_{\mathbb{R}^m}$. □

In the previous proposition we have seen (11) as first order implicit–explicit time discretization of (15). However, we stress again the fact that $\mathbf{C}_{\mathbf{uu}}^{n+1}$ is actually computed with information at time level n, and therefore explicitly, while it appears in the term which is treated implicitly. Nevertheless, $\mathbf{C}_{\mathbf{uu}}^{n+1}$ can be written in terms of the average quantities, which typically change in a small time scale, justifying therefore the use of an explicit evaluation.

Corollary 3.2. *Assume that $\mathcal{G} = \mathbf{G}$, with $\mathbf{G} \in \mathcal{L}(\mathbb{R}^d, \mathbb{R}^K)$, with $\mathcal{L}(\mathbb{R}^d, \mathbb{R}^K)$ space of linear operators mapping $\mathbb{R}^d$ to $\mathbb{R}^K$. The DAE system (15) can be written in terms of the gradient of the least squares functional Φ (7), obtaining*

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{u}^j &= -\mathbf{C}_{\mathbf{uu}}(\nabla_{\mathbf{u}}\Phi(\mathbf{u}^j, \mathbf{y}) + \mathbf{J}_{\mathcal{A}}^{\intercal}(\mathbf{u}^j)\boldsymbol{\lambda}^j) \\ \mathbf{0}_{\mathbb{R}^m} &= \mathcal{A}(\mathbf{u}^j(t)). \end{aligned} \tag{18}$$

Proof. Due to the linearity of the model $\mathbf{G}$, we observe that $\mathbf{C}_{\mathbf{uw}} = \mathbf{C}_{\mathbf{uu}}\mathbf{G}^{\intercal}$ and use $\nabla_{\mathbf{u}}\Phi(\mathbf{u}^j, \mathbf{y}) = -\mathbf{G}^{\intercal}\boldsymbol{\Gamma}(\mathbf{y} - \mathbf{G}\mathbf{u}^j)$ to rewrite (15) as (18). □

In view of the discussion in section 2.2, it is easy to show that in the case of a linear equality constraint $\mathcal{A}(\mathbf{u}) = \mathbf{A}\mathbf{u} \in \mathbb{R}^m$, the DAE system (18) reduces to

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{u}^j = -\mathbf{C}_{\mathbf{uu}}\nabla_{\mathbf{u}}\Phi(\mathbf{u}^j, \mathbf{y}) \tag{19}$$

which is the preconditioned gradient flow equation for the least square functional Φ, studied e.g. in [4, 6, 7]. We expect that (19) still allows each ensemble member to satisfy the linear constraint if the initial condition is feasible, again as consequence of the subspace property of the EnKF which holds also in the continuous dynamics [6].

We observe that the differential equation in (18) can be written as

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{u}^j &= -\mathbf{C}_{\mathbf{uu}}\nabla_{\mathbf{u}}\Psi(\mathbf{u}^j, \boldsymbol{\lambda}^j, \mathbf{y}) \\ \Psi(\mathbf{u}^j, \boldsymbol{\lambda}^j, \mathbf{y}) &= \Phi(\mathbf{u}^j, \mathbf{y}) + \sum_{k=1}^{m} \boldsymbol{\lambda}_{(k)}^j \mathcal{A}_{(k)}(\mathbf{u}^j) \end{aligned} \tag{20}$$

and therefore it still has the structure of a preconditioned gradient type flow, where the flow Φ is perturbed along the direction of a linear combination of the constraints. However, while Φ is convex, Ψ is not necessarily convex even if the constraint is since the method lacks information about the sign of the multipliers.

3.1.1. Analysis in the case of a linear model. DAE systems are usually characterized by two indices, namely the perturbation index and the differentiation index. Special DAEs are the Hessenberg-type, having the property that perturbation and differentiation indices coincide. In particular, a DAE is said to be of high order index if their perturbation and differentiation

indices are greater than or equal to 2. Under suitable sufficient conditions it is possible to guarantee that a DAE has indices equal to 1 [26].

Proposition 3.3. *Let $I = [t_0, T] \subset \mathbb{R}$, $t_0 < T$ be a compact time interval. Let $\mathbf{x} : I \to \mathbb{R}^{d_\mathbf{x}}$ and $\mathbf{y} : I \to \mathbb{R}^{d_\mathbf{y}}$. Consider the semi-explicit DAE system*

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}(t) &= \mathbf{f}(t, \mathbf{x}(t), \mathbf{y}(t)), \\ \mathbf{0}_{\mathbb{R}^{d_\mathbf{y}}} &= \mathbf{g}(t, \mathbf{x}(t), \mathbf{y}(t)). \end{aligned}$$

Assume that

(*a*) *$\boldsymbol{f}$ is Lipschitz continuous with respect to $\boldsymbol{x}$ and $\boldsymbol{y}$ with Lipschitz constant $L_\mathbf{f}$ uniformly with respect to t;*

(*b*) *$\boldsymbol{g}$ is continuously differentiable and $\mathbf{J}^\mathbf{y}_\mathbf{g} = [\nabla_\mathbf{y} g_1, \ldots, \nabla_\mathbf{y} g_{d_\mathbf{y}}]^\intercal$ is non-singular and bounded for all $(t, \mathbf{x}, \mathbf{y}) \in I \times \mathbb{R}^{d_\mathbf{x}} \times \mathbb{R}^{d_\mathbf{y}}$.*

Then, the semi-explicit DAE system has perturbation index 1. If additionally the inverse of $\mathbf{J}^\mathbf{y}_\mathbf{g}$ is bounded for all $(t, \mathbf{x}, \mathbf{y}) \in I \times \mathbb{R}^{d_\mathbf{x}} \times \mathbb{R}^{d_\mathbf{y}}$, then the semi-explicit DAE system has differentiation index 1.

Using proposition 3.3, the following results hold true for (18), in which we recall that the dependence of the algebraic equation on the multipliers is given implicitly by the differential variables, namely $\mathcal{A}(\mathbf{u}^j(t))$ is in fact $\mathcal{A}(\mathbf{u}^j(t; \boldsymbol{\lambda}^j))$.

Corollary 3.4. *Consider the semi-explicit DAE system (18) for each fixed $j = 1, \ldots, J$. Assume that condition (b) in proposition 3.3 holds for the vector valued function $\mathcal{A}$ with respect to the algebraic variables $\boldsymbol{\lambda}$. Then, system (18) has perturbation index 1. If additionally the inverse of $\mathbf{J}^{\boldsymbol{\lambda}^j}_\mathcal{A}$ is bounded for all $(\mathbf{u}^j, \boldsymbol{\lambda}^j) \in \mathbb{R}^d \times \mathbb{R}^m$, then system (18) has differentiation index 1.*

Proof. Continuous differentiability of the vector valued function $\mathcal{A}$ with respect to the dynamical variable $\mathbf{u}^j$ gives a sufficient condition to the right-hand side of the dynamical equation in (18) to be Lipschitz continuous with respect to $\mathbf{u}^j$ and $\boldsymbol{\lambda}^j$, for each fixed $j = 1, \ldots, J$. Then, the statement follows as application of proposition 3.3. □

Corollary 3.5. *For each fixed $j = 1, \ldots, J$, consider the DAE system (18) with initial condition $\mathbf{u}^j(0) = \mathbf{u}^{j,0}$. Let $(\tilde{\mathbf{u}}^j, \tilde{\boldsymbol{\lambda}}^j)$ be the solution of the perturbed system*

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\tilde{\mathbf{u}}^j &= -\mathbf{C}_{\tilde{\mathbf{u}}\tilde{\mathbf{u}}}\nabla_{\tilde{\mathbf{u}}}\Psi(\tilde{\mathbf{u}}^j, \tilde{\boldsymbol{\lambda}}^j, \mathbf{y}) + \boldsymbol{\delta}_1(t), \quad \tilde{\mathbf{u}}^j(0) = \tilde{\mathbf{u}}^{j,0} \\ \mathbf{0}_{\mathbb{R}^m} &= \mathcal{A}(\tilde{\mathbf{u}}^j) + \boldsymbol{\delta}_2(t) \end{aligned}$$

on $t \in [0, T] \subset \mathbb{R}$ and with Ψ defined in (20). Then, under the assumptions of corollary 3.4, $\exists L_1, L_2 \geqslant 0$ such that the following bound holds true:

$$\|\mathbf{u}^j(t) - \tilde{\mathbf{u}}^j(t)\| \leqslant \left(\|\mathbf{u}^{j,0} - \tilde{\mathbf{u}}^{j,0}\| + T \left(L_1 L_2 \max_{0 \leqslant \tau \leqslant t} \|\boldsymbol{\delta}_2(\tau)\| + \max_{0 \leqslant \tau \leqslant t} \|\boldsymbol{\delta}_1(\tau)\| \right) \right) \times \exp\left(L_1(1 + L_2)t\right)$$

Proof. The assumptions of corollary 3.4 guarantee that the implicit function theorem can be applied to the algebraic equation $\mathbf{0}_{\mathbb{R}^m} = \mathcal{A}(\tilde{\mathbf{u}}^j) + \boldsymbol{\delta}_2$. Therefore, we can solve for $\boldsymbol{\lambda}^j \in \mathbb{R}^m$, $\forall t \in [0, T]$, $\tilde{\mathbf{u}}^j \in \mathbb{R}^d$, obtaining $\tilde{\boldsymbol{\lambda}}^j = \boldsymbol{\Lambda}(\tilde{\mathbf{u}}^j, \boldsymbol{\delta}_2)$. Moreover, $\boldsymbol{\Lambda}$ is locally Lipschitz continuous with respect to $\tilde{\mathbf{u}}^j$ and $\boldsymbol{\delta}_2$ with Lipschitz constant L_2. Then for the multipliers we get the bound

$$\|\boldsymbol{\lambda}^j(t) - \tilde{\boldsymbol{\lambda}}^j(t)\| = \|\boldsymbol{\Lambda}(\mathbf{u}^j(t), \mathbf{0}_{\mathbb{R}^m}) - \boldsymbol{\Lambda}(\tilde{\mathbf{u}}^j(t), \boldsymbol{\delta}_2)\| \leqslant L_2 \left(\|\mathbf{u}^j(t) - \tilde{\mathbf{u}}^j(t)\| + \|\boldsymbol{\delta}_2(t)\| \right).$$

Let L_1 be the Lipschitz constant of the right-hand side of the dynamical equation. We have

$$\begin{aligned}
\|\mathbf{u}^j(t) - \tilde{\mathbf{u}}^j(t)\| &\leqslant \|\mathbf{u}^{j,0} - \tilde{\mathbf{u}}^{j,0}\| + L_1 \int_0^t \|\mathbf{u}^j(\tau) - \tilde{\mathbf{u}}^j(\tau)\| \\
&\quad + \|\boldsymbol{\lambda}^j(\tau) - \tilde{\boldsymbol{\lambda}}^j(\tau)\| \, \mathrm{d}\tau + \| \int_0^t \boldsymbol{\delta}_1(\tau) \, \mathrm{d}\tau \| \\
&\leqslant \|\mathbf{u}^{j,0} - \tilde{\mathbf{u}}^{j,0}\| + L_1 (1 + L_2) \int_0^t \|\mathbf{u}^j(\tau) - \tilde{\mathbf{u}}^j(\tau)\| \, \mathrm{d}\tau \\
&\quad + L_1 L_2 \int_0^t \|\boldsymbol{\delta}_2(\tau)\| \, \mathrm{d}\tau + \int_0^t \|\boldsymbol{\delta}_1(\tau)\| \, \mathrm{d}\tau.
\end{aligned}$$

Then the statement follows easily by application of the Gronwall's lemma. □

The DAE system (18) is derived by starting from the first order necessary optimality conditions stated in proposition 2.5. We recall that these conditions are also sufficient if the feasible set is convex, see corollary 2.6. Now we study the large time behavior of the solution of the DAE system (18).

