On the ensemble Kalman inversion under inequality constraints

Matei Hanu[1],[*] and Simon Weissmann[2]

[1] Freie Universität Berlin, Fachbereich Mathematik und Informatik, 14195 Berlin, Germany
[2] Universität Mannheim, Institut für Mathematik, 68138 Mannheim, Germany

E-mail: matei.hanu@fu-berlin.de and simon.weissmann@uni-mannheim.de

Abstract

The ensemble Kalman inversion (EKI), a recently introduced optimisation method for solving inverse problems, is widely employed for the efficient and derivative-free estimation of unknown parameters. Specifically in cases involving ill-posed inverse problems and high-dimensional parameter spaces, the scheme has shown promising success. However, in its general form, the EKI does not take constraints into account, which are essential and often stem from physical limitations or specific requirements. Based on a log-barrier approach, we suggest adapting the continuous-time formulation of EKI to incorporate convex inequality constraints. We underpin this adaptation with a theoretical analysis that provides lower and upper bounds on the ensemble collapse, as well as convergence to the constraint optimum for general nonlinear forward models. Finally, we showcase our results through two examples involving partial differential equations.

Keywords: ensemble Kalman inversion, Tikhonov regularisation, derivative-free optimisation, convex inequality constraints

1. Introduction

Mathematical models have been employed to describe a wide range of physical, biological, and social systems and processes, enabling the analysis and prediction of their behaviors. When applying a model to a specific system, calibration becomes crucial, aligning the model with

---

[*] Author to whom any correspondence should be addressed.

observational data. This calibration, often referred to as inversion, serves as the foundation for various applications such as numerical weather prediction, medical image processing, and numerous machine learning methods. In the realm of inversion techniques, two prominent categories are variational/optimisation-based approaches and Bayesian/statistical approaches. This work focuses on a method that bridges these two approaches—the ensemble Kalman inversion (EKI) framework introduced in [27, 35]. EKI utilizes an ensemble Kalman-Bucy filter iteratively to address inverse problems.

Nevertheless, the fundamental version of EKI lacks the capability to integrate additional constraints on parameters. Such constraints frequently emerge in various applications due to additional insights into the system. Since the estimation of the EKI is usually not feasible, incorporating constraints into the EKI is a significant task. The subsequent discussion will specifically address the efficient integration of convex inequality constraints to the EKI. One approach to incorporate constraints to the EKI has been done in [14]. Here, the authors introduce a projection method in the discrete-time EKI and derive a continuous-time limit. Since the resulting continuous-time limit admits discontinuities in the right hand side, the authors proposed a smoothed system using a log-barrier approach and derive convergence for linear forward models.

In the following paper, we extend the log-barrier approach from [14] to a broader class of convex inequality constraints. Moreover, we also incorporate Tikhonov regularisation and provide a convergence analysis for nonlinear forward models.

1.1. Literature overview

From its very beginning [20], the ensemble Kalman filter (EnKF) has found extensive use in both inverse problems and data assimilation scenarios. Its widespread application is attributed to its easy implementation and resilience with respect to small ensemble sizes [4, 5, 25–27, 31]. Stability has been addressed in works like [40, 41], and convergence analysis, grounded in the continuous-time limit of EKI, has been developed in [8, 9, 11, 35, 36]. However, achieving convergence results in the parameter space often necessitates some form of regularisation. We primarily focus on Tikhonov regularisation, extensively analyzed for EKI in [15], for instance. Recent advancements include further analysis on Tikhonov regularisation for stochastic EKI and adaptive Tikhonov strategies to enhance the original variant [45]. In the context of large ensemble sizes, a mean-field limit analysis is presented in [12, 18].

The addition of constraints to the EKI has been analysed more and more in recent years, for example [1, 24]. An extensive survey of existing methodologies for handling linear and nonlinear constraints in Kalman-based methods is available in [2, 39]. One popular method is to project the estimates of the EKI into the feasible set [28, 43]. The advantage of this method that the theory can be expanded to non-linear constraints.

Many of these variations find motivation by interpreting the updates in Kalman-based methods as solutions to corresponding optimisation problems. For additional insights, refer to [2].

For our analysis we consider preconditioned gradient methods that are based on [6, 7, 37, 38]. The idea of preconditioning by the sample covariance matrix is also applied in other particle methods for inverse problems such as interacting Langevin dynamic [19, 21, 22], Consensus based sampling [13] and Fokker–Planck dynamics [33, 34]. In these applications the preconditioning through the sample covariance is often motivated by the property of remaining invariant under affine coordinate transformations which is beneficial for solving ill-conditioned optimisation problems.

Finally, we will also incorporate covariance inflation into our algorithm. While the ensemble collapse, which is the convergence of the particles to their mean value, leads to an improvement in the gradient approximation, it also leads to a degeneration of the preconditioner and therefore the EKI may get stuck in a solution that is far from global optimality. Variance inflation is a tool which allows us to control the speed of the collapse [3, 40].

1.2. Preliminaries

In the present paper, we consider an inverse problem of the following form

$$y = G(u) + \eta. \tag{1.1}$$

The goal is to recover the unknown parameter $u \in X$, where $y \in \mathbb{R}^K$ denotes the observed data and $\eta \sim \mathcal{N}(0, \Gamma)$ is Gaussian additive observational noise with $\Gamma \in \mathbb{R}^{K \times K}$ symmetric positive definite. Moreover, we consider a possibly nonlinear forward map $G : X \to \mathbb{R}^K$ mapping from the parameter space X to the observation space $\mathbb{R}^K$. Throughout the manuscript we assume that the parameter space is finite dimensional given by $X := \mathbb{R}^d$. We follow a minimisation based approach to solve the inverse problem, where our goal is to find a minimiser of the Tikhonov regularised potential

$$\Phi^{\text{reg}}(u) := \frac{1}{2} \|G(u) - y\|_\Gamma^2 + \frac{\lambda}{2} \|u\|_{C_0}^2. \tag{1.2}$$

Here, $C_0 \in \mathbb{R}^{d \times d}$ denotes a symmetric positive definite regularisation matrix and $\lambda > 0$ is a regularisation parameter. Moreover, given a symmetric positive definite matrix $\Sigma \in \mathbb{R}^{K \times K}$ we define the rescaled norm $\|x\|_\Sigma := \langle x, \Sigma^{-1} x \rangle$, $x \in \mathbb{R}^K$, where $\langle \cdot, \cdot \rangle$ denotes the euclidean inner product over $\mathbb{R}^K$. Moreover, we will use $\| \cdot \|_F$ to denote the Frobenius norm. In the following, assume that the observation y, the regularisation matrix C_0 and the regularisation parameter λ are given and we do not consider a parameter choice rule for λ and C_0. Hence, we suppress the dependence of Φ^{reg} on these quantities. In this work we follow the approach of applying EKI as derivative-free optimisation method for finding a minimiser of the potential Φ^{reg}.

1.3. Inverse problem under constraints

In many practical scenarios the unknown parameter $u \in X$ is subjected to physical constraints. In what follows, we will assume that the set of feasible parameters is given as set of inequality constraints of the form

$$\Omega = \{u \in X : h_j(u) \leqslant 0, j = 1, \ldots, m\}, \tag{1.3}$$

where $h_j : \mathbb{R}^d \to \mathbb{R}$ are convex functions for all $j = 1, \ldots, m$. Note that one important example would be the set of linear inequality constraints

$$\Omega = \{u \in X : \langle c_j, u \rangle + \delta_j \leqslant 0, j = 1, \ldots, m\},$$

where $c_j = \pm e_j$, $\delta_j \in \mathbb{R}$, $j = 1, \ldots, m$ and e_j denotes the j-th unit vector. Then δ_j denotes an upper and lower bound on the j-th component of u. Defining $h_j(u) = \langle c_j, u \rangle + \delta_j$ for all $j = 1, \ldots, m$ we are back in the representation (1.3). Our goal is to solve the inverse problem (1.1) under convex-constraints $u \in \Omega$ by solving the constrained optimisation problem

$$\min_{u \in \Omega} \Phi^{\text{reg}}(u). \tag{1.4}$$

We make the following assumption on the considered objective function Φ^{reg}.