Let us introduce the following notation. We define $\overline{\mathbf{m}}(t)$ and $\overline{\overline{\mathbf{m}}}(t)$ be the mean and the second moment of the ensembles, respectively, at time $t \geqslant 0$, namely

$$\overline{\mathbf{m}}(t) = \frac{1}{J} \sum_{j=1}^{J} \mathbf{u}^j(t), \quad \overline{\overline{\mathbf{m}}}(t) = \frac{1}{J} \sum_{j=1}^{J} \mathbf{u}^j(t) \otimes \mathbf{u}^j(t).$$

Further, for each $j = 1, \ldots, J$, we define

$$\mathbf{e}^j(t) = \mathbf{u}^j(t) - \overline{\mathbf{m}}(t), \quad \mathbf{r}^j(t) = \mathbf{u}^j(t) - \mathbf{u}^* \tag{21}$$

be the ensemble spread and the residual to a value $\mathbf{u}^*$, respectively. We observe that the evolution in time of $\overline{\mathbf{m}}$ and $\overline{\overline{\mathbf{m}}}$ are governed by

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\overline{\mathbf{m}}(t) &= \frac{1}{J}\sum_{j=1}^{J}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{u}^j(t) = -\frac{1}{J}\sum_{j=1}^{J}\mathbf{C}_{\mathbf{uu}}\left(\nabla_{\mathbf{u}}\Phi(\mathbf{u}^j,\mathbf{y}) + \mathbf{J}_{\mathcal{A}}^{\intercal}(\mathbf{u}^j)\boldsymbol{\lambda}^j\right)\\
&= -\mathbf{C}_{\mathbf{uu}}\left(\nabla_{\mathbf{u}}\Phi\left(\frac{1}{J}\sum_{j=1}^{J}\mathbf{u}^j,\mathbf{y}\right) + \frac{1}{J}\sum_{j=1}^{J}\mathbf{J}_{\mathcal{A}}^{\intercal}(\mathbf{u}^j)\boldsymbol{\lambda}^j\right)\\
&= -\mathbf{C}_{\mathbf{uu}}\left(\nabla_{\mathbf{u}}\Phi(\overline{\mathbf{m}},\mathbf{y}) + \sum_{k=1}^{m}\frac{1}{J}\sum_{j=1}^{J}\boldsymbol{\lambda}_{(k)}^j\nabla_{\mathbf{u}}\mathcal{A}_{(k)}(\mathbf{u}^j)\right),\\
\frac{\mathrm{d}}{\mathrm{d}t}\overline{\overline{\mathbf{m}}}(t) &= \frac{1}{J}\sum_{j=1}^{J}\mathbf{u}^j(t)\otimes\mathbf{u}^j(t) = \frac{2}{J}\sum_{j=1}^{J}\left(\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{u}^j(t)\right)\otimes\mathbf{u}^j(t)\\
&= -\frac{2}{J}\sum_{j=1}^{J}\mathbf{C}_{\mathbf{uu}}\left(\nabla_{\mathbf{u}}\Phi(\mathbf{u}^j,\mathbf{y}) + \mathbf{J}_{\mathcal{A}}^{\intercal}(\mathbf{u}^j)\boldsymbol{\lambda}^j\right)\otimes\mathbf{u}^j(t)\\
&= -2\mathbf{C}_{\mathbf{uu}}\left(\nabla_{\mathbf{u}}\Phi\left(\frac{1}{J}\sum_{j=1}^{J}\mathbf{u}^j\otimes\mathbf{u}^j,\mathbf{y}\otimes\frac{1}{J}\sum_{j=1}^{J}\mathbf{u}^j\right) + \frac{1}{J}\sum_{j=1}^{J}\mathbf{J}_{\mathcal{A}}^{\intercal}(\mathbf{u}^j)\boldsymbol{\lambda}^j\otimes\mathbf{u}^j\right)\\
&= -2\mathbf{C}_{\mathbf{uu}}\left(\nabla_{\mathbf{u}}\Phi\left(\overline{\overline{\mathbf{m}}},\mathbf{y}\otimes\overline{\mathbf{m}}\right) + \frac{1}{J}\sum_{j=1}^{J}\mathbf{J}_{\mathcal{A}}^{\intercal}(\mathbf{u}^j)\boldsymbol{\lambda}^j\otimes\mathbf{u}^j\right)
\end{aligned}
\tag{22}
$$

where we have used the linearity of $\nabla_{\mathbf{u}}\Phi$. While the evolution in time of $\mathbf{e}^j$ is governed by

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{e}^j(t) &= -\mathbf{C}_{\mathbf{uu}}\left(\mathbf{G}^{\intercal}\boldsymbol{\Gamma}\mathbf{G}\mathbf{e}^j + \mathbf{J}_{\mathcal{A}}^{\intercal}(\mathbf{u}^j)\boldsymbol{\lambda}^j - \frac{1}{J}\sum_{\ell=1}^{J}\mathbf{J}_{\mathcal{A}}^{\intercal}(\mathbf{u}^\ell)\boldsymbol{\lambda}^\ell\right)\\
&= -\mathbf{C}_{\mathbf{uu}}\left(\mathbf{G}^{\intercal}\boldsymbol{\Gamma}\mathbf{G}\mathbf{e}^j + \sum_{k=1}^{m}\boldsymbol{\lambda}_{(k)}^j\nabla_{\mathbf{u}}\mathcal{A}_{(k)}(\mathbf{u}^j) - \sum_{k=1}^{m}\frac{1}{J}\sum_{\ell=1}^{J}\boldsymbol{\lambda}_{(k)}^\ell\nabla_{\mathbf{u}}\mathcal{A}_{(k)}(\mathbf{u}^\ell)\right),
\end{aligned}
\tag{23}
$$

$j = 1,\ldots,J$, respectively. The covariance matrix can be written in terms of the mean and the second moment of the ensembles as

$$
\mathbf{C}_{\mathbf{uu}}(t) : \left(\mathbf{u}^1(t),\ldots,\mathbf{u}^J(t)\right) \in \mathbb{R}^d\times\cdots\times\mathbb{R}^d \longmapsto \overline{\overline{\mathbf{m}}}(t) - \overline{\mathbf{m}}(t)\otimes\overline{\mathbf{m}}(t) \in \mathbb{R}^{d\times d}.
\tag{24}
$$

Equations (22) and (23) are coupled with J algebraic equations imposing the validity of the constraint for each ensemble member. In particular, we stress the fact that the systems for the ensemble mean, second moment and spread are not closed, due to the presence of the multipliers and since the constraint is nonlinear and needs to be satisfied by each ensemble member. This requires the knowledge of the evolution in time of the ensembles. Further, it is not ensured that the mean of the ensemble satisfies the constraint.

We prove the following results for the residuals in the space of the control.

Proposition 3.6. *Let $\mathbf{u}^{j,0} \in \mathbb{R}^d$ be an initial condition of the DAE system (18) such that $\mathcal{A}(\mathbf{u}^{j,0}) = \mathbf{0}_{\mathbb{R}^m}$, for all $j = 1, \dots, J$. Then there exists a steady state $(\mathbf{u}^{j,\infty}, \boldsymbol{\lambda}^{j,\infty}) \in \mathbb{R}^d \times \mathbb{R}^m$ of the DAE dynamics which is a KKT point of the minimization problem*

$$\min_{\mathbf{u}\in\mathbb{R}^d} \Phi(\mathbf{u}, \mathbf{y}) \quad \textit{subject to } \mathcal{A}(\mathbf{u}) = \mathbf{0}_{\mathbb{R}^m}$$

for a given $\mathbf{y} \in \mathbb{R}^K$. Moreover, if $\mathcal{A}(\mathbf{u}) = \mathbf{0}_{\mathbb{R}^m}$ is a convex feasible domain, then $\mathbf{u}^{j,\infty}$ provides a first order approximation of $\mathbf{u}^*$ being an optimal solution of the minimization problem.

Proof. We have that a steady state $(\mathbf{u}^{j,\infty}, \boldsymbol{\lambda}^{j,\infty})$ of the DAE system (18) solves

$$\mathbf{0}_{\mathbb{R}^d} = \nabla_{\mathbf{u}}\Phi(\mathbf{u}^{j,\infty}, \mathbf{y}) + \sum_{k=1}^{m} \boldsymbol{\lambda}_{(k)}^{j,\infty} \nabla_{\mathbf{u}}\mathcal{A}_{(k)}(\mathbf{u}^{j,\infty}), \quad \mathbf{0}_{\mathbb{R}^m} = \mathcal{A}(\mathbf{u}^{j,\infty}).$$

For a given $\mathbf{y} \in \mathbb{R}^K$, this is a KKT system and they are the first order necessary optimality conditions for $\mathbf{u}^{j,\infty}$ as solution to the minimization problem $\min_{\mathbf{u}\in\mathbb{R}^d}\Phi(\mathbf{u}, \mathbf{y})$ subject to $\mathcal{A}(\mathbf{u}) = \mathbf{0}_{\mathbb{R}^m}$. If the set $\mathcal{A}(\mathbf{u}) = \mathbf{0}_{\mathbb{R}^m}$ is convex, the KKT conditions are also sufficient. Then $\mathbf{u}^{j,\infty}$ is a solution of the minimization problem. □

Since the typical estimator of the EnKF is provided by the mean of the ensemble, we discuss properties of the ensemble mean.

Proposition 3.7. *Let $\mathbf{u}^0 = \{\mathbf{u}^{j,0}\}_{j=1}^J$ be an initial condition of the DAE system (18) such that $\mathcal{A}(\mathbf{u}^{j,0}) = \mathbf{0}_{\mathbb{R}^m}$, for all $j = 1, \dots, J$. Then there exists a steady state $(\overline{\mathbf{m}}^\infty, \boldsymbol{\lambda}^{1,\infty}, \dots, \boldsymbol{\lambda}^{J,\infty}) \in \mathbb{R}^d \times \mathbb{R}^m \times \cdots \times \mathbb{R}^m$ of the dynamical system (22) of the first moment $\overline{\mathbf{m}}$ which is a KKT point of the minimization problem*

$$\min_{\mathbf{u}^1,\dots,\mathbf{u}^J} \Phi(\overline{\mathbf{m}}, \mathbf{y}) \quad \textit{subject to } \mathcal{A}(\mathbf{u}^1) = \cdots = \mathcal{A}(\mathbf{u}^J) = \mathbf{0}_{\mathbb{R}^m}$$

for a given $\mathbf{y} \in \mathbb{R}^K$. Moreover, if $\mathcal{A}(\mathbf{u}) = \mathbf{0}_{\mathbb{R}^m}$ is a convex feasible domain, then $\overline{\mathbf{m}}^\infty$ provides a first order approximation of an optimal solution of the minimization problem.

Proof. We have that a steady state $(\overline{\mathbf{m}}^\infty, \boldsymbol{\lambda}^{1,\infty}, \dots, \boldsymbol{\lambda}^{J,\infty})$ of (22) solves

$$\mathbf{0}_{\mathbb{R}^d} = \nabla_{\mathbf{u}}\Phi(\overline{\mathbf{m}}^\infty, \mathbf{y}) + \sum_{k=1}^{m} \frac{1}{J}\sum_{j=1}^{J} \boldsymbol{\lambda}_{(k)}^{j,\infty} \nabla_{\mathbf{u}}\mathcal{A}_{(k)}(\mathbf{u}^{j,\infty}), \quad \mathbf{0}_{\mathbb{R}^m} = \mathcal{A}(\mathbf{u}^{J,\infty}) = \cdots = \mathcal{A}(\mathbf{u}^{1,\infty})$$

where the $\mathbf{u}^{j,\infty}$'s are the steady states of (18). For a given $\mathbf{y} \in \mathbb{R}^K$, this is a KKT system and they are the first order necessary optimality conditions for $\overline{\mathbf{m}}^\infty$ as solution to the minimization problem $\min_{\mathbf{u}}\Phi(\overline{\mathbf{m}}, \mathbf{y})$ subject to $\mathcal{A}(\mathbf{u}^1) = \cdots = \mathcal{A}(\mathbf{u}^J) = \mathbf{0}$. If the set $\mathcal{A}(\mathbf{u}) = \mathbf{0}_{\mathbb{R}^m}$ is convex, the KKT conditions are also sufficient. Then $\overline{\mathbf{m}}^\infty$ is a solution of the minimization problem. □

Remark 3.8. The optimization problems in propositions 3.6 and 3.7 are the same if the constraint is linear.