Assumption 1.1. The functional Φ^{reg} is $C^2(X, \mathbb{R}_+)$ as well as

(i) μ-strongly convex, i.e. there exists $\mu > 0$ such that

$$\Phi^{\text{reg}}(x_1) - \Phi^{\text{reg}}(x_2) \geqslant \langle \nabla \Phi^{\text{reg}}(x_2), x_1 - x_2 \rangle + \frac{\mu}{2} \|x_1 - x_2\|^2, \quad \text{for all } x_1, x_2 \in X.$$

(ii) L-smooth, i.e. there exists $L > 0$ such that the gradient $\nabla \Phi^{\text{reg}}$ is global L-Lipschitz continuous:

$$\|\nabla \Phi^{\text{reg}}(x_1) - \nabla \Phi^{\text{reg}}(x_2)\| \leqslant L \|x_1 - x_2\| \quad \text{for all } x_1, x_2 \in X.$$

We note that while the above assumptions are formulated globally, our theoretical analysis requires these assumptions only locally in Ω. The smoothness property implies the following useful descent condition

$$\Phi^{\text{reg}}(x_1) - \Phi^{\text{reg}}(x_2) \leqslant \langle \nabla \Phi^{\text{reg}}(x_2), x_1 - x_2 \rangle + \frac{L}{2} \|x_1 - x_2\|^2, \quad \text{for all } x_1, x_2 \in X$$

which is a standard property used to prove convergence of first order optimisation methods. Moreover, since Φ^{reg} is assumed to be μ-strongly convex it also satisfies the Polyak-Łojasiewic (PL) inequality of form

$$\nu \|\nabla \Phi^{\text{reg}}(x)\|^2 \geqslant \Phi^{\text{reg}}(x) - \Phi^{\text{reg}}(x^*),$$

for some $\nu > 0$ and all $x \in X$, where x^* is the unique global minimiser of Φ^{reg}. We provide a detailed derivation in lemma A.1. It is well-known that under L-smoothness and the above PL inequality the gradient descent method converges linearly towards the unique minimiser x^* [29].

Remark 1.2. Usually the strong convexity of (1.2), that we assume in assumption 1.1 can be achieved through large enough regularisation parameter λ. However, due to this large choice of λ the computed solution of the regularised problem might be irrelevant for the initial potential $\frac{1}{2} \|G(u) - y\|_\Gamma^2$. There have been several discussions on the strong convexity of Tikhonov regularisation as well as other forms of regularisations for which we refer to [17, 30].

In section 5.2, we present a nonlinear example with Tikhonov regularisation to obtain a strongly convex potential.

Note that under assumption 1.1 the optimisation problem (1.4) is convex and therefore any point $u^* \in \Omega$ that satisfies the Karuhn-Kash-Tucker (KKT) conditions, i.e. there exists $\lambda^* \in \mathbb{R}^m$ such that $u^* \in \Omega$ satisfying

- $\nabla \Phi^{\text{reg}}(u^*) + \sum_{j=1}^m \lambda_j^* \nabla h_j(u^*) = 0$,
- $\lambda_j^* h_j(u^*) \leqslant 0$ for all $j \in \{1, \dots, m\}$,
- $\lambda_j^* \geqslant 0$ for all $j \in \{1, \dots, m\}$ and $\sum_{j=1}^m \lambda_j^* h_j(u^*) = 0$,

is the unique global minimiser of Φ^{reg}, also called the KKT point of (1.4).

1.4. Our contribution

In this manuscript we will apply the EKI as derivative-free optimisation method for solving (1.2) under convex inequality constraints on $u \in X$. The EKI for solving optimisation problems under box-constraints has been introduced in [14], where the authors provide a convergence analysis for linear forward operators without regularisation. The purpose of this work is to make use of the gradient flow structure of EKI presented in [44] for extending the convergence analysis of linear EKI under box-constraints to nonlinear EKI under convex inequality constraints. Moreover, our proposed scheme allows to incorporate Tikhonov regularisation. We make the following contribution:

- We suggest a new adaptation of EKI, enabling the integration of convex inequality constraints on the unknown parameters using a log-barrier penalty approach. The adaptation incorporates Tikhonov regularisation as well as covariance inflation.
- Under strong convexity and smoothness, we provide a convergence analysis of our adaptation, where we analyse feasibility, ensemble collapse and convergence to the unique KKT-point.
- We demonstrate our findings through two examples based on partial differential equations (PDEs). In order to keep the implementation of the proposed scheme efficient we apply an adaptively increasing penalty parameter during the algorithm.

The paper is organised as follows. We introduce our adaptation in section 2. In section 3 we quantify the ensemble collapse of our scheme as well as verify the feasibility of the computed solutions. We analyse the convergence of our scheme in section 4; before presenting our numerical experiments in section 5. Finally, we summarise our work with a conclusion in section 6.

2. Ensemble Kalman inversion

We consider the EKI to solve the inverse problem (1.1), where we will focus on the continuous-time limit of the scheme, cp. [36]. The framework is the following.

Firstly, we define an initial ensemble $u_0 = (u_0^{(j)})_{j=1,\dots,J}$, $u_0^{(j)} \in X$, $j = 1, \dots, J$, of size $J \geqslant 2$. In the continuous-time formulation of EKI, the particle system $u_t = (u_t^{(j)})_{j=1,\dots,J}$ then moves according to the dynamical system

$$\frac{\mathrm{d}u_t^{(j)}}{\mathrm{d}t} = -\widehat{C}_t^{u,G} \Gamma^{-1} \left(G\left(u_t^{(j)}\right) - y \right), \qquad j = 1, \dots, J \tag{2.1}$$

with initial state u_0. Here, we have defined the following empirical means and covariances within the particle system u_t, $t \geqslant 0$,

$$\widehat{C}_t^{u,G} := \frac{1}{J} \sum_{j=1}^{J} (u_t^{(j)} - \overline{u}_t) \otimes (G(u_t^{(j)}) - (\overline{G}(u_t)),$$

$$\overline{u}_t = \frac{1}{J} \sum_{j=1}^{J} u_t^{(j)}, \quad \overline{G}(u_t) = \frac{1}{J} \sum_{j=1}^{J} G(u_t^{(j)}).$$

Following [15] we can incorporate Tikhonov regularisation into the particle system using the time evolution

$$\frac{\mathrm{d}u_t^{(j)}}{\mathrm{d}t} = -\widehat{C}_t^{u,G}\Gamma^{-1}\left(G\left(u_t^{(j)}\right) - y\right) - \widehat{C}\left(u_t\right)\lambda^{1/2}C_0^{-1}u_t^{(j)}, \qquad j = 1,\ldots,J, \tag{2.2}$$

where $\lambda > 0$ and $C_0 \in \mathbb{R}^{d \times d}$ symmetric positive definite. The values correspond to the regularisation parameters introduced in (1.2). We note that both dynamical systems (2.1) and (2.2) may be implemented more efficiently by utilising the representations

$$\widehat{C}_t^{u,G}\Gamma^{-1}(G(u_t^{(j)}) - y) = \frac{1}{J}\sum_{k=1}^J \langle G(u_t^{(j)}) - y, G(u_t^{(k)} - \bar{G}_t\rangle_\Gamma (u_t^{(k)} - \bar{u}),$$

$$\widehat{C}(u_t)C_0^{-1}u_t^{(j)} = \langle u_t^{(j)}, u_t^{(k)} - \bar{u}_t\rangle_{C_0}(u_t^{(k)} - \bar{u}),$$

which avoids the explicit computation of the sample covariance matrix.

In case of a linear forward map G the EKI dynamics (2.2) can be written as system of gradient flows

$$\frac{\mathrm{d}u_t^{(j)}}{\mathrm{d}t} = -\widehat{C}\left(u_t\right)\nabla\Phi^{\mathrm{reg}}\left(u^{(j)}\right). \tag{2.3}$$

which are coupled through an adaptive preconditioner given by the empirical covariance matrix $\widehat{C}(u_t)$. In case of a nonlinear forward operators, this representation in general only holds approximately. Indeed, by Taylors theorem it can be justified that $\widehat{C}_t^{u,G} \approx \widehat{C}(u_t)DG^*(u^{(j)})$, where DG denotes the (Frechet) derivative of G (see [44, lemma 4.5]). It follows that (2.2) can be viewed as derivative-free approximation of the preconditioned gradient flow (2.3) also in the nonlinear setting. To be more precise, we make the following sufficient assumptions on the nonlinear forward map to justify the approximate gradient flow structure of EKI.

Assumption 2.1. The forward operator G is locally Lipschitz continuous, with constant $c_{\mathrm{lip}} > 0$ and satisfies the linear approximation property

$$G\left(x_1\right) = G\left(x_2\right) + DG\left(x_2\right)\left(x_1 - x_2\right) + \mathrm{Res}\left(x_1,x_2\right), \tag{2.4}$$

for all $x_1, x_2 \in X$. The approximation error is bounded by

$$\|\mathrm{Res}\left(x_1,x_2\right)\|_2 \leqslant b_{\mathrm{res}}\|x_1 - x_2\|_2^2, \tag{2.5}$$

where $b_{\mathrm{res}} > 0$ is independent of x_1 and x_2.

2.1. Ensemble Kalman inversion under box constraints

In the following, we will revisit the EKI Algorithm under box-constraints and modify the formulation with the goal of minimising (1.2) under box-constraints. We refer to [14, section 3] for more details. In the following, we consider the constrained optimisation problem

$$\min_{u \in \mathcal{B}} \Phi\left(u\right), \quad \Phi\left(u\right) := \frac{1}{2}\|G\left(u\right) - y\|_\Gamma^2, \tag{2.6}$$

where $\mathcal{B} = \{u \in \mathbb{R}^d : a_i \leqslant u_i \leqslant b_i, i = 1,\ldots,m\}, m \leqslant d,$ denotes the considered box. In [14] the authors incorporate the box-constraints into the algorithm using an element-wise projection into the box defined as

$$
(\mathcal{P}(u))_i = \begin{cases} a_i, & \text{if } u_i < a_i, \\ u_i, & \text{if } u_i \in [a_i, b_i], \quad i = 1,\ldots,m, \\ b_i, & \text{if } u_i > b_i, \end{cases}
$$
$$
(\mathcal{P}(u))_i = u_i \quad i = m+1,\ldots,n.
$$

Following the idea of projected gradient methods [6, 38], EKI under box-constraints in discrete time proceeds by projecting the ensemble of particles into the box as introduced in [14]. The authors define the following variables

$$
u_{n,p}^{(j)} = \mathcal{P}\left(u_n^{(j)}\right), \quad \bar{u}_{\mathcal{P}} = \frac{1}{J}\sum_{j=1}^{J}\left(u_{n,p}^{(j)}\right), \quad \bar{G}_{\mathcal{P}} = \frac{1}{J}\sum_{j=1}^{J}G\left(u_{n,p}^{(j)}\right),
$$

as well as the empirical covariance matrices

$$
\widehat{C}_{n,\mathcal{P}}^{u,u} := \frac{1}{J}\sum_{j=1}^{J}\left(u_{n,p}^{(j)} - \bar{u}_{\mathcal{P}}\right) \otimes \left(u_{n,p}^{(j)} - \bar{u}_{\mathcal{P}}\right),
$$

$$
\widehat{C}_{n,\mathcal{P}}^{u,G} := \frac{1}{J}\sum_{j=1}^{J}\left(u_{n,p}^{(j)} - \bar{u}_{\mathcal{P}}\right) \otimes \left(G\left(u_{n,p}^{(j)}\right) - \bar{G}_{\mathcal{P}}\right),
$$

$$
\widehat{C}_{n,\mathcal{P}}^{G,G} := \frac{1}{J}\sum_{j=1}^{J}\left(G\left(u_{n,p}^{(j)}\right) - \bar{G}_{\mathcal{P}}\right) \otimes \left(G\left(u_{n,p}^{(j)}\right) - \bar{G}_{\mathcal{P}}\right),
$$

The update formula of the EKI under box-constraints then consists of a prediction step and a projection onto the box $\mathcal{B}$.

$$
\begin{cases} \tilde{u}_{n+1,p}^{(j)} = u_{n,p}^{(j)} + \widehat{C}_{n,\mathcal{P}}^{u,G}\left(\widehat{C}_{n,\mathcal{P}}^{G,G} + h^{-1}\Gamma\right)^{-1}\left(y - G\left(u_{n,p}^{(j)}\right)\right), \\ u_{n+1,p}^{(j)} = \mathcal{P}\left(\tilde{u}_{n+1,p}^{(j)}\right). \end{cases} \tag{2.7}
$$

Taking the limit $h \to 0$ one obtains the continuous-time formulation of EKI under box-constraints given by a coupled system of ODEs

$$
\left(\frac{\mathrm{d}u_t^{(j)}}{\mathrm{d}t}\right)_i = \begin{cases} v_i(u_t^{(j)}), & (u_t^{(j)})_i \in (a_i, b_i), \\ \mathbb{1}_{[0,\infty)}(v_i(u_t^{(j)}))v_i(u_t^{(j)}), & (u_t^{(j)})_i = a_i, \quad i = 1,\ldots,m, \\ \mathbb{1}_{[-\infty,0)}(v_i(u_t^{(j)}))v_i(u_t^{(j)}), & (u_t^{(j)})_i = b_i, \end{cases}
$$
$$
\left(\frac{\mathrm{d}u_t^{(j)}}{\mathrm{d}t}\right)_i = u_i \quad i = m+1,\ldots,n, \tag{2.8}
$$

where

$$
v_i\left(u_t^{(j)}\right) = \left[-\widehat{C}_t^{u,G}\Gamma^{-1}\left(G\left(u_t^{(j)}\right) - y\right)\right]_i,
$$

which in the linear setting again simplifies to

$$v_i \left(u_t^{(j)} \right) = \left[-\widehat{C}(u_t) \nabla \Phi \left(u_t^{(j)} \right) \right]_i .$$

Roughly speaking, the dynamical system (2.8) evolves similarly to EKI (2.1), but forces the particles to stay within the box by deactivating directions pointing outside of the box at the boundary. The resulting system of ODEs occurs with two cruical problems, which are discontinuities of the RHS of (2.8) and the fact that straightforward preconditioning may lead to forcing directions which are no descent direction with respect to the potential Φ. The latter one is a well known problem for preconditioned gradient methods [6]. To overcome this issue, the authors propose to consider a smoothed coupled system of ODEs given by

$$\frac{\mathrm{d}u_t^{(j)}}{\mathrm{d}t} = \tau v \left(u_t^{(j)} \right) + \sum_{i=1}^{2m} \frac{1}{h_i \left(u_t^{(j)} \right)} \nabla h_i \left(u_t^{(j)} \right), \tag{2.9}$$

where $h_i(u) = a_i - u_i$ for $i = 1, \ldots, m$ and $h_{i+m}(u) = u_i - b_i$ for $i = 1, \ldots, m$. The purpose of introducing this form of smoothing is the resulting connection to the unconstrained optimisation problem

$$\min_{u \in \mathbb{R}^d} \tau \Phi(u) - \sum_{i=1}^{2m} \log(-h_i(u)) .$$

using a log-barrier penalty approach. Indeed, assuming that the forward model is linear and the constraints are described by boxes, one can prove that the solution of (2.8) solves (2.10) in the long-time limit [14, theorem 3.4]. In the following, we want to generalize this result to the Tikhonov regularised optimisation problem under more general convex inequality constraints.

2.2. Tikhonov regularised EKI under convex inequality constraints

Motivated by the continuous-time formulation of EKI under box constraints using the smoothed system (2.9) we are now ready to present our considered dynamical system for solving

$$\min_{u \in \Omega} \Phi^{\mathrm{reg}}(u) , \tag{2.10}$$

where $\Omega = \{u \in X : h_j(u) \leqslant 0, \ j = 1, \ldots, m\}$ describes our set of inequality constraints. We emphasize that the formulation of (2.10) can also be applied to standard EKI without Tikhonov regularisation by setting $\lambda = 0$. However, our theoretical justification crucially depends on $\lambda > 0$.

To be more precise, we make the following assumptions on the feasible set Ω for convex and continuously differentiable functions $h_j : X \to \mathbb{R}$.

Assumption 2.2. We assume that for each $j = 1, \ldots, m$ the function $h_j : X \to \mathbb{R}$ is convex and continuously differentiable. Moreover, we assume that the interior of Ω is non-empty, i.e. there exists $u \in X$ such that $h_j(u) < 0$ for each $j = 1, \ldots, m$.

Remark 2.3. Note that by assumption 2.2 it follows that for any $u \in X$ with $h_j(u) = 0$ we have that $\nabla h_j(u) \neq 0$. If $\nabla h_j(u) = 0$ for some $u \in X$ with $h_j(u) = 0$, by convexity of h_j it would follow that $\min_{u \in X} h_j(u) = 0$, which is in contradiction to assumption 2.2.