From proposition 3.7 it is clear that the steady state of the mean of the ensembles depends on the steady states of the ensembles. Moreover, proposition 3.7 does not guarantee that the mean of the ensembles at equilibrium satisfies the constraint. A sufficient condition to guarantee that

$\mathcal{A}(\overline{\mathbf{m}}^{\infty}) = \mathbf{0}_{\mathbb{R}^m}$ is that the KKT point belongs to the set $\mathbf{C}_{\mathbf{uu}}^{\infty} = \mathbf{0}_{\mathbb{R}^{d\times d}}$, which is a set of possible equilibria for (18) and (22). In fact, we note that a concentration of the particles at any point of $\mathbb{R}^d$ is a stationary solution of the dynamics. In this case $\|\mathbf{e}^{j,\infty}\| = 0$, for all $j = 1, \ldots, J$. Therefore the question whether or not all the equilibria are necessary in the kernel of $\mathbf{C}_{\mathbf{uu}}^{\infty}$ is to be discussed. This is true in the unconstrained ensemble Kalman filter as proved in [11].

The following counterexample shows that not all the equilibrium solutions of (18) belong to the set $\mathbf{C}_{\mathbf{uu}} = \mathbf{0}_{\mathbb{R}^{d\times d}}$. We consider the case of a one-dimensional control and two ensembles $u, v \in \mathbb{R}$, so that $d = 1$ and $J = 2$. In addition, we take a scalar quadratic and convex constraint $\mathcal{A}(u) = \frac{1}{2}h_1u^2 + h_2u$, with $h_1 > 0$. In order to study the steady state of the mean and the second moment of the ensembles, we need to couple the evolution equation (22) with the system of DAEs (18), obtaining

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}u &= -C_{uv}\left(-G^{\intercal}\Gamma(y - Gu) + h_1\lambda u + h_2\lambda\right),\\
\frac{\mathrm{d}}{\mathrm{d}t}v &= -C_{uv}\left(-G^{\intercal}\Gamma(y - Gv) + h_1\mu v + h_2\mu\right),\\
0 &= \mathcal{A}(u) = \mathcal{A}(v),\\
\frac{\mathrm{d}}{\mathrm{d}t}\overline{m} &= -C_{uv}\left(-G^{\intercal}\Gamma(y - G\overline{m}) + \frac{h_1}{2}(\lambda u + \mu v) + h_2(\lambda + \mu)\right),\\
\frac{\mathrm{d}}{\mathrm{d}t}\overline{\overline{m}} &= -2C_{uv}\left(-G^{\intercal}\Gamma(y\overline{m} - G\overline{\overline{m}}) + \frac{h_1}{2}\left(\lambda u^2 + \mu v^2\right) + h_2(\lambda u + \mu v)\right).
\end{aligned} \tag{25}$$

Here, λ and μ are the two multipliers related to u and v, respectively. We compute the steady states of (25) by solving $\dot{u} = 0$, $\dot{v} = 0$, $\mathcal{A}(u) = 0$, $\mathcal{A}(v) = 0$, $\dot{\overline{m}} = 0$, $\dot{\overline{\overline{m}}} = 0$ obtaining the nullclines in the phase space $(u, v, \lambda, \mu, \overline{m}, \overline{\overline{m}})$. Solutions are provided by either $C_{uv} = \overline{\overline{m}} - \overline{m}^2 = 0$ or

$$\begin{aligned}
u &= (G^{\intercal}\Gamma G + h_1\lambda)^{-1}(G^{\intercal}\Gamma y - h_2\lambda),\\
v &= (G^{\intercal}\Gamma G + h_1\mu)^{-1}(G^{\intercal}\Gamma y - h_2\mu),\\
\overline{m} &= (G^{\intercal}\Gamma G)^{-1}\left(G^{\intercal}\Gamma y - \frac{h_1}{2}(\lambda u + \mu v) - \frac{h_2}{2}(\lambda + \mu)\right),\\
\overline{\overline{m}} &= (G^{\intercal}\Gamma G)^{-1}\left(G^{\intercal}\Gamma y\overline{m} - \frac{h_1}{2}\left(\lambda u^2 + \mu v^2\right) - \frac{h_2}{2}(\lambda u + \mu v)\right),
\end{aligned} \tag{26}$$

where λ and μ are solutions $\mathcal{A}(u) = 0$ and $\mathcal{A}(v) = 0$, respectively. The equilibrium or fixed points are the intersections of the nullclines and therefore the question becomes whether or not $(\overline{m}, \overline{\overline{m}})$ always belongs to the set where $\overline{\overline{m}} - \overline{m}^2 = 0$. It is easy to observe that this is possible if and only if $\lambda = \mu$, since then

$$\begin{aligned}
\overline{m} &= (G^{\intercal}\Gamma G + h_1\lambda)^{-1}(G^{\intercal}\Gamma y - h_2\lambda),\\
\overline{\overline{m}} &= (G^{\intercal}\Gamma G + h_1\lambda)^{-1}(G^{\intercal}\Gamma y - h_2\lambda)\overline{m}.
\end{aligned}$$

We can then conclude that this counterexample shows that, even in the simplest one-dimensional setting, the constrained EnKF provides feasible solutions to the constrained

optimization problem which do not collapse. This consideration opens the question on guaranteeing that the collapse of the ensemble to the mean occurs, i.e. the equilibria lie on the set where $\|\mathbf{e}^{j,\infty}\| = 0$, for $j = 1, \ldots, J$.

A sufficient condition for the existence of a monotonic decay to a bound for the ensemble spread is guaranteed by the following result, which holds in the space of the control.

Proposition 3.9. *Let $\mathbf{u}^{j,0} \in \mathbb{R}^d$, $j = 1, \ldots, J$, be an initial condition of the DAE system (18) such that $\mathcal{A}(\mathbf{u}^{j,0}) = \mathbf{0}_{\mathbb{R}^m}$. Assume that $\mathcal{A} : \mathbb{R}^d \to \mathbb{R}^m$ is convex and quadratic. Then the quantity $\frac{1}{J}\sum_{j=1}^{J} \|\mathbf{e}^j(t)\|^2$ is decreasing in time and thus in particular we have $\frac{1}{J}\sum_{j=1}^{J} \|\mathbf{e}^j(t)\|^2 \leqslant \frac{1}{J}\sum_{j=1}^{J} \|\mathbf{e}^j(0)\|^2$, for $t \geqslant 0$, provided that $\boldsymbol{\lambda}^j(t) = \boldsymbol{\Lambda}(t) \geqslant 0$, $\forall j = 1, \ldots, J$ and $t > 0$.*

Proof. For the sake of simplicity we consider a scalar constraint, i.e. $m = 1$. To prove the statement, it is sufficient to show that $\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{J}\sum_{j=1}^{J} \|\mathbf{e}^j(t)\|^2 \leqslant 0$. Using (23), we compute

$$\begin{aligned}
\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{J}\sum_{j=1}^{J} \|\mathbf{e}^j(t)\|^2 &= \frac{1}{J}\sum_{j=1}^{J} \left\langle \mathbf{e}^j(t), \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{e}^j(t) \right\rangle \\
&= -\frac{1}{J}\sum_{j=1}^{J} \left\langle \mathbf{e}^j(t), \mathbf{C}_{\mathbf{uu}}\mathbf{G}^\intercal\boldsymbol{\Gamma}\mathbf{G}\mathbf{e}^j(t) \right\rangle \\
&\quad - \frac{1}{J}\sum_{j=1}^{J} \left\langle \mathbf{e}^j(t), \mathbf{C}_{\mathbf{uu}} \left(\lambda^j \nabla_{\mathbf{u}}\mathcal{A}(\mathbf{u}^j) - \frac{1}{J}\sum_{\ell=1}^{J} \lambda^\ell \nabla_{\mathbf{u}}\mathcal{A}(\mathbf{u}^\ell) \right) \right\rangle .
\end{aligned}$$

Let us consider the first term in the right-hand side. Using the structure of $\mathbf{C}_{\mathbf{uu}}$ we have

$$-\frac{1}{J}\sum_{j=1}^{J} \left\langle \mathbf{e}^j(t), \mathbf{C}_{\mathbf{uu}}\mathbf{G}^\intercal\boldsymbol{\Gamma}\mathbf{G}\mathbf{e}^j(t) \right\rangle = -\frac{1}{J^2}\sum_{k,j=1}^{J} \left\langle \mathbf{e}^j(t), \mathbf{e}^k(t) \right\rangle \left\langle \mathbf{e}^k(t), \mathbf{G}^\intercal\boldsymbol{\Gamma}\mathbf{G}\mathbf{e}^j(t) \right\rangle .$$

Since $\mathbf{G}^\intercal\boldsymbol{\Gamma}\mathbf{G}$ is symmetric and positive semidefinite, it is possible to find a set of eigenpairs $(\mu^i, \mathbf{v}^i)_{i=1}^d$ with $\mu^i \geqslant 0$ and $\mathbf{v}^i$ orthonormal basis such that $\mathbf{e}^j = \sum_{i=1}^{d} \alpha^{ij}\mathbf{v}^i$, $\forall j = 1, \ldots, J$, and $\mathbf{G}^\intercal\boldsymbol{\Gamma}\mathbf{G}\mathbf{v}^i = \mu^i\mathbf{v}^i$, for all $i = 1, \ldots, d$. Then

$$-\frac{1}{J^2}\sum_{k,j=1}^{J} \left\langle \mathbf{e}^j(t), \mathbf{e}^k(t) \right\rangle \left\langle \mathbf{e}^k(t), \mathbf{G}^\intercal\boldsymbol{\Gamma}\mathbf{G}\mathbf{e}^j(t) \right\rangle = -\frac{1}{J^2}\sum_{k,j=1}^{J} \left(\sum_{i=1}^{d} \sqrt{\mu^i}\alpha^{ik}\alpha^{ij} \right)^2 \leqslant 0.$$

For the second term in the right-hand side, using the assumption $\lambda^j(t) = \Lambda(t) \geqslant 0$ and the convexity of the constraint, we obtain

$$-\frac{1}{J}\sum_{j=1}^{J} \left\langle \mathbf{e}^j(t), \mathbf{C}_{\mathbf{uu}} \left(\lambda^j \nabla_{\mathbf{u}}\mathcal{A}(\mathbf{u}^j) - \frac{1}{J}\sum_{\ell=1}^{J} \lambda^\ell \nabla_{\mathbf{u}}\mathcal{A}(\mathbf{u}^\ell) \right) \right\rangle = -\frac{\Lambda}{J}\sum_{j=1}^{J} \left\langle \mathbf{e}^j(t), \mathbf{C}_{\mathbf{uu}}\nabla_{\mathbf{u}}\mathcal{A}(\mathbf{e}^j) \right\rangle \leqslant 0.$$

□

Under stronger assumptions, it is possible to prove convergence of the residual in the control space.