We firstly reformulate the constrained optimisation problem as unconstrained problem using the log-barrier penalty approach

$$\min_{u \in \mathbb{R}^d} \tau \Phi^{\mathrm{reg}}(u) - \sum_{i=1}^m \log(-h_i(u)),$$

where $\tau > 0$ is a penalty parameter. This optimisation problem can equivalently be rewritten through

$$\min_{u \in \mathbb{R}^d} \Phi^{\mathrm{reg}}(u) - \frac{1}{\tau} \sum_{i=1}^m \log(-h_i(u)), \tag{2.11}$$

We define $\Phi^b(u) = \Phi^{\mathrm{reg}}(u) - \frac{1}{\tau} \sum_{i=1}^m \log(-h_i(u))$. Observe that Φ^b is strongly convex in Ω due to Φ^{reg} being strongly convex and the assumption that h_i are convex. Hence, for any $\tau > 0$ there exists a unique global minimiser $u_*^\tau \in \Omega$ of Φ^b. Note that $\Phi^b(u)$ is not well-defined for $u \notin \Omega \setminus \partial\Omega$, and hence we are working with the convention that $\Phi^b(u) = +\infty$ for $u \notin \Omega \setminus \partial\Omega$. Moreover, using duality arguments one can even bound

$$\Phi^{\mathrm{reg}}(u_*^\tau) - \Phi^{\mathrm{reg}}(u_*) \leqslant \frac{m}{\tau}, \tag{2.12}$$

where u_* denotes the unique minimiser of the constrained problem (2.10), we refer to [10, section 11.2] for more details.

Remark 2.4. We can see in (2.12) that the computed solution from solving problem (2.11) is approaching the true KKT point u_* as we increase our penalty parameter τ. However, as τ increases the computation of the minimiser of (2.11) using an iterative optimisation method becomes more expensive. This is a well-known problem for penalty methods which is due to small steps of the applied scheme which are needed close to the boundary in order to ensure feasibility. Depending on the implemented optimisation scheme, one may need to check feasibility in each iteration and successively decrease the step size. In practical implementations one usually solves the optimisation problem for fixed but sequentially increasing values of the penalty parameter $\tau = \tau_k$. Here, one may use the solution of the preceding experiment as initial point for the next sequence.

We will further discuss how to apply this methodology in EKI in section 5.1. In regard to our algorithm we propose a method to use an adaptively increasing choice of τ which keeps the computational run-time low.

In the following, we set $\lambda = 1$ for simplicity. For solving (2.11) we incorporate the gradient descent direction of the log-barrier penalty into the dynamical system of Tikhonov regularised EKI leading to the dynamical system

$$\frac{\mathrm{d}u_t^{(j)}}{\mathrm{d}t} = \left[-\widehat{C}_t^{u,G} \Gamma^{-1} \left(G\left(u_t^{(j)}\right) - y\right) - \widehat{C}(u_t) C_0^{-1} u_t^{(j)} \right] + \frac{1}{\tau} \widehat{C}(u_t) \sum_{i=1}^m \frac{1}{h_i(\bar{u}_t)} \nabla h_i(\bar{u}_t), \tag{2.13}$$

for $\tau > 0$ and $j = 1, \ldots, J$. Let us emphasize that our considered dynamical system differs from (2.9) in more than just the additional regularisation term. Firstly, we are incorporating the log-barrier function with respect to $\bar{u}_t$ instead of for each particle $u_t^{(j)}$. This leads structural advantage when theoretically analysing the dynamical behavior of the ensemble mean. As result, the feasibility of the scheme with respect to the convex inequality constraints can only

be guaranteed for the ensemble mean. However, in practical scenarios it may be necessary to impose feasibility for each particle which could be guaranteed by the dynamical system

$$\frac{\mathrm{d}u_t^{(j)}}{\mathrm{d}t} = \left[-\widehat{C}_t^{u,G} \Gamma^{-1} \left(G\left(u_t^{(j)}\right) - y \right) - \widehat{C}\left(u_t\right) C_0^{-1} u_t^{(j)} \right] + \frac{1}{\tau} \widehat{C}\left(u_t\right) \sum_{i=1}^{m} \frac{1}{h_i\left(u_t^{(j)}\right)} \nabla h_i\left(u_t^{(j)}\right).$$

In what follows, we will focus on the dynamical system given by (2.13). As second difference to (2.9), we introduce a preconditioning of the gradient descent direction resulting from the log-barrier term by the empirical covariance. This preconditioning ensures the well-known subspace property of ensemble Kalman methods. This means, that the particle system solving (2.13) remains in the linear subspace spanned by the initial ensemble. We define

$$\mathcal{S} := u_0^{\perp} + \mathrm{span}\left\{ u_0^{(j)} - \bar{u}_0, j = 1, \ldots, J \right\},$$

where $u_0^{\perp} = \bar{u}_0 - \mathcal{P}_E \bar{u}_0$ with the projection matrix $\mathcal{P}_E = E(E^T E)^{-1} E^T$ onto $E \in \mathbb{R}^{d \times J}$ denotes matrix with columns consisting of the centered initial particles, i.e. $E = \left[u_0^{(1)} - \bar{u}_0, \ldots, u_0^{(J)} - \bar{u}_0 \right]$. Therefore, the constrained optimisation problem (2.10) changes to

$$\min_{u \in \Omega \cap \mathcal{S}} \Phi^{\mathrm{reg}}\left(u\right), \tag{2.14}$$

in case $J \leqslant d$.

Lemma 2.5. *Let $(u_0^{(j)})_{j=1,\ldots,J}$ be the initial ensemble. Then any solution $(u_t^{(j)}, t \geqslant 0)$ of (2.13) remains in the affine subspace $\mathcal{S}$, i.e. $u_t^{(j)} \in \mathcal{S}$ for all $t \geqslant 0$ and $j = 1, \ldots, J$.*

Proof. The proof of the first term has been shown in [15]. For the latter part we obtain

$$\frac{1}{\tau} \widehat{C}\left(u_t\right) \sum_{i=1}^{m} \frac{1}{h_i\left(u_t\right)} \nabla h_i\left(u_t\right) = \frac{1}{\tau} \frac{1}{J} \sum_{j=1}^{J} \sum_{i=1}^{m} \frac{1}{h_i\left(\bar{u}_t\right)} \langle u_t^{(j)} - \bar{u}_t, \nabla h_i\left(\bar{u}_t\right) \rangle \left(u_t^{(j)} - \bar{u}_t \right).$$

Hence, the latter term also remains in the space spanned by the centered particles and therefore the particles remain in the affine space $\mathcal{S}$. $\qquad \square$

Finally, we make the following assumption which is necessary to ensure feasibility and even the possibility for implementation. We need to ensure that the dynamical system is initialized with a feasible particle system.

Assumption 2.6. We assume that the mean of the initial ensemble $(u_0^{(j)})_{j=1,\ldots,J}$ lies in the interior of the feasible set Ω, i.e. we assume that $h_i(\bar{u}_0) < 0$ for all $i = 1, \ldots, m$. Furthermore, we assume that $\|u_0^{(j)}\| \leqslant c$ for all $j = 1, \ldots, J$ and some $c > 0$.

If the initial ensemble is not feasible, we may project the particles onto Ω. Hence, we assume without loss of generality that assumption 2.6 is satisfied.

2.3. Covariance inflation

One fundamental key step in analysing EKI algorithms is to quantify the ensemble collapse, which is the degeneration of the spread in the particle system. While the gradient approximation improves when the particles are close to each other, the preconditioner degenerates as well

and the scheme may get stuck when being far away from the optimal solution. The dynamics of the ensemble mean is given by

$$\frac{\mathrm{d}\bar{u}_t}{\mathrm{d}t} = \left[-\widehat{C}_t^{u,G}\Gamma^{-1}\left(\bar{G}(u_t) - y\right) - \widehat{C}(u_t)C_0^{-1}\bar{u}_t \right] + \frac{1}{\tau}\widehat{C}(u_t)\sum_{i=1}^{m}\frac{1}{h_i(\bar{u}_t)}\nabla h_i(\bar{u}_t).$$

For motivating EKI as derivative-free optimisation method, we split the dynamical system into a (preconditioned) gradient flow and an approximation error, written as

$$\frac{\mathrm{d}\bar{u}_t}{\mathrm{d}t} = \left[-\widehat{C}(u_t)\nabla\Phi^{\mathrm{reg}}(\bar{u}_t) \right] + \frac{1}{\tau}\widehat{C}(u_t)\sum_{i=1}^{m}\frac{1}{h_i(\bar{u}_t)}\nabla h_i(\bar{u}_t) + \mathrm{Err}\left(u_t^{(1)},\ldots,u_t^{(J)}\right)$$

$$= -\widehat{C}(u_t)\nabla\Phi^{b}(\bar{u}_t) + \mathrm{Err}\left(u_t^{(1)},\ldots,u_t^{(J)}\right).$$

For EKI without constraints one can indeed verify that the ensemble collapses but not too fast [44]. As result, the approximation error $\mathrm{Err}(u_t^{(1)},\ldots,u_t^{(J)})$ will degenerate in time, while the lower bound on $\widehat{C}(u_t)$ will ensure the convergence of the scheme.

In order to obtain more flexibility in controlling the speed of collapse from below and above, we are going to introduce covariance inflation. In particular, we are applying the covariance inflation introduced in [44, section 6] which inflates the particle system without changing the dynamical behavior of the ensemble mean. Indeed, it has been observed theoretically as well as numerically that this type of covariance inflation enhances the convergence speed.

The centered particles of (2.13) without covariance inflation satisfy

$$\frac{\mathrm{d}\left(u_t^{(j)} - \bar{u}_t\right)}{\mathrm{d}t} = -\widehat{C}_t^{u,G}\Gamma^{-1}\left(G\left(u_t^{(j)}\right) - \bar{G}(u_t)\right) - \widehat{C}(u_t)C_0^{-1}\left(u_t^{(j)} - \bar{u}_t\right),$$

where both terms force the particles to collapse. In order to reduce the contracting forces, we wish to relax these forces by changing the dynamical system of the ensemble spread to

$$\frac{\mathrm{d}\left(u_t^{(j)} - \bar{u}_t\right)}{\mathrm{d}t} = -(1-\rho_t)\widehat{C}_t^{u,G}\Gamma^{-1}\left(G\left(u_t^{(j)}\right) - \bar{G}(u_t)\right) - (1-\beta_t)\widehat{C}(u_t)C_0^{-1}\left(u_t^{(j)} - \bar{u}_t\right),$$

for $\rho_t,\beta_t \in [0,1)$ scaling the inflation strength. However, we aim to introduce this inflation without changing the dynamical behavior of the ensemble mean itself, which is used as optimisation scheme. One possible way of achieving the covariance inflation without changing the dynamical behavior of the ensemble mean is to consider the particle evolution of form

$$\frac{\mathrm{d}u_t^{(j)}}{\mathrm{d}t} = p_{\rho,\beta}\left(u_t^{(j)}\right) + \frac{1}{\tau}\widehat{C}(u_t)\sum_{i=1}^{m}\frac{1}{h_i(\bar{u}_t)}\nabla h_i(\bar{u}_t), \tag{2.15}$$

where

$$p_{\rho,\beta}\left(u_t^{(j)}\right) = -\widehat{C}_t^{u,G}\Gamma^{-1}\left(G\left(u_t^{(j)}\right) - y\right) - \widehat{C}(u_t)C_0^{-1}u_t^{(j)}$$
$$+ \rho_t\widehat{C}_t^{u,G}\Gamma^{-1}\left(G\left(u_t^{(j)}\right) - \bar{G}(u_t)\right) + \beta_t\widehat{C}(u_t)C_0^{-1}\left(u_t^{(j)} - \bar{u}_t\right), \tag{2.16}$$

with $0 \leqslant \rho_t \leqslant 1$ and $0 \leqslant \beta_t \leqslant 1$ for all $t \geqslant 0$. This incorporation of covariance inflation may be seen as artificial. However, it has a good intuition from an optimisation point of view. While in

the original EKI each particle is driven by its own direction, the formulation (2.15) additionally moves each particle into a joint direction. This interpretation can be seen more clearly when assuming $\rho_t = \beta_t$ and rewriting (2.16) by

$$p_{\rho,\beta}\left(u_t^{(j)}\right) = -(1-\rho_t)\left[\widehat{C}_t^{u,G}\Gamma^{-1}\left(G\left(u_t^{(j)}\right) - y\right) + \widehat{C}(u_t)C_0^{-1}u_t^{(j)}\right]$$
$$- \rho_t\left[\widehat{C}_t^{u,G}\Gamma^{-1}\left(\bar{G}(u_t) - y\right) + \widehat{C}(u_t)C_0^{-1}\bar{u}_t\right],$$

where the latter term is equal for all $j = 1, \ldots, J$. In our theoretical analysis, we will consider an inflation factor $\rho_t \to 1$ as $t \to \infty$ and turn of the inflation for the regularisation term through setting $\beta_t = 0$. Note that the presented results straightforwardly extend to $\beta_t > 0$. As stated above the evolution of the ensemble mean remains

$$\frac{d\bar{u}_t}{dt} = -\widehat{C}_t^{u,G}\Gamma^{-1}\left(\bar{G}(u_t) - y\right) - \widehat{C}(u_t)C_0^{-1}\bar{u}_t + \frac{1}{\tau}\widehat{C}(u_t)\sum_{i=1}^{m}\frac{1}{h_i(\bar{u}_t)}\nabla h_i(\bar{u}_t)$$

$$= v(u_t) + \frac{1}{\tau}\widehat{C}(u_t)\sum_{i=1}^{m}\frac{1}{h_i(\bar{u}_t)}\nabla h_i(\bar{u}_t),$$

where we defined

$$v(u_t) := -\widehat{C}_t^{u,G}\Gamma^{-1}\left(\bar{G}(u_t) - y\right) - \widehat{C}(u_t)C_0^{-1}\bar{u}_t. \tag{2.17}$$

However, the evolution of the ensemble spread changes to

$$\frac{d\left(u_t^{(j)} - \bar{u}_t\right)}{dt} = -(1-\rho_t)\widehat{C}_t^{u,G}\Gamma^{-1}\left(G\left(u_t^{(j)}\right) - \bar{G}(u_t)\right) - \widehat{C}(u_t)C_0^{-1}\left(u_t^{(j)} - \bar{u}_t\right).$$

We emphasize that the solutions of the inflated flow obviously still satisfy the subspace property (see [44]). Note again, the dynamics of the centered particles with and without covariance inflation are independent of τ, which will later lead to a speed of collapse independent of τ. We summarize the incorporation of variance inflation in the following assumption.

Assumption 2.7. We consider the covariance inflated dynamics given by (2.16) where $0 \leqslant \rho_t < 1$ with $\rho_t \to 1$ as $t \to \infty$ and set $\beta_t = 0$ for all $t > 0$.

Remark 2.8. This form of covariance inflation has been proposed in [44] to reduce the speed of ensemble collapse, we are utilizing this approach to remain the most suitable rate of collapse of order $\frac{1}{t}$.

Furthermore, we highlight this effect in the numerical experiment in section 5.3, where we compare the results using no covariance inflation, a fixed (but tuned) covariance inflation and our proposed adaptive covariance inflation. We observe that adaptively increasing the covariance inflation from 0 to 1 yields the most stable and efficient performance. As a result, we observe significant acceleration of convergence for increasing covariance inflation and for the adaptive approach a reduction of the computational time. We note that many other theoretical and practical papers apply alternative forms of covariance inflation to encounter the ensemble collapse, e.g. [16, 32].

3. Ensemble collapse and well-posedness

We define the centered particles as $e_t^{(j)} = u_t^{(j)} - \bar{u}_t$. Next, we define the following Lyapunov function $V_e(t) = \frac{1}{J}\sum_{j=1}^{J}\frac{1}{2}\|e_t^{(j)}\|^2$ describing the deviation of the ensemble of particles from

its mean and we will analyse its behaviour. To be more precise, we will prove that the ensemble collapse with rate $V_e(t) \in \mathcal{O}(1/t)$.