Proposition 3.10. *Let $\mathbf{u}^{j,0} \in \mathbb{R}^d$, $j = 1, \ldots, J$, be an initial condition of the DAE system (18) such that $\mathcal{A}(\mathbf{u}^{j,0}) = \mathbf{0}_{\mathbb{R}^m}$. Assume that $\mathcal{A}(\mathbf{u}) = \mathbf{0}_{\mathbb{R}^m}$ is a convex feasible domain and $\mathcal{A} : \mathbb{R}^d \to \mathbb{R}^m$ is convex and quadratic. Assume that $\mathbf{G}^\intercal\boldsymbol{\Gamma}\mathbf{G}$ is positive definite. Then $\lim_{t\to\infty}\frac{1}{J}\sum_{j=1}^{J} \|\mathbf{r}^j(t)\|^2 = 0$, provided that $\boldsymbol{\lambda}^j(t) = \boldsymbol{\Lambda}(t) \geqslant 0$, $\forall j = 1, \ldots, J$ and $t > 0$.*

Proof. By assumption $\mathbf{G}^\intercal\boldsymbol{\Gamma}\mathbf{G}$ is positive definite and thus we have a unique global minimizer $\mathbf{u}^*$ of the constrained minimization problem

$$\min_{\mathbf{u}\in\mathbb{R}^d}\Phi(\mathbf{u},\mathbf{y}) \quad \text{subject to } \mathcal{A}(\mathbf{u}) = \mathbf{0}_{\mathbb{R}^m}$$

for a given $\mathbf{y}\in\mathbb{R}^K$. For simplicity and without loss of generality we consider the case of a scalar constraint, i.e. $m = 1$. We study

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2J}\sum_{j=1}^{J}\|\mathbf{r}^j(t)\|^2 &= \frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2J}\sum_{j=1}^{J}\|\mathbf{u}^j(t)-\mathbf{u}^*\|^2 = \frac{1}{J}\sum_{j=1}^{J}\left\langle \mathbf{u}^j(t)-\mathbf{u}^*, \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{u}^j(t)\right\rangle\\
&= -\frac{1}{J}\sum_{j=1}^{J}\left\langle \mathbf{u}^j(t)-\mathbf{u}^*, \mathbf{C}_{\mathbf{uu}}\left(\nabla_{\mathbf{u}}(\mathbf{u}^j,\mathbf{y}) + \lambda^j\nabla_{\mathbf{u}}\mathcal{A}(\mathbf{u})|_{\mathbf{u}=\mathbf{u}^j(t)}\right)\right\rangle\\
&= -\frac{1}{J}\sum_{j=1}^{J}\left\langle \mathbf{u}^j-\mathbf{u}^*, \mathbf{C}_{\mathbf{uu}}\mathbf{G}^\intercal\boldsymbol{\Gamma}\mathbf{G}(\mathbf{u}^j-\mathbf{u}^*)\right\rangle\\
&\quad -\frac{\Lambda}{J}\sum_{j=1}^{J}\left\langle \mathbf{u}^j-\mathbf{u}^*, \lambda^j\mathbf{C}_{\mathbf{uu}}\nabla_{\mathbf{u}}\left(\mathcal{A}(\mathbf{u})|_{\mathbf{u}=\mathbf{u}^j} - \mathcal{A}(\mathbf{u}^*)\right)\right\rangle < 0.\end{aligned}$$

□

The assumptions that $\mathbf{G}^\intercal\boldsymbol{\Gamma}\mathbf{G}$ and the empirical covariance $\mathbf{C}_{\mathbf{uu}}$ are positive definite are strong and in general not satisfied. To deal with the positive definiteness of $\mathbf{C}_{\mathbf{uu}}$ usually a constant or time-dependent inflation of the covariance is considered [4, 17, 27].

3.2. Continuous limit in the number of ensembles

Typically, the EnKF method is applied for a fixed and finite ensemble size. It is clear that the computational and memory cost of the method increases with the number of the ensembles, but there is a substantial gain in accuracy. The analysis of the method was also studied in the large ensemble limit, see e.g. [8–11, 28–31]. In this limit a slow down of the ensemble spread has been observed. In this section, we derive the corresponding mean-field limit of the continuous time equation (18) and provide an analysis of the resulting PDE equation.

We follow the classical formal derivation to formulate a mean-field equation of a particle system, see [32–35]. Let us denote by

$$f = f(t,\mathbf{u},\boldsymbol{\lambda}) : \mathbb{R}^+\times\mathbb{R}^d\times\mathbb{R}^m \to \mathbb{R}^+ \tag{27}$$

the compactly supported on $\mathbb{R}^d\times\mathbb{R}^m$ probability density of the pair $(\mathbf{u},\boldsymbol{\lambda})$ at time t and introduce the following moments of f at time t with respect to $\mathbf{u}$ and $\boldsymbol{\lambda}$, respectively, as

$$\begin{bmatrix}\mathbf{m}_1(t)\\ \boldsymbol{\Lambda}_1(t)\end{bmatrix} = \iint_{\mathbb{R}^d\times\mathbb{R}^m}\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\lambda}\end{bmatrix} f(t,\mathbf{u},\boldsymbol{\lambda})\,\mathrm{d}\mathbf{u}\,\mathrm{d}\boldsymbol{\lambda}, \quad \begin{bmatrix}\mathbf{m}_2(t)\\ \boldsymbol{\Lambda}_2(t)\end{bmatrix} = \iint_{\mathbb{R}^d\times\mathbb{R}^m}\begin{bmatrix}\mathbf{u}\otimes\mathbf{u}\\ \boldsymbol{\lambda}\otimes\boldsymbol{\lambda}\end{bmatrix} f(t,\mathbf{u},\boldsymbol{\lambda})\,\mathrm{d}\mathbf{u}\,\mathrm{d}\boldsymbol{\lambda} \tag{28}$$

Since $(\mathbf{u},\boldsymbol{\lambda})\in\mathbb{R}^d\times\mathbb{R}^m$, the empirical measure is given by

$$f(t,\mathbf{u},\boldsymbol{\lambda}) = \frac{1}{J}\sum_{j=1}^{J}\delta(\mathbf{u}^j-\mathbf{u})\delta(\boldsymbol{\lambda}^j-\boldsymbol{\lambda}). \tag{29}$$

This formulation allows for a representation of the covariance operator (24) as

$$\begin{aligned}\mathcal{C}(t) &= \iint_{\mathbb{R}^d\times\mathbb{R}^m} \mathbf{u}\otimes\mathbf{u} f(t,\mathbf{u},\boldsymbol{\lambda})\,\mathrm{d}\mathbf{u}\,\mathrm{d}\boldsymbol{\lambda} - \iint_{\mathbb{R}^d\times\mathbb{R}^m} \mathbf{u} f(t,\mathbf{u},\boldsymbol{\lambda})\,\mathrm{d}\mathbf{u}\,\mathrm{d}\boldsymbol{\lambda} \iint_{\mathbb{R}^d\times\mathbb{R}^m} \mathbf{u} f(t,\mathbf{u},\boldsymbol{\lambda})\,\mathrm{d}\mathbf{u}\,\mathrm{d}\boldsymbol{\lambda}\\ &= \mathbf{m}_2(t) - \mathbf{m}_1(t)\otimes\mathbf{m}_1(t) \geqslant 0.\end{aligned}$$

Let us denote $\varphi(\mathbf{u},\boldsymbol{\lambda}) \in C_0^\infty(\mathbb{R}^d\times\mathbb{R}^m)$ a test function. We compute

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\langle f,\varphi\rangle &= \frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d}\frac{1}{J}\sum_{j=1}^{J}\delta(\mathbf{u}-\mathbf{u}^j)\delta(\boldsymbol{\lambda}^j-\boldsymbol{\lambda})\varphi(\mathbf{u},\boldsymbol{\lambda})\,\mathrm{d}\mathbf{u}\,\mathrm{d}\boldsymbol{\lambda}\\ &= -\frac{1}{J}\sum_{j=1}^{J}\nabla_{\mathbf{u}}\varphi(\mathbf{u}^j,\boldsymbol{\lambda}^j)\cdot\mathcal{C}\left(\nabla_{\mathbf{u}}\Phi(\mathbf{u}^j,\mathbf{y}) + \mathbf{J}_{\mathcal{A}}^{\intercal}(\mathbf{u}^j)\boldsymbol{\lambda}^j\right)\\ &= -\iint_{\mathbb{R}^d\times\mathbb{R}^m}\nabla_{\mathbf{u}}\varphi(\mathbf{u},\boldsymbol{\lambda})\cdot\mathcal{C}\left(\nabla_{\mathbf{u}}\Phi(\mathbf{u},\mathbf{y}) + \mathbf{J}_{\mathcal{A}}^{\intercal}(\mathbf{u})\boldsymbol{\lambda}\right) f(t,\mathbf{u})\,\mathrm{d}\mathbf{u}\,\mathrm{d}\boldsymbol{\lambda}\end{aligned}$$

coupled to the mean-field of the algebraic constraint. Finally the mean-field kinetic equation corresponding to the DAE system (18) reads:

$$\begin{aligned}\partial_t f(t,\mathbf{u},\boldsymbol{\lambda}) - \nabla_{\mathbf{u}}\cdot\left[\mathcal{C}(t)(\nabla_{\mathbf{u}}\Phi(\mathbf{u},\mathbf{y}) + \mathbf{J}_{\mathcal{A}}^{\intercal}(\mathbf{u})\boldsymbol{\lambda})f(t,\mathbf{u},\boldsymbol{\lambda})\right] &= 0\\ \iint_{\mathbb{R}^d\times\mathbb{R}^m}\mathcal{A}(\mathbf{u})f(t,\mathbf{u},\boldsymbol{\lambda})\,\mathrm{d}\mathbf{u}\,\mathrm{d}\boldsymbol{\lambda} &= \mathbf{0}_{\mathbb{R}^m}.\end{aligned} \tag{30}$$

Proposition 3.11. *Let $f(t,\mathbf{u},\boldsymbol{\lambda})$ be a solution in distributional sense of the mean-field equation (30) at $t>0$ for compactly supported initial probability distribution $f(t=0,\mathbf{u},\boldsymbol{\lambda})$. Then, $f(\mathbf{u},\boldsymbol{\lambda}) = \delta(\mathbf{u}-\mathbf{v})\delta(\boldsymbol{\lambda}-\boldsymbol{\mu})$ is a steady state solution in distributional sense of (30) provided that either $\mathcal{C} = \mathbf{0}_{\mathbb{R}^{d\times d}}$ or $(\mathbf{v},\boldsymbol{\mu})$ is a KKT point of the minimization problem*

$$\min_{\mathbf{u}\in\mathbb{R}^d}\Phi(\mathbf{u},\mathbf{y}) \quad \textit{subject to } \mathcal{A}(\mathbf{u}) = \mathbf{0}_{\mathbb{R}^m}.$$

Proof. Let $\varphi(\mathbf{u},\boldsymbol{\lambda}) \in C_0^\infty(\mathbb{R}^d\times\mathbb{R}^m)$ be a test function. Then, weak steady state solutions, say $f^\infty(\mathbf{u},\boldsymbol{\lambda})$, to (30) satisfy formally

$$\begin{aligned}\iint_{\mathbb{R}^d\times\mathbb{R}^m}\nabla_{\mathbf{u}}\varphi(\mathbf{u},\boldsymbol{\lambda})\cdot\mathcal{C}\left(\nabla_{\mathbf{u}}\Phi(\mathbf{u},\mathbf{y}) + \mathbf{J}_{\mathcal{A}}^{\intercal}(\mathbf{u})\boldsymbol{\lambda})\right)\,\mathrm{d}f^\infty(\mathbf{u},\boldsymbol{\lambda}) &= 0\\ \iint_{\mathbb{R}^d\times\mathbb{R}^m}\varphi(\mathbf{u},\boldsymbol{\lambda})\mathcal{A}(\mathbf{u})\,\mathrm{d}f^\infty(\mathbf{u},\boldsymbol{\lambda}) &= \mathbf{0}_{\mathbb{R}^m}\end{aligned}$$

Substituting f^∞ with $f(\mathbf{u},\boldsymbol{\lambda}) = \delta(\mathbf{u}-\mathbf{v})\delta(\boldsymbol{\lambda}-\boldsymbol{\mu})$, we easily obtain that the above conditions are satisfied when either $\mathcal{C} = \mathbf{0}_{\mathbb{R}^{d\times d}}$ or $\nabla_{\mathbf{u}}\Phi(\mathbf{u},\mathbf{y})|_{\mathbf{u}=\mathbf{v}} = -\left(\mathbf{J}_{\mathcal{A}}^{\intercal}(\mathbf{u})\boldsymbol{\lambda}\right)|_{\mathbf{u}=\mathbf{v},\boldsymbol{\lambda}=\boldsymbol{\mu}}$ and simultaneously $\mathcal{A}(\mathbf{u})|_{\mathbf{u}=\mathbf{v}} = \mathbf{0}_{\mathbb{R}^m}$. □

The previous proposition states that, as in the discrete case, not all the steady states are in the kernel of $\mathcal{C}$ due to the presence of the multipliers.

It is clear that the mean-field equation (30) does not provide a closed differential system of the moments. In order to show an energy decay estimate, we perform a preliminary moment analysis in the simple setting of a one-dimensional control and scalar constraint. Thus, similarly

to the previous section, we assume that $(u, \lambda) \in \mathbb{R} \times \mathbb{R}$ and consider a strictly convex quadratic constraint $\mathcal{A}(u) \in \mathbb{R}$. We write

$$\nabla_u \Phi(u, y) = b_1 u + b_2, \quad \nabla_u \mathcal{A}(u) = h_1 u + h_2,\ h_1 > 0.$$

Then by the second equation in (30) we have

$$0 = \iint_{\mathbb{R}^2} \left(\frac{1}{2} h_1 u^2 f(t, u, \lambda) + h_2 u f(t, u, \lambda) \right) \mathrm{d}u\, \mathrm{d}\lambda = \frac{1}{2} h_1 m_2(t) + h_2 m_1(t)$$

which implies the following link between the first and the second moment due to the constraint:

$$m_2(t) = -2 \frac{h_2}{h_1} m_1(t). \tag{31}$$

Then, it suffices to study the evolution of the first moment in order to obtain information on the evolution of the second moment. Using the first equation in (30) we derive the following evolution equation for the first moment:

$$\frac{\mathrm{d}}{\mathrm{d}t} m_1(t) = -C \iint_{\mathbb{R}^2} \nabla_u \Phi(u, y) f(t, u, \lambda)\, \mathrm{d}u\, \mathrm{d}\lambda - C \iint_{\mathbb{R}^2} (h_1 u \lambda f(t, u, \lambda) + h_2 \lambda f(t, u, \lambda))\, \mathrm{d}u\, \mathrm{d}\lambda. \tag{32}$$

The evolution of the multiplier is computed from (30) obtaining for each $k \geqslant 0$

$$\frac{\mathrm{d}}{\mathrm{d}t} \iint_{\mathbb{R}^2} \lambda^k f(t, u, \lambda)\, \mathrm{d}u\, \mathrm{d}\lambda = 0 \Rightarrow \iint_{\mathbb{R}^2} \lambda^k f(t, u, \lambda)\, \mathrm{d}u\, \mathrm{d}\lambda = \iint_{\mathbb{R}^2} \lambda^k f_0(u, \lambda)\, \mathrm{d}u\, \mathrm{d}\lambda =: \Lambda_k.$$

For an arbitrary small positive quantity ϵ we have

$$\iint_{\mathbb{R}^2} (\sqrt{\epsilon} u) \frac{\lambda}{\sqrt{\epsilon}} f(t, u, \lambda)\, \mathrm{d}u\, \mathrm{d}\lambda \leqslant \frac{\epsilon}{2} m_2(t) + \frac{1}{2\epsilon} \Lambda_2.$$

Using the above relation and (31), from (32) we obtain the following bound for the evolution equation of the first moment:

$$\frac{\mathrm{d}}{\mathrm{d}t} m_1(t) \leqslant L_\epsilon \mathcal{C}(t) \left(m_1(t) + \frac{K_\epsilon}{L_\epsilon} \right), \quad L_\epsilon = \epsilon h_2 - b_1,\ K_\epsilon = -b_2 - \frac{h_1}{2\epsilon} \Lambda_2 - h_2 \Lambda_1$$

with $L_\epsilon < 0$ for ϵ sufficiently small. Defining $\tilde{m}_1(t) = m_1(t) + \frac{K_\epsilon}{L_\epsilon}$, by application of the Gronwall lemma we have

$$\frac{\mathrm{d}}{\mathrm{d}t} \tilde{m}_1(t) \leqslant L_\epsilon \mathcal{C}(t) \tilde{m}_1(t) \Rightarrow \tilde{m}_1(t) \leqslant \tilde{m}_1(0) \exp\left(L_\epsilon \int_0^t \mathcal{C}(\tau)\, \mathrm{d}\tau \right) \to 0.$$

Finally,

$$m_1(t) \leqslant -\frac{K_\epsilon}{L_\epsilon} = \left| \frac{K_\epsilon}{L_\epsilon} \right|$$

and the mean is hence bounded. Consequently, also the second moment m_2 is bounded by the relation (31). This shows that, according to proposition 3.11, not necessarily all the steady states are Dirac distributions on $\mathbb{R}^d \times \mathbb{R}^m$.

4. Numerical experiments

The simulations are performed for the case of linear models $\mathbf{G}$ by solving the system of DAEs (18). The details on the scheme and on the setting of the experiments are as follows. Many numerical methods for DAEs are known and based on Runge–Kutta and BDF methods, see e.g. the monographs [36, 37]. Here, for the sake of simplicity, we employ straightforwardly first order implicit–explicit (IMEX) time integration coupled to Newton's method. The IMEX discretization of (18) with time step Δt then reads as

$$\begin{aligned} \mathbf{u}^{j,n+1} &= \mathbf{u}^{j,n} - \Delta t \mathbf{C}^n_{\mathbf{uu}} \nabla_{\mathbf{u}} \Phi(\mathbf{u}^{j,n}, \mathbf{y}) - \Delta t \mathbf{C}^n_{\mathbf{uu}} \mathbf{J}_{\mathcal{A}}(\mathbf{u}^{j,n+1}) \boldsymbol{\lambda}^{j,n+1} \\ \mathbf{0}_{\mathbb{R}^m} &= \mathcal{A}(\mathbf{u}^{j,n+1}). \end{aligned} \tag{33}$$

This represents a possible high dimensional system of nonlinear equations for $\mathbf{u}^{j,n+1} \in \mathbb{R}^d$ and $\boldsymbol{\lambda}^{j,n+1} \in \mathbb{R}^m$, for $j = 1, \ldots, J$, to be solved at each time integration step n. Inspired by the derivation of (18) we consider the covariance matrix $\mathbf{C}_{\mathbf{uu}}$ computed explicitly at time level n. This is motivated by the fact that $\mathbf{C}_{\mathbf{uu}}$ can be written in terms of moments. Observe that then the IMEX discretization is (18) in the limit $\Delta t \to 0^+$.

The differential point of view allows us to employ an adaptive time step to avoid stability issues. In the numerical experiments we use

$$\Delta t \leqslant \frac{1}{\max_i \left(| (\Re(\mu_i) | \right)}$$

where the μ_i's are the eigenvalues of $\mathbf{C}^n_{\mathbf{uu}} \mathbf{G}^\intercal \boldsymbol{\Gamma} \mathbf{G}$. As we observe that $\mathbf{C}^n_{\mathbf{uu}} \mathbf{G}^\intercal \boldsymbol{\Gamma} \mathbf{G}$ is characterized by large spectral radius at initial time that reduces over time, the adaptive computation of Δt allows also to reach equilibrium in less time steps compared to the choice of a fixed Δt.

We are interested to solve the inverse problem described in section 2.2 aimed to find the force function of the linear elliptic equation in (14). This is a typical example used in the mathematical literature to test the property of the ensemble Kalman filter [6, 11, 16, 19]. We consider the same setup as in section 2.2, but in this case we take into account a nonlinear constraint. In particular, we focus on a scalar ($m = 1$) quadratic constraint $\mathcal{A} : \mathbb{R}^d \to \mathbb{R}$ of the form $\mathcal{A}(\mathbf{u}) = \frac{1}{2}\mathbf{u}^\intercal \mathbf{A} \mathbf{u} + \mathbf{u}^\intercal \mathbf{b}$, with $\mathbf{A} \in \mathbb{R}^{d \times d}$ and $\mathbf{b} \in \mathbb{R}^d$ given. In this setting, the gradient of $\mathcal{A}$ is linear, $\nabla_{\mathbf{u}} \mathcal{A} = \mathbf{A}\mathbf{u} + \mathbf{b}$. Therefore, (33) can be explicitly solved by $\mathbf{u}^{j,\,n+1}$ obtaining

$$\mathbf{u}^{j,n+1} = \left(\mathbf{I} + \Delta t \lambda^{j,n+1} \mathbf{C}^n_{\mathbf{uu}} \right)^{-1} \left(\mathbf{u}^{j,n} - \Delta t \mathbf{C}_{\mathbf{uu}} \nabla_{\mathbf{u}} \Phi(\mathbf{u}^{j,n}, \mathbf{y}) - \Delta t \lambda^{j,n+1} \mathbf{C}_{\mathbf{uu}} \mathbf{b} \right), \tag{34}$$

where $\mathbf{I} \in \mathbb{R}^{d \times d}$ is the identity matrix. Then, the scalar multiplier $\lambda^{j,n+1}$ is computed by solving the nonlinear equation $\mathcal{A}(\mathbf{u}^{j,n+1}) = 0$ for each $j = 1, \ldots, J$ at each integration step. The multipliers are finally inserted into (34) to determine the final update of the feasible ensembles at time level $n + 1$.

In each example the exact force function is chosen as $u(x) = \sin(\pi x)$ and then suitable modified in order to satisfy the constraint. Also, the initial condition $\mathbf{u}^j(0) = \mathbf{u}^{j,0}$ on the ensembles satisfies the constraints. The vector $\mathbf{b}$, characterizing the linear term in the constraint, is always randomly sampled from a Gaussian distribution with zero mean and standard deviation $\frac{1}{2}$.

Information on the simulation results is presented in the following norms

$$E(t) = \frac{1}{J} \sum_{j=1}^{J} \|\mathbf{e}^j(t)\|^2, \quad R(t) = \frac{1}{J} \sum_{j=1}^{J} \|\mathbf{r}^j(t)\|^2 \tag{35}$$

at each iteration, where the quantities $\mathbf{e}^j(t)$ and $\mathbf{r}^j(t)$ are the spread and the residual of the ensembles, respectively, as defined by (21). The residual is measured by taking $\mathbf{u}^*$ as the truth solution $\mathbf{u}^\dagger$.

Another additional important quantities is given by the misfit which allows to measure the quality of the solution at each iteration. The misfit for the jth sample is defined as

$$\boldsymbol{\vartheta}^j(t) = \mathbf{G}\mathbf{r}^j(t) - \boldsymbol{\eta}, \tag{36}$$

where $\mathbf{G}$ is again the finite difference discretization of the continuous linear operator defining the elliptic PDE (14). By using (36) we finally look at

$$\vartheta(t) = \frac{1}{J}\sum_{j=1}^{J} \|\boldsymbol{\vartheta}^j(t)\|^2. \tag{37}$$

Driving this quantity to zero leads to over-fitting of the solution. For this reason, usually it is suitable introducing a stopping criterion which avoids this effect. In the following we will consider the discrepancy principle which check and stop the simulation if the condition $\vartheta \leqslant \|\boldsymbol{\eta}\|^2$ is satisfied, with $\boldsymbol{\eta}$ measurement noise as described in the problem setup in section 2.2.