Lemma 3.1 (upper bound). *Let assumptions* 2.6 *and* 2.7 *hold. Furthermore, let* $(u_0^{(j)})_{j=1,\ldots,J} \in \Omega$ *be the initial particle system and let* $(u_t^{(j)}, t \geqslant 0)$, $j = 1,\ldots,J$, *be a solution of* (2.15). *Then for all* $t \geqslant 0$ *it holds true that*

$$V_e(t) \leqslant \frac{1}{\frac{2\sigma_{\min}}{J} t + V_e(0)^{-1}} \,.$$

Proof. The outline of the proof follows similarly to [44, lemma 4.3]. The time evolution of V_e is given by

$$\begin{aligned}
\frac{dV_e(t)}{dt} =& \frac{2}{J} \sum_{j=1}^{J} \langle u_t^{(j)} - \bar{u}_t, \frac{d\left(u_t^{(j)} - \bar{u}_t\right)}{dt} \rangle \\
=& - (1 - \rho_t) \frac{2}{J^2} \sum_{k,j=1}^{J} \langle u_t^{(j)} - \bar{u}_t, \left(u_t^{(k)} - \bar{u}_t\right) \left(G\left(u_t^{(k)}\right) - \bar{G}(u_t)\right)^T \Gamma^{-1} \left(G\left(u_t^{(j)}\right) - \bar{G}(u_t)\right) \rangle \\
& - \frac{2}{J^2} \sum_{k,j=1}^{J} \langle u_t^{(j)} - \bar{u}_t, \left(u_t^{(k)} - \bar{u}_t\right) \left(u_t^{(k)} - \bar{u}_t\right)^T C_0^{-1} \left(u_t^{(j)} - \bar{u}_t\right) \rangle \,.
\end{aligned}$$

We observe that

$$\frac{1}{J^2} \sum_{k,j=1}^{J} \langle u_t^{(j)} - \bar{u}_t, \left(u_t^{(k)} - \bar{u}_t\right) \left(G\left(u_t^{(k)}\right) - \bar{G}(u_t)\right)^T \Gamma^{-1} \left(G\left(u_t^{(j)}\right) - \bar{G}(u_t)\right) \rangle = \|\widehat{C}_t^{u,G} \Gamma^{-1/2}\|_F^2$$

and similarly

$$\frac{2}{J^2} \sum_{k,j=1}^{J} \langle u_t^{(j)} - \bar{u}_t, u_t^{(k)} - \bar{u}_t \rangle \langle u_t^{(k)} - \bar{u}_t C_0^{-1} \left(u_t^{(j)} - \bar{u}_t\right) \rangle = \|\widehat{C}_t(u_t) C_0^{-1/2}\|_F^2$$

such that

$$\begin{aligned}
\frac{dV_e(t)}{dt} =& - 2(1 - \rho_t) \|\widehat{C}_t^{u,G} \Gamma^{-1/2}\|_F^2 - 2\|\widehat{C}_t(u_t) C_0^{-1/2}\|_F^2 \\
\leqslant& - \frac{2\sigma_{\min}}{J} \left(\frac{1}{J} \sum_{j=1}^{J} \|u_t^{(j)} - \bar{u}_t\|^2\right)^2 = - \frac{2\sigma_{\min}}{J} V_e(t)^2,
\end{aligned}$$

where $\sigma_{\min}$ denotes the smallest eigenvalue of C_0^{-1}. Then the claim follows by Lyapunov theory. □

As discussed above, in order to prove convergence of the scheme as optimisation method we will need to lower bound the covariance matrix which is used as preconditioner. We will provide the following lower bound on the smallest eigenvalue of $(\widehat{C}(u_t))_{t \geqslant 0}$ which prevents the ensemble to collapse too fast.

Lemma 3.2 (lower bound). *Let assumptions* 2.6 *and* 2.7 *hold. Moreover, let* $(u_0^{(j)})_{j=1,\ldots,J} \in \Omega$ *be the initial particle system and let* $(u_t^{(j)}, t \geqslant 0), j = 1, \ldots, J,$ *be a solution of* (2.15). *Moreover, we assume that*

$$\eta_0 := \min_{z \in \mathcal{S}, \|z\|=1} \langle z, C(u_0), z \rangle > 0.$$

Then, for each $z \in \mathcal{S}$ *with* $\|z\| = 1$ *there exists a* $t_\circ > 0$ *such that.*

$$\langle z, \widehat{C}(u_t) z \rangle \geqslant \frac{w}{(c + w \max(k_1, 0))(t + b)}$$

for all $t \geqslant t_\circ$*. The constants are* $a = \frac{(1-\rho_{t_\circ})c_{\mathrm{lip}}^2 \lambda_{\max} J}{\sigma_{\min}}, b = \frac{V_e(0)^{-1} J}{2\sigma_{\min}}, c = 2\sigma_{\max}, k_1 = \frac{(1-a)-\eta_{t_\circ} cb}{(1-a)b^a \eta_{t_\circ}}$ *and* $w = 1 - a > 0$*. Here* $\sigma_{\min}$ *and* $\sigma_{\max}$ *denote the smallest and largest eigenvalue of* C_0^{-1} *and* $\lambda_{\max}$ *denotes the largest eigenvalue of* Γ*.*

Proof. The outline of the proof follows similarly to [44, lemma 4.4], where we include an adaptation for allowing adaptive covariance inflation $(\rho_t, t \geqslant 0)$. Consider the dynamics of the empirical covariance matrix

$$\frac{\mathrm{d}\widehat{C}(u_t)}{\mathrm{d}t} = \frac{1}{J} \sum_{j=1}^{J} \frac{\mathrm{d}e_t^{(j)}}{\mathrm{d}t} \left(e_t^{(j)}\right)^T + \frac{1}{J} \sum_{j=1}^{J} e_t^{(j)} \left(\frac{\mathrm{d}e_t^{(j)}}{\mathrm{d}t}\right)^T$$

$$= -2(1-\rho_t)\widehat{C}_t^{u,G}\Gamma^{-1}\widehat{C}_t^{G,u} - 2\widehat{C}(u_t)C_0^{-1}\widehat{C}(u_t).$$

Next let $z \in \mathcal{S}$ with $\|z\| = 1$, then we note that

$$\langle z, \frac{\mathrm{d}\widehat{C}(u_t)}{\mathrm{d}t} z \rangle = -2(1-\rho_t)\|\widehat{C}_t^{G,u}z\|_\Gamma^2 - 2\|\widehat{C}(u_t)z\|_{C_0}. \tag{3.1}$$

Observe that

$$\|\widehat{C}_t^{G,u}z\|_\Gamma^2 \leqslant \|\widehat{C}_t^{G,u}z\|^2\|\Gamma^{-1}\| \leqslant \|\widehat{C}_t^{G,u}z\|^2 \lambda_{\max},$$

where $\lambda_{\max}$ denotes the largest eigenvalue of Γ^{-1}. We apply Cauchy-Schwarz and Hölder's inequality to derive the following bound

$$\|\widehat{C}^{G,u}z\|^2 = \langle \widehat{C}^{G,u}z, \widehat{C}^{G,u}z \rangle = \langle \frac{1}{J}\sum_{j=1}^{J}\langle u^{(j)} - \bar{u}, z \rangle \left(G\left(u^{(j)}\right) - \bar{G}\right), \frac{1}{J}\sum_{i=1}^{J}\langle u^{(i)} - \bar{u}, z \rangle \left(G\left(u^{(i)}\right) - \bar{G}\right) \rangle$$

$$= \frac{1}{J^2}\sum_{i,j=1}^{J}\langle u^{(j)} - \bar{u}, z \rangle\langle u^{(i)} - \bar{u}, z \rangle\langle G\left(u^{(j)}\right) - \bar{G}, G\left(u^{(i)}\right) - \bar{G} \rangle$$

$$\leqslant \frac{1}{J^2}\sum_{i,j=1}^{J}\langle u^{(j)} - \bar{u}, z \rangle\langle u^{(i)} - \bar{u}, z \rangle\|G\left(u^{(j)}\right) - \bar{G}\|\|G\left(u^{(i)}\right) - \bar{G}\|$$

$$= \left(\frac{1}{J}\sum_{j=1}^{J}\langle u^{(j)} - \bar{u}, z \rangle\|G\left(u^{(j)}\right) - \bar{G}\|\right)^2$$

$$\leqslant \left(\left(\frac{1}{J}\sum_{j=1}^{J}\langle u^{(j)} - \bar{u}, z \rangle^2\right)^{\frac{1}{2}}\left(\frac{1}{J}\sum_{j=1}^{J}\|G(u^{(j)}) - \bar{G}\|^2\right)^{\frac{1}{2}}\right)^2,$$

which implies that

$$\|\widehat{C}_t^{G,u}z\|_\Gamma^2 \leqslant c_{\text{lip}}^2 \lambda_{\max} V_e(t) \langle z, C(u)z\rangle,$$

since

$$\langle u^{(j)} - \bar{u}, z\rangle^2 = z^T \left(u^{(j)} - \bar{u}\right)\left(u^{(j)} - \bar{u}\right)^T z.$$