4.1. Quadratic convex constraint

The first situation we consider is the case of a strictly convex constraint by taking $\mathbf{A} = \mathbf{I} \in \mathbb{R}^{d\times d}$, i.e.,

$$\mathcal{A}(\mathbf{u}) = \sum_{k=1}^{d} \mathbf{u}_{(k)}^2 + \mathbf{b}_{(k)}\mathbf{u}_{(k)}.$$

In order to allow the force function $u(x)$ to satisfy the constraint, we artificially modify it on the grid introduced on the domain $[0, \pi]$. In particular, we determine $\{\mathbf{u}_{(k)}\}_{k=1}^{\frac{d}{2}}$ in such a way $\mathbf{u}_{(k)}^2 + \mathbf{b}_{(k)}\mathbf{u}_{(k)} = -\mathbf{u}_{(d-k+1)}^2 - \mathbf{b}_{(d-k+1)}\mathbf{u}_{(d-k+1)}$, for $k = 1, \ldots, \frac{d}{2}$. The initial ensembles are sampled from a multivariate normal distribution and then also artificially modified to satisfy the constraint. We consider the same setup as in section 2.2, thus a noise level $\gamma = 0.01$, $d = K = 2^8$.

In figure 2 we compute the solution obtained with $J = 160$ ensembles. The left plot shows the solution $p(x)$ of the PDE (14) and noisy observations. The reconstruction of $p(x)$ is provided in the center panel and it is obtained by using the mean of the ensembles represented in the right panel. The gray areas gives information on the spread due to the ensembles. The constrained ensemble Kalman filter accurately reconstructs both the control and its projection through the PDE model.

The analysis of the solution is performed with three values of the ensemble size $J \in \{40, 80, 160\}$, see figure 3. In particular, we consider the behavior in time of the residual, the misfit, the ensemble collapse and the value of the constraint computed on the ensemble mean. We observe that increasing the value of the ensembles allows a more accurate reconstruction. In fact, the solution is able to meet the discrepancy principle, i.e. the misfit reaches the noise level, and the residual values decreases. However, the collapse to the mean slows down causing the increase of the ensemble spread. This in turn does not allow the mean to satisfy exactly the constraint when the simulation stops due to the discrepancy principle.

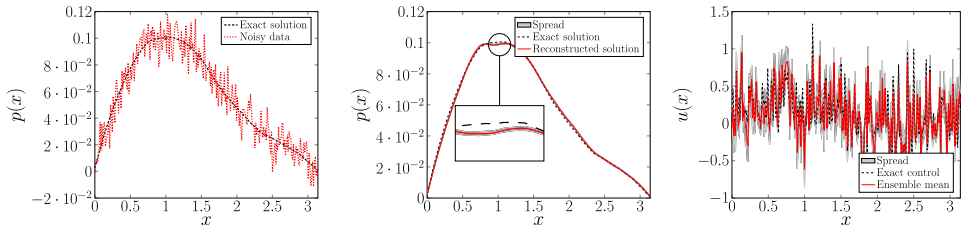

Figure 2. Solution provided by the constrained ensemble Kalman filter using $J = 160$ ensembles on the strictly convex nonlinear constraint.

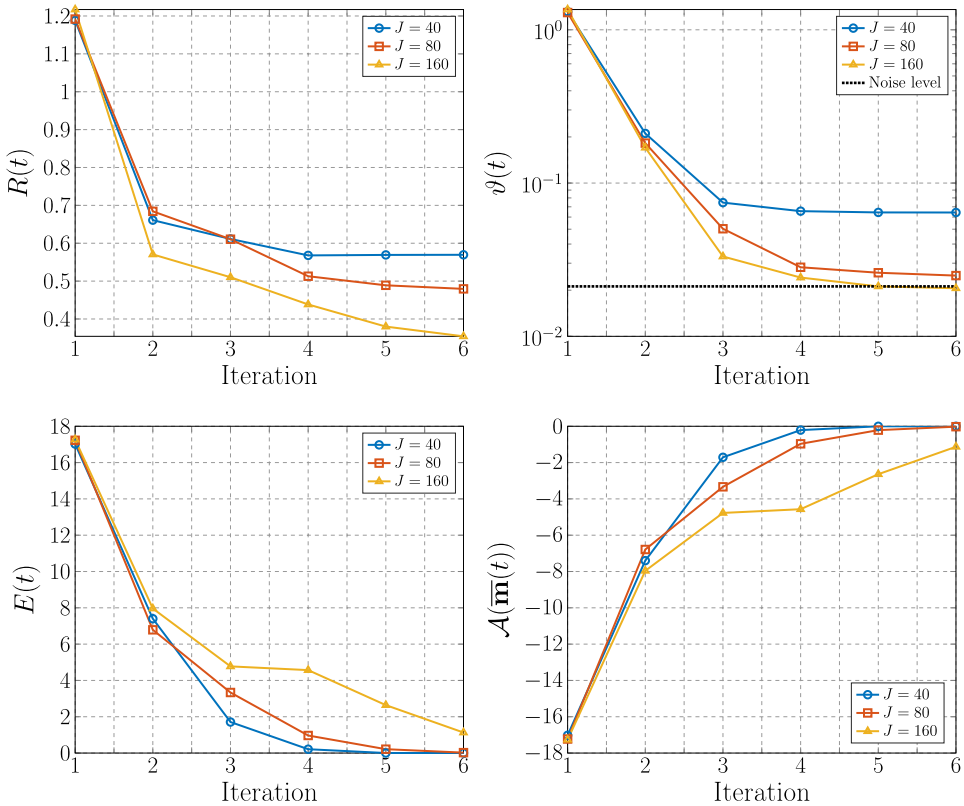

Figure 3. Time evolution of the residual (top left), the misfit (top right), the ensemble collapse (bottom left) and the value of the constraint computed on the ensemble mean (bottom right) for the strictly convex nonlinear constraint.

In figure 4 we show the evolution of the minimum value of the multipliers, noticing that it remains always positive. Moreover, the spread of the multipliers to their mean decays to zero. This two results justify the ensemble spread observed in figure 3 and reflect the analysis provided in section 3.1.1, cf the discussion after proposition 3.7 and the statement of proposition 3.9

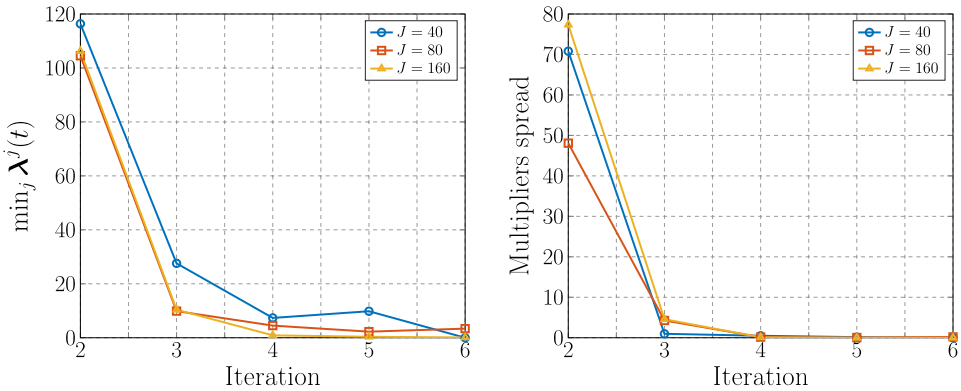

Figure 4. Time evolution of the minimum value of the multipliers (left) and of the multipliers spread around their mean (right) for the strictly convex nonlinear constraint.

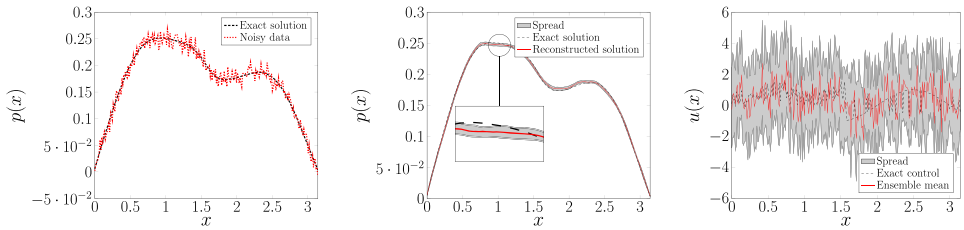

Figure 5. Solution provided by the constrained ensemble Kalman filter using $J = 160$ ensembles on the non-convex nonlinear constraint.

4.2. Quadratic non-convex constraint

The second situation we consider is the case of a non-convex constraint by taking $\mathbf{A} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I} \end{bmatrix} \in \mathbb{R}^{d\times d}$. Then the constraint reads as

$$\mathcal{A}(\mathbf{u}) = \sum_{k=1}^{\frac{d}{2}} \mathbf{u}_{(k)}^2 + \mathbf{b}_{(k)}\mathbf{u}_{(k)} + \sum_{\ell=\frac{d}{2}+1}^{d} \mathbf{b}_{(k)}\mathbf{u}_{(k)} - \mathbf{u}_{(\ell)}^2.$$

We use the same setup as in section 4.1. Now the force function $u(x)$ is artificially modified to satisfy the constraint by determining $\{\mathbf{u}_{(k)}\}_{k=1}^{\frac{d}{2}}$ in such a way $\mathbf{u}_{(k)}^2 + \mathbf{b}_{(k)}\mathbf{u}_{(k)} = \mathbf{u}_{(d-k+1)}^2 - \mathbf{b}_{(d-k+1)}\mathbf{u}_{(d-k+1)}$, for $k = 1, \ldots, \frac{d}{2}$. We proceed similarly for the initial ensembles.

In figure 5 we compute the solution obtained with $J = 160$ ensembles. The left plot shows the solution $p(x)$ of the PDE (14) and the noisy observations. The reconstruction of $p(x)$ is provided in the center panel and it is obtained by using the mean of the ensembles represented in the right panel. The gray areas gives information on the spread due to the ensembles. Compared to the previous example, here we notice that the ensembles are widely spread around the exact control function, cf the right plot in figure 5. The method does not reproduce a very accurate control function but the application of the model still allows to obtain a good reconstruction of the solution.

We perform the analysis of the method with five values of the ensemble size $J \in \{40, 80, 160, 320, 640\}$, see figure 6. Again we consider the behavior in time of the residual, the

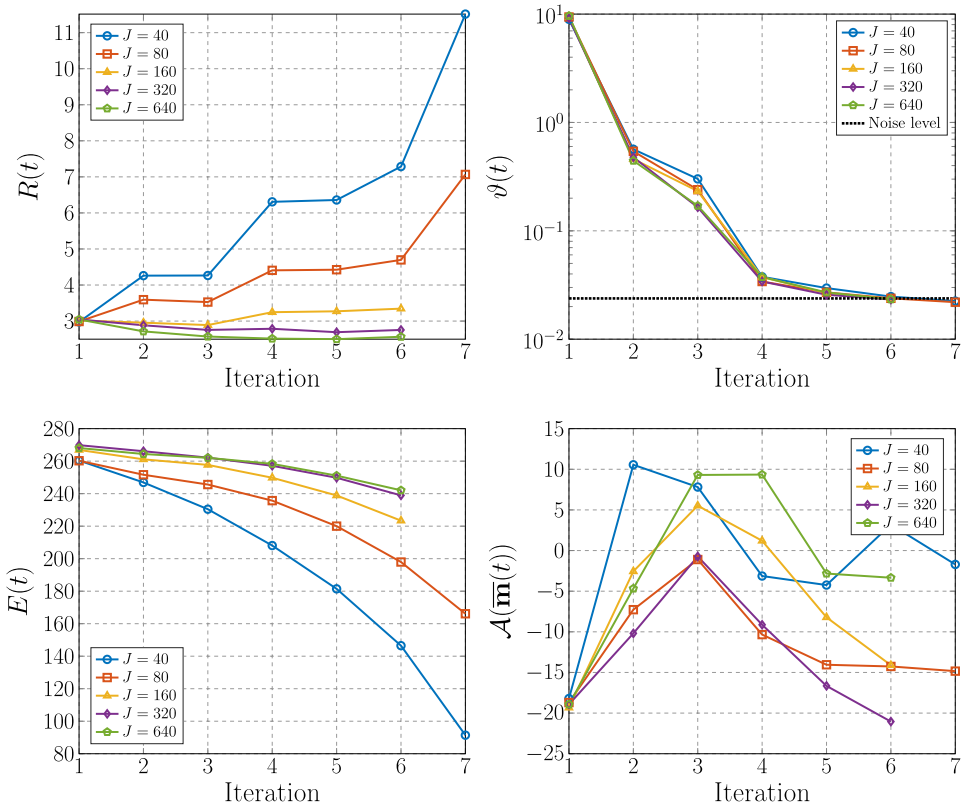

Figure 6. Time evolution of the residual (top left), the misfit (top right), the ensemble collapse (bottom left) and the value of the constraint computed on the ensemble mean (bottom right) for the non-convex nonlinear constraint.

misfit, the ensemble collapse and the value of the constraint computed on the ensemble mean. We observe that the ensemble spread is decreasing in time very slowly, even with a small number of ensembles. This result also affects the value of the constraint computed on the mean of the ensembles, which is far from zero. The residual shows a decreasing behavior in time only with very large number of ensembles. However, in all cases the misfit is able to meet the noise level and therefore the discrepancy principle holds. These results can be motivated by the use of a non-convex constraint, and the method is providing a solution which minimizes the least square functional but the mean does not exactly satisfy the constraint.

5. Summary and perspectives

In this paper, inspired by [16], we have focused on the formulation of the ensemble Kalman filter to solve constrained inverse problems. We have worked in the setting of optimization theory deriving first order necessary optimality conditions for the case of equality constraints in the space of the controls. We have observed that the method relaxes to the unconstrained one when linear equality constraints are considered. Therefore, we have mainly focused on nonlinear constraints and analyzed the method by computing continuous limits, in time and in the regime of infinitely many ensembles. The numerical results have shown that the method is able to provide solution to constrained inverse problems with quadratic convex and non-convex equality constraints.

We recall that in order to set up the mathematical formulation of the problem studied in this work, we have considered finite dimensional Hilbert spaces and equality constraints. As future perspectives it would be of interest to extend this study to arbitrary and possibly infinite dimensional Hilbert spaces, and to consider also inequality constraints.

Acknowledgments

This research is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC-2023 Internet of Production—390621612 and supported also by DFG HE5386/15.

Appendix A. Detailed computations in the proof of proposition 2.5

- Woodbury matrix identity:

$$(\mathbf{M}+\mathbf{U}\mathbf{N}\mathbf{V})^{-1}=\mathbf{M}^{-1}-\mathbf{M}^{-1}\mathbf{U}(\mathbf{N}^{-1}+\mathbf{V}\mathbf{M}^{-1}\mathbf{U})^{-1}\mathbf{V}\mathbf{M}^{-1},$$

  for all matrices $\mathbf{M}$, $\mathbf{U}$, $\mathbf{N}$, $\mathbf{V}$.
- Application of the Woodbury matrix identity to $\left(\mathbf{H}^{\intercal}\boldsymbol{\Gamma}\mathbf{H}+\mathbf{C}_{\varepsilon}^{n+1^{-1}}\right)^{-1}$:

$$\left(\mathbf{H}^{\intercal}\boldsymbol{\Gamma}\mathbf{H}+\mathbf{C}_{\varepsilon}^{n+1^{-1}}\right)^{-1}=\mathbf{C}_{\varepsilon}^{n+1}-\mathbf{C}_{\varepsilon}^{n+1}\mathbf{H}^{\intercal}\left(\boldsymbol{\Gamma}^{-1}+\mathbf{H}\mathbf{C}_{\varepsilon}^{n+1}\mathbf{H}^{\intercal}\right)^{-1}\mathbf{H}\mathbf{C}_{\varepsilon}^{n+1}$$

- From equation (12) to equation (13):

$$\begin{aligned}
\mathbf{v}^{j,n+1}=&\left(\mathbf{H}^{\intercal}\boldsymbol{\Gamma}\mathbf{H}+\mathbf{C}_{\varepsilon}^{n+1^{-1}}\right)^{-1}\left(\mathbf{H}^{\intercal}\boldsymbol{\Gamma}\mathbf{y}^{n+1}+\mathbf{C}_{\varepsilon}^{n+1^{-1}}\hat{\mathbf{v}}^{j,n+1}\right)\\
&-\left(\mathbf{H}^{\intercal}\boldsymbol{\Gamma}\mathbf{H}+\mathbf{C}_{\varepsilon}^{n+1^{-1}}\right)^{-1}\mathbf{J}_{\tilde{\mathcal{A}}}^{\intercal}(\mathbf{v}^{j,n+1})\boldsymbol{\lambda}^{j,n+1}\\
=&\left(\mathbf{H}^{\intercal}\boldsymbol{\Gamma}\mathbf{H}+\mathbf{C}_{\varepsilon}^{n+1^{-1}}\right)^{-1}\mathbf{C}_{\varepsilon}^{n+1^{-1}}\hat{\mathbf{v}}^{j,n+1}+\left(\mathbf{H}^{\intercal}\boldsymbol{\Gamma}\mathbf{H}+\mathbf{C}_{\varepsilon}^{n+1^{-1}}\right)^{-1}\mathbf{H}^{\intercal}\boldsymbol{\Gamma}\mathbf{y}^{n+1}\\
&-\left(\mathbf{H}^{\intercal}\boldsymbol{\Gamma}\mathbf{H}+\mathbf{C}_{\varepsilon}^{n+1^{-1}}\right)^{-1}\mathbf{J}_{\tilde{\mathcal{A}}}^{\intercal}(\mathbf{v}^{j,n+1})\boldsymbol{\lambda}^{j,n+1}\\
=&\left(\mathbf{C}_{\varepsilon}^{n+1}-\mathbf{C}_{\varepsilon}^{n+1}\mathbf{H}^{\intercal}\left(\boldsymbol{\Gamma}^{-1}+\mathbf{H}\mathbf{C}_{\varepsilon}^{n+1}\mathbf{H}^{\intercal}\right)^{-1}\mathbf{H}\mathbf{C}_{\varepsilon}^{n+1}\right)\mathbf{C}_{\varepsilon}^{n+1^{-1}}\hat{\mathbf{v}}^{j,n+1}\\
&+\mathbf{C}_{\varepsilon}^{n+1}\mathbf{H}^{\intercal}\left(\mathbf{H}\mathbf{C}_{\varepsilon}^{n+1}\mathbf{H}^{\intercal}+\boldsymbol{\Gamma}^{-1}\right)^{-1}\mathbf{y}^{n+1}\\
&-\left(\mathbf{C}_{\varepsilon}^{n+1}-\mathbf{C}_{\varepsilon}^{n+1}\mathbf{H}^{\intercal}\left(\boldsymbol{\Gamma}^{-1}+\mathbf{H}\mathbf{C}_{\varepsilon}^{n+1}\mathbf{H}^{\intercal}\right)^{-1}\mathbf{H}\mathbf{C}_{\varepsilon}^{n+1}\right)\mathbf{J}_{\tilde{\mathcal{A}}}^{\intercal}(\mathbf{v}^{j,n+1})\boldsymbol{\lambda}^{j,n+1}\\
=&\hat{\mathbf{v}}^{j,n+1}-\mathbf{C}_{\varepsilon}^{n+1}\mathbf{H}^{\intercal}\left(\boldsymbol{\Gamma}^{-1}+\mathbf{H}\mathbf{C}_{\varepsilon}^{n+1}\mathbf{H}^{\intercal}\right)^{-1}\mathbf{H}\hat{\mathbf{v}}^{j,n+1}\\
&+\mathbf{C}_{\varepsilon}^{n+1}\mathbf{H}^{\intercal}\left(\mathbf{H}\mathbf{C}_{\varepsilon}^{n+1}\mathbf{H}^{\intercal}+\boldsymbol{\Gamma}^{-1}\right)^{-1}\mathbf{y}^{n+1}\\
&-\left(\mathbf{C}_{\varepsilon}^{n+1}-\mathbf{C}_{\varepsilon}^{n+1}\mathbf{H}^{\intercal}\left(\boldsymbol{\Gamma}^{-1}+\mathbf{H}\mathbf{C}_{\varepsilon}^{n+1}\mathbf{H}^{\intercal}\right)^{-1}\mathbf{H}\mathbf{C}_{\varepsilon}^{n+1}\right)\mathbf{J}_{\tilde{\mathcal{A}}}^{\intercal}(\mathbf{v}^{j,n+1})\boldsymbol{\lambda}^{j,n+1}\\
=&\hat{\mathbf{v}}^{j,n+1}+\mathbf{C}_{\varepsilon}^{n+1}\mathbf{H}^{\intercal}\left(\mathbf{H}\mathbf{C}_{\varepsilon}^{n+1}\mathbf{H}^{\intercal}+\boldsymbol{\Gamma}^{-1}\right)^{-1}(\mathbf{y}^{n+1}-\mathbf{H}\hat{\mathbf{v}}^{j,n+1})\\
&-\left(\mathbf{C}_{\varepsilon}^{n+1}-\mathbf{C}_{\varepsilon}^{n+1}\mathbf{H}^{\intercal}\left(\mathbf{H}\mathbf{C}_{\varepsilon}^{n+1}\mathbf{H}^{\intercal}+\boldsymbol{\Gamma}^{-1}\right)^{-1}\mathbf{H}\mathbf{C}_{\varepsilon}^{n+1}\right)\mathbf{J}_{\tilde{\mathcal{A}}}^{\intercal}(\mathbf{v}^{j,n+1})\boldsymbol{\lambda}^{j,n+1}
\end{aligned}$$

- From equation (13) to the constrained EnKF updated formula (11).

  We multiply equation (13) by the observation matrix $\mathbf{H}^{\perp}$

$$\mathbf{u}^{j,n+1} = \mathbf{H}^{\perp}\mathbf{v}^{j,n+1} = \mathbf{H}^{\perp}\hat{\mathbf{v}}^{j,n+1} + \mathbf{H}^{\perp}\mathbf{C}_{\varepsilon}^{n+1}\mathbf{H}^{\intercal}\left(\mathbf{H}\mathbf{C}_{\varepsilon}^{n+1}\mathbf{H}^{\intercal} + \mathbf{\Gamma}^{-1}\right)^{-1}(\mathbf{y}^{n+1} - \mathbf{H}\hat{\mathbf{v}}^{j,n+1})$$
$$- \mathbf{H}^{\perp}\left(\mathbf{C}_{\varepsilon}^{n+1} - \mathbf{C}_{\varepsilon}^{n+1}\mathbf{H}^{\intercal}\left(\mathbf{H}\mathbf{C}_{\varepsilon}^{n+1}\mathbf{H}^{\intercal} + \mathbf{\Gamma}^{-1}\right)^{-1}\mathbf{H}\mathbf{C}_{\varepsilon}^{n+1}\right)\mathbf{J}_{\tilde{\mathcal{A}}}^{\intercal}(\mathbf{v}^{j,n+1})\boldsymbol{\lambda}^{j,n+1}.$$

and, since the map $\mathbf{M} \to \mathbf{M}^{-1}$ is continuous over the set of invertible matrices, letting $\varepsilon \to 0$

$$\mathbf{u}^{j,n+1} = \mathbf{H}^{\perp}\hat{\mathbf{v}}^{j,n+1} + \mathbf{H}^{\perp}\mathbf{C}^{n+1}\mathbf{H}^{\intercal}\left(\mathbf{H}\mathbf{C}^{n+1}\mathbf{H}^{\intercal} + \mathbf{\Gamma}^{-1}\right)^{-1}(\mathbf{y}^{n+1} - \mathbf{H}\hat{\mathbf{v}}^{j,n+1})$$
$$- \mathbf{H}^{\perp}\left(\mathbf{C}^{n+1} - \mathbf{C}^{n+1}\mathbf{H}^{\intercal}\left(\mathbf{H}\mathbf{C}^{n+1}\mathbf{H}^{\intercal} + \mathbf{\Gamma}^{-1}\right)^{-1}\mathbf{H}\mathbf{C}^{n+1}\right)\mathbf{J}_{\tilde{\mathcal{A}}}^{\intercal}(\mathbf{v}^{j,n+1})\boldsymbol{\lambda}^{j,n+1}.$$