We denote by η_t the smallest eigenvalue of $\widehat{C}(u_t)$ and by $\phi(t)$ the corresponding unit-norm eigenvector. We will show that $\eta_t > 0$ for all $t \geqslant 0$. The following holds

$$0 = \frac{\mathrm{d}\|\phi(t)\|^2}{\mathrm{d}t} = 2\langle \phi(t), \frac{\mathrm{d}\phi(t)}{\mathrm{d}t}\rangle,$$

and hence,

$$
\begin{aligned}
\frac{\mathrm{d}\eta_t}{\mathrm{d}t} &= \frac{\mathrm{d}\langle \phi(t), \widehat{C}(u_t)\phi(t)\rangle}{\mathrm{d}t} \\
&= \langle \phi(t), \frac{\mathrm{d}\widehat{C}(u_t)\phi(t)}{\mathrm{d}t}\rangle + \langle \frac{\mathrm{d}\phi(t)}{\mathrm{d}t}, \widehat{C}(u_t)\phi(t)\rangle \\
&= \langle \phi(t), \frac{\mathrm{d}\widehat{C}(u_t)}{\mathrm{d}t}\phi(t)\rangle + 2\eta_t\langle \frac{\mathrm{d}\phi(t)}{\mathrm{d}t}, \phi(t)\rangle \\
&\geqslant -2(1-\rho_t)c_{\text{lip}}^2 V_e(t)\lambda_{\max}\langle \phi(t), \widehat{C}(u_t)\phi(t)\rangle - 2\sigma_{\max}\|\widehat{C}(u_t)\phi(t)\|^2, \\
&= -2(1-\rho_t)c_{\text{lip}}^2 V_e(t)\lambda_{\max}\eta_t - 2\sigma_{\max}\eta_t^2 \\
&\geqslant -\frac{2(1-\rho_t)c_{\text{lip}}^2\lambda_{\max}}{\frac{2\sigma_{\min}}{J}t + V_e(0)^{-1}}\eta_t - 2\sigma_{\max}\eta_t^2,
\end{aligned}
$$

where we used in the third step the symmetry of $\widehat{C}(u_t)$ and in the last lemma 3.1. Hence we have an ODE of the form

$$\frac{\mathrm{d}\eta_t}{\mathrm{d}t} \geqslant -\frac{a_t}{t+b}\eta_t - c\eta_t^2,$$

where $a_t = \frac{(1-\rho_t)c_{\text{lip}}^2\lambda_{\max}J}{\sigma_{\min}}, b = \frac{V_e(0)^{-1}J}{2\sigma_{\min}}, c = 2\sigma_{\max}$. Since $\rho_t \to 1$ as $t \to \infty$, there exists a $t_\circ$ such that $a_t \leqslant a \in (0,1)$ for all $t \geqslant t_\circ$. The solution of the ODE

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = -\frac{a}{t+b}x(t) - cx(t)^2$$

is given by

$$x(t) = \frac{1-a}{c(t+b) + (1-a)k_1(t+b)^a}, \quad x(0) = \eta_{t_\circ} > 0,$$

where k_1 is chosen such that the initial condition is satisfied (see lemma A.3). Since $a \in (0,1)$ we have that $x(t) > 0$ for all $t \geqslant 0$. Hence, we can run our ODE in time and choose $t_\circ$ as our initial starting point. For such values of a we have that $(t+b)^a \leqslant (t+b)$ such that we obtain the lower bound

$$x(t) \geqslant \frac{1-a}{c(b+t) + (1-a)\max(k_1,0)(t+b)} = \frac{w}{(c+w\max(k_1,0))(t+b)},$$

where $w = 1 - a > 0$, which guarantees by comparison argument that

$$\eta_t \geqslant \frac{w}{(c + w \max(k_1, 0))(t + b)}$$

for all $t \geqslant t_\circ$. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Remark 3.3. We emphasize that the constants describing the upper and lower bound of the ensemble collapse as well as $t_\circ$ from lemma 3.2 are independent of the penalty parameter τ.

Our next result shows that any solution $u(t)$ of (2.15), remains in the feasible set Ω for all $t \geqslant 0$ and therefore, there exists a unique global solution. We provide the proof of the statement in appendix A.2.

Proposition 3.4 (existence). *Let assumptions* 1.1, 2.1, 2.2, 2.6 *and* 2.7 *hold and let* $(u_0^{(j)})_{j=1,\ldots,J}$ *be the initial particle system. We assume that the feasible set is bounded, i.e. that there exists $R > 0$ such that $\Omega \subseteq \mathcal{B}_R(0)$. Moreover, we assume that $\|\nabla h_i(u)\| \leqslant C(R)$ for some constant $C(R) > 0$ depending on R and all $u \in \mathcal{B}_R(0)$. Then the mean of any solution $(u_t^{(j)}, t \geqslant 0)$ of (2.15) remains in the feasible set Ω, i.e. $\bar{u}_t \in \Omega$ for all $t \geqslant 0$. Moreover, there exists a unique solution of (2.15).*

Due to proposition 3.4 under assumption 2.6 the mean of computed solutions is always feasible and bounded. Hence, we also obtain that $\|u_t^{(j)}\| \leqslant B$ for some $B > 0$, all $t \geqslant 0$ and $j = 1, \ldots, J$.

4. Gradient flow structure and convergence analysis

The dynamics of the particles given by the RHS of (2.2) can be approximated by the preconditioned gradient flow

$$-\widehat{C}(u_t) \nabla \Phi^{\text{reg}}\left(u^{(j)}\right).$$

This approximation is getting more precise, as the ensemble collapse is happening [44, lemma 4.5, lemma 6.4]. We consider in the following the approximation of (2.13) for the mean of our particles given in [44, lemma 6.4].

Lemma 4.1 (gradient flow approximation). *Consider the ensemble $\{u_t^{(j)}, j = 1, \ldots, J\}$ and assume that assumption* 2.1 *holds, then the mean of the particles satisfies*

$$\|\widehat{C}_t^{u,G} \Gamma^{-1}(\bar{G}_t - y) - \widehat{C}(u_t) \nabla \Phi(\bar{u})\| \leqslant b_1 J \sqrt{\Phi(\bar{u})} V_e(t)^{\frac{3}{2}}, \qquad (4.1)$$

where b_1 is independent of J.

We are now ready to formulate our main result of convergence.

Theorem 4.2 (main convergence). *Suppose that assumptions* 1.1, 2.1, 2.2, 2.6 *and* 2.7 *are satisfied. Let u_τ^* be the unique global minimiser of (2.11), let $(u_0^{(j)})_{j=1,\ldots,J} \in \Omega$ be the initial particle system with $J > d$ and let $(u_t^{(j)}, t \geqslant 0)$, $j = 1, \ldots, J$, be a solution of (2.15). Then there exist $c_1, c_2 > 0$ and $\bar{t} \geqslant 0$ such that*

$$\Phi^b(\bar{u}_t) - \Phi^b(u_*^\tau) \leqslant \left(\frac{c_1}{t + c_2}\right)^{\frac{1}{\gamma}},$$

for all $t \geqslant \bar{t}$, where $\gamma := c + w \max(k_1, 0) > 0$ is independent of t and τ. The constants c, w and k_1 are defined in lemma 3.2. *In particular, we have that*

$$\lim_{t \to \infty} \Phi^b(\bar{u}_t) - \Phi^b(u_*^\tau) = 0.$$

Proof. Define $V(t) = \Phi^b(\bar{u}_t) - \Phi^b(u_*^\tau)$ and $Err(t) = \widehat{C}_t^{u,G}\Gamma^{-1}(\bar{G}_t - y) - \widehat{C}(u_t)\nabla\Phi(\bar{u})$, then we have

$$
\begin{aligned}
\frac{dV(t)}{dt} &= \langle \nabla\Phi^b(\bar{u}_t), \frac{d\bar{u}_t}{dt} \rangle \\
&= \langle \nabla\Phi^b(\bar{u}_t), \left[-\widehat{C}_t^{u,G}\Gamma^{-1}(\bar{G}(u_t) - y) - \widehat{C}(u_t)C_0^{-1}\bar{u}_t \right] + \frac{1}{\tau}\widehat{C}(u_t)\sum_{i=1}^m \frac{1}{h_i(\bar{u}_t)}\nabla h_i(\bar{u}_t) \rangle \\
&= \langle \nabla\Phi^b(\bar{u}_t), -Err(t) \rangle \\
&\quad + \langle \nabla\Phi^b(\bar{u}_t), \widehat{C}(u_t)\left(-\nabla\Phi(\bar{u}_t) - C_0^{-1}\bar{u}_t + \frac{1}{\tau}\sum_{i=1}^m \frac{1}{h_i(\bar{u}_t)}\nabla h_i(\bar{u}_t) \right) \rangle \\
&= \langle \nabla\Phi^b(\bar{u}_t), -Err(t) \rangle + \langle \nabla\Phi^b(\bar{u}_t), -\widehat{C}(u_t)\nabla\Phi^b(\bar{u}_t) \rangle.
\end{aligned}
$$