Noticing that

$$\mathbf{H}^{\perp}\hat{\mathbf{v}}^{j,n+1} = \mathbf{u}^{j,n}, \quad \mathbf{H}^{\perp}\mathbf{C}^{n+1}\mathbf{H}^{\intercal} = \mathbf{C}_{\mathbf{uw}}^{n+1}, \quad \mathbf{H}\mathbf{C}^{n+1}\mathbf{H}^{\intercal} = \mathbf{C}_{\mathbf{ww}}^{n+1}, \quad \mathbf{H}\hat{\mathbf{v}}^{j,n+1} = \mathcal{G}(\mathbf{u}^{j,n})$$
$$\mathbf{H}\mathbf{C}^{n+1}\mathbf{J}_{\tilde{\mathcal{A}}}^{\intercal}(\mathbf{v}^{j,n+1}) = \mathbf{C}_{\mathbf{uw}}^{n+1\intercal}\mathbf{J}_{\mathcal{A}}^{\intercal}(\mathbf{u}^{j,n+1}), \quad \mathbf{H}^{\perp}\mathbf{C}^{n+1}\mathbf{J}_{\tilde{\mathcal{A}}}^{\intercal}(\mathbf{v}^{j,n+1}) = \mathbf{C}_{\mathbf{uu}}^{n+1}\mathbf{J}_{\mathcal{A}}^{\intercal}(\mathbf{u}^{j,n+1})$$

we finally obtain

$$\mathbf{u}^{j,n+1} = \mathbf{u}^{j,n} + \mathbf{C}_{\mathbf{uw}}^{n+1}(\mathbf{C}_{\mathbf{ww}}^{n+1} + \mathbf{\Gamma}^{-1})^{-1}(\mathbf{y}^{n+1} - \mathcal{G}(\mathbf{u}^{j,n}))$$
$$+ \mathbf{C}_{\mathbf{uw}}^{n+1}(\mathbf{C}_{\mathbf{ww}}^{n+1} + \mathbf{\Gamma}^{-1})^{-1}\mathbf{C}_{\mathbf{uw}}^{n+1\intercal}\mathbf{J}_{\mathcal{A}}^{\intercal}(\mathbf{u}^{j,n+1})\boldsymbol{\lambda}^{j,n+1} - \mathbf{C}_{\mathbf{uu}}^{n+1}\mathbf{J}_{\mathcal{A}}^{\intercal}(\mathbf{u}^{j,n+1})\boldsymbol{\lambda}^{j,n+1}.$$

Appendix B. List of most used symbols

Symbol	Description
$\mathbb{R}$, $\mathbb{N}$	Spaces of real numbers and positive integers
$\mathbf{y} \in \mathbb{R}^K$, $K \in \mathbb{N}$	Observations or measurements
$\mathbf{u} \in \mathbb{R}^d$, $d \in \mathbb{N}$	Unknown control
$\mathcal{G} : \mathbb{R}^d \to \mathbb{R}^K$	Forward model
$\mathbf{G} : \mathbb{R}^d \to \mathbb{R}^K$	Forward linear model
$\mathbf{\Gamma}^{-1} \in \mathbb{R}^{K\times K}$	Covariance matrix of the noise
$\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Gamma}^{-1})$	Normally distributed measurement noise
$\Phi(\mathbf{u}, \mathbf{y}) = \frac{1}{2}\|\mathbf{\Gamma}^{\frac{1}{2}}(\mathbf{y} - \mathcal{G}(\mathbf{u}))\|^2$	Least square functional
$\mathcal{A} : \mathbb{R}^d \to \mathbb{R}^m$, $m \in \mathbb{N}$	Constraint functions
$\tilde{\mathcal{A}} : \mathbb{R}^{d+K} \to \mathbb{R}^m$	Extension of the constraint functions to the space $\mathbb{R}^{d+K}$
$\mathbf{A} \in \mathbb{R}^{m\times d}$	Linear constraint functions
$\mathbf{J}_{\mathcal{A}} \in \mathbb{R}^{m\times d}$	Jacobian matrix of the constraints
$\mathbf{u}^j \in \mathbb{R}^d, \quad j \in \{1, \ldots, J\}$	jth ensemble member
$\boldsymbol{\lambda}^j \in \mathbb{R}^{\cdot m}, \quad j \in \{1, \ldots, J\}$	Multiplier related to the jth ensemble member
$\mathbf{C}_{\mathbf{uu}} \in \mathbb{R}^{d\times d}$	Empirical covariance of the ensemble
$\mathbf{C}_{\mathbf{uw}} \in \mathbb{R}^{d\times K}$	Empirical covariance of the ensemble and its image with respect the forward model
$\mathbf{C}_{\mathbf{ww}} \in \mathbb{R}^{K\times K}$	Empirical covariance of the image of the ensemble with respect the forward model
$\overline{\mathbf{m}} \in \mathbb{R}^d$	First moment (or mean) of the ensemble
$\overline{\overline{\mathbf{m}}} \in \mathbb{R}^{d\times d}$	Second moment of the ensemble
$\mathbf{e}^j \in \mathbb{R}^d$	jth ensemble spread to the mean
$\mathbf{r}^j \in \mathbb{R}^d$	jth ensemble residual
$\boldsymbol{\vartheta}^j \in \mathbb{R}^K$	jth ensemble misfit

ORCID iDs

Giuseppe Visconti https://orcid.org/0000-0003-2899-7167

References

[1] Engl H W, Hanke M and Neubauer A 1996 *Regularization of Inverse Problems* (Berlin: Springer) vol 375

[2] Evensen G 1994 Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics *J. Geophys. Res.* **99** 10143–62

[3] Bloemker D, Schillings C and Wacker P 2018 A strongly convergent numerical scheme from ensemble Kalman inversion *SIAM J. Numer. Anal.* **56** 2537–62

[4] Bloemker D, Schillings C, Wacker P and Weissman S 2019 Well posedness and convergence analysis of the ensemble Kalman inversion *Inverse Problems* **35** 085007

[5] Chada N K, Stuart X T and Tong A M 2019 Tikhonov regularization within ensemble Kalman inversion submitted (arXiv:1901.10382)

[6] Schillings C and Stuart A M 2017 Analysis of the ensamble Kalman filter for inverse problems *SIAM J. Numer. Anal.* **55** 1264–90

[7] Schillings C and Stuart A M 2018 Convergence analysis of ensemble Kalman inversion: the linear, noisy case *Appl. Anal.* **97** 107–23

[8] Carrillo J A and Vaes U 2019 Wasserstein stability estimates for covariance-preconditioned Fokker–Planck equations submitted (arXiv:1910.07555)

[9] Ding Z and Li Q 2019 Ensemble Kalman inversion: mean-field limit and convergence analysis (arXiv:1908.05575)

[10] Garbuno-Inigo A, Hoffmann F, Li W and Stuart A M 2019 Interacting Langevin diffusions: gradient structure and ensemble Kalman sampler (arXiv:1903.08866)

[11] Herty M and Visconti G 2019 Kinetic methods for inverse problems *Kinet. Relat. Models* **12** 1109–30

[12] Janjiè T, McLaughlin D, Cohn S E and Verlaan M 2014 Conservation of mass and preservation of positivity with ensemble-type Kalman filter algorithms *Mon. Weather Rev.* **142** 755–73

[13] Schwenzer M, Visconti G, Ay M, Bergs T, Herty M and Abel D 2019 Identifying trending coefficients with an ensemble Kalman filter submitted (arXiv:2001.10861)

[14] Teixeira B O S, Târres L A B, Aguirre L A and Bernstein D S 2010 On unscented Kalman filtering with state interval constraints *J. Process Control* **20** 45–57

[15] Aravkin A Y, Burke J V and Pillonetto G 2014 Optimization viewpoint on Kalman smoothing with applications to robust and sparse estimation *Compressed Sensing & Sparse Filtering* (Berlin: Springer) pp 237–80

[16] Albers D J, Blancquart P-A, Levine M E, Seylabi E E and Stuart A M 2019 Ensemble Kalman methods with constraints *Inverse Problems* **35** 095007

[17] Chada N K, Schillings C and Weissmann S 2019 On the incorporation of box-constraints for ensemble Kalman inversion *Foundations of Data Science* **1** 433

[18] Anderson J L 2001 An ensemble adjustment Kalman filter for data assimilation *Mon. Weather Rev.* **129** 2884–903

[19] Iglesias M, Law K and Stuart A M 2013 Analysis of the Ensemble Kalman methods for inverse problems *Inverse Problems* **29** 045001

[20] Fiacco A V and McCormick G P 1968 *Nonlinear Programming: Sequential Unconstrained Minimization Techniques* (New York: Wiley)

[21] Kuhn H W and Tucker A W 1951 Nonlinear programming *Proc. of the 2nd Berkeley Symp. on Mathematical Statistics and Probability* (Berkeley: University of California Press) pp 481–92

[22] Nikaidô H and Isoda K 1955 Note on non-cooperative convex games *Pac. J. Math.* **5** 807–15

[23] Bressan A 2010 *Noncooperative Differential Games A Tutorial Technical Report* (Department of Mathematics, Penn State University)

[24] Facchinei F and Pang J-S 2003 *Finite-Dimensional Variational Inequalities and Complementarity Problems* (Berlin: Springer)

[25] Schwartz A 2017 *Lecture Game Theory Technical Report* (Technische Universität Darmstadt)

[26] Gerdts M 2012 *Optimal Control of ODEs and DAEs* (Berlin: De Gruyter)

[27] Tong X T, Majda A and Kelly D 2016 Nonlinear stability and ergodicity of ensemble based Kalman filters *Nonlinearity* **29** 657

[28] Ernst O G, Sprungk B and Starkloff H-J 2015 Analysis of the ensemble and polynomial chaos Kalman filters in bayesian inverse problems *SIAM/ASA J. Uncertain. Quantification* **3** 823–51

[29] Kwiatkowski E and Mandel J 2015 Convergence of the square root ensemble Kalman filter in the large ensemble limit *SIAM/ASA J. Uncertain. Quantification* **3** 1–17
[30] Law K J H, Tembine H and Tempone R 2016 Deterministic mean-field ensemble Kalman filtering *SIAM J. Sci. Comput.* **38** A1251–79
[31] Le Gland F, Monbet V and Tran V-D 2009 Large sample asymptotics for the ensemble Kalman filter *Research Report RR-7014* (INRIA)
[32] Carrillo J A, Fornasier M, Toscani G and Vecil F 2010 Mathematical modeling of collective behavior in socio-economic and life sciences, chapter particle, kinetic, and hydrodynamic models of swarming *Modeling and Simulation in Science, Engineering and Technology* (Boston: Birkhäuser) pp 297–336
[33] Ha S-Y and Tadmor E 2008 From particle to kinetic and hydrodynamic descriptions of flocking *Kinet. Relat. Models* **3** 415–35
[34] Pareschi L and Toscani G 2013 *Interacting Multiagent Systems. Kinetic Equations and Monte Carlo Methods* (Oxford: Oxford University Press)
[35] Toscani G 2006 Kinetic models of opinion formation *Commun. Math. Sci.* **4** 481–96
[36] Brenan K E, Campbell S L S L and Petzold L R 1996 *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations* 2nd edn (Philadelphia, PA: SIAM)
[37] Hairer E and Wanner G 1996 *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems* (Springer Series in Computational Mathematics) 2nd edn