The first term can be approximated by Cauchy–Schwarz and using (4.1)

$$
\begin{aligned}
\langle \nabla\Phi^b(\bar{u}_t), -Err(t) \rangle &\leqslant \|\nabla\Phi^b(\bar{u}_t)\| \|Err(t)\| \leqslant \|\nabla\Phi^b(\bar{u}_t)\| b_1 J \sqrt{\Phi(\bar{u}_t)} V_e(t)^{\frac{3}{2}} \\
&\leqslant \|\nabla\Phi^b(\bar{u}_t)\| b_1 J \sqrt{c_{\text{lip}}\|\bar{u}_t\|} V_e(t)^{\frac{3}{2}}.
\end{aligned}
$$

For the second term we obtain through lemma 3.1 and the PL-inequality, for which we give more details in lemma A.1,

$$-\langle \nabla\Phi^b(\bar{u}_t), \widehat{C}(u_t)\nabla\Phi^b(\bar{u}_t) \rangle \leqslant -\eta_t \|\nabla\Phi^b(\bar{u}_t)\|^2 \leqslant -2\mu\eta_t\left(\Phi^b(\bar{u}_t) - \Phi^b(u_*^\tau) \right),$$

where $\eta_t > 0$ denotes the smallest eigenvalue of $\widehat{C}(u_t)$. Using Young's inequality it follows that $\|\nabla\Phi^b(\bar{u}_t)\| \leqslant \frac{1}{2}\|\nabla\Phi^b(\bar{u}_t)\|^2 + \frac{1}{2}$, such that we obtain

$$
\begin{aligned}
\frac{dV(t)}{dt} &\leqslant -\eta_t\|\nabla\Phi^b(\bar{u}_t)\|^2 + \|\nabla\Phi^b(\bar{u}_t)\| b_1 J\sqrt{c_{\text{lip}}\|\bar{u}_t\|} V_e(t)^{\frac{3}{2}} \\
&\leqslant -\frac{w}{(c + w\max(k_1, 0))(t + b)}\|\nabla\Phi^b(\bar{u}_t)\|^2 \\
&\quad + \left(\frac{1}{2}\|\nabla\Phi^b(\bar{u}_t)\|^2 + \frac{1}{2} \right) b_1 J\sqrt{c_{\text{lip}}} B\left(\frac{1}{\frac{2\sigma_{\min}}{J}t + V_e(0)^{-1}} \right)^{\frac{3}{2}}
\end{aligned}
$$

for all $t \geqslant t_\circ$, where we have used lemma 3.2. For sufficiently large $\bar{t} \geqslant 0$ we have that

$$\frac{1}{2}\|\nabla\Phi^b(\bar{u}_t)\|^2\left(b_1 J\sqrt{c_{lip}} B\left(\frac{1}{\frac{2\sigma_{\min}}{J}t + V_e(0)^{-1}} \right)^{\frac{3}{2}} - \frac{w}{2(c + w\max(k_1, 0))(t + b)} \right) \leqslant 0,$$

for all $t \geqslant \bar{t}$, such that

$$
\frac{\mathrm{d}V(t)}{\mathrm{d}t} \leqslant -\frac{w}{2\left(c + w\max\left(k_1, 0\right)\right)\left(t + b\right)} \|\nabla\Phi^b\left(\bar{u}_t\right)\|^2 + \frac{1}{2}b_1 J\sqrt{c_{\mathrm{lip}}B}\left(\frac{1}{\frac{2\sigma_{\min}}{J}t + V_e\left(0\right)^{-1}}\right)^{\frac{3}{2}}
$$

$$
\leqslant -\frac{w\mu}{\left(c + w\max\left(k_1, 0\right)\right)\left(t + b\right)}V(t) + \frac{1}{2}b_1 J\sqrt{c_{\mathrm{lip}}B}\left(\frac{1}{\frac{2\sigma_{\min}}{J}t + V_e\left(0\right)^{-1}}\right)^{\frac{3}{2}}
$$

for all $t \geqslant \max\{\bar{t}, t_\circ\}$. Using Gronwall's lemma we obtain that

$$
V(t) = \Phi^b\left(\bar{u}_t\right) - \Phi^b\left(u_*^\tau\right) \leqslant \left(\frac{c_1}{t + c_2}\right)^{\frac{1}{\gamma}},
$$

for all $t \geqslant \max\{\bar{t}, t_\circ\}$, where $\gamma := c + w\max(k_1, 0)$ is independent of t and τ. □

Remark 4.3. The assumption $J > d$ in theorem 4.2 is rather constraining, and usually not fulfilled in practical situations. However, in the case $J \leqslant d$ we have observed that the well-known subspace property lemma 2.5 is satisfied. In that case we can consider convergence to a KKT-point u^* of (2.14) and the results in theorem 4.2 can be generalized straightforwardly to this setting. We refer to section 4.3 in [44] for more details.

5. Numerical experiments

We now conduct two numerical experiments to test our methodology. In the first experiment, we aim to estimate the source term in a 1D parabolic PDE, using measurements obtained from its solution. However, this experiment is a linear inverse problem, therefore we incorporate a non-linear perturbation term into the potential (1.2). Subsequently, in the second experiment, our goal is to estimate the log-diffusion coefficient in a 2D elliptic PDE once again using measurements of its solution. In the implementation we solve equation (2.15) using MATLABs ODE solver ode45.

5.1. Implementation of the penalty approach

We conclude the latter experiment with an implementation of an adaptive penalty parameter τ. Let u_* be the KKT-point of (2.10) and $u_*^\tau \in \mathbb{R}^d$ the unique global minimiser of Φ^b. Then the error (2.12) decreases as the penalty parameter τ increases.

We will compute the solutions of (2.15) for several fixed penalty parameters τ and compute the errors to u_* in the parameter space as well as the observations space. In proposition 3.4 we have verified that each solution of (2.15) remains feasible for fixed τ. However, in numerical implementations this cannot be guaranteed anymore. Depending on the chosen numerical solver one may need to validate feasibility in each discretisation step and successively decrease the applied step size. Indeed, we observe that the ODE (2.15) becomes more and more stiff with increasing τ. Hence, our applied ODE solver has to take smaller step sizes in order to ensure feasibility and the algorithm becomes computationally slow. Taking this into account, we also consider an adaptive choice of the penalty parameter τ, which is increasing over time, when solving the ODE (2.15) in section 5.3.1. To be more precise, we consider the dynamical system

$$\frac{\mathrm{d}u_t^{(j)}}{\mathrm{d}t} = p_{\rho_t,0}\left(u_t^{(j)}\right) + \frac{1}{\tau_t}\widehat{C}(u_t)\sum_{i=1}^{m}\frac{1}{h_i(\bar{u}_t)}\nabla h_i(\bar{u}_t), \tag{5.1}$$

with $(\tau_t)_{t\geqslant 0}$ denoting our adaptive penalty parameter choice with $\lim_{t\to\infty}\tau_t = \infty$. We will observe in section 5.3.1 that this approach leads to significant reduction of the overall runtime.

5.2. Pseudolinear example

In the following section, we present a nonlinear example with Tikhonov regularisation to obtain a strongly convex potential. The example is based on a linear forward model where we add a nonlinear perturbation to the model leading to a nonlinear inverse problem. The magnitude of the nonlinear term is controlled by some $\varepsilon > 0$. Then the regularisation parameter can be chosen depending on ε, i.e. $\lambda(\varepsilon)$, such that we can keep the regularisation at a low level and still obtain a strongly-convex functional by controlling the magnitude of the nonlinear term.

We consider an inverse problem of the form (1.2) where $G(u) = Au + \varepsilon[\sin(u_1),\dots,\sin(u_K)]^T$, where $A \in \mathbb{R}^{K\times K}$ is symmetric and positive semi-definite and $\varepsilon > 0$. We assume that $\sup_{u\in\Omega}\|u\| \leqslant B$ and let $A_{\max} = \max_{i,j}|A_{i,j}|$, $y_{\max} = \max_i|y_i|$. For $\lambda > ((4+B)a_{\max} + y_{\max})\varepsilon + \varepsilon^2$, the Tikhonov regularised loss function

$$\Phi^{\mathrm{reg}}(u) = \frac{1}{2}\|G(u) - y\|^2 + \frac{\lambda}{2}\|u\|^2$$

is strongly convex and L-smooth, see lemma A.2.

For the linear model A we consider a one-dimensional heat equation given by the differential equation

$$\begin{aligned}
\frac{\partial u(x,t)}{\partial t} - \frac{\partial^2 u(t,x)}{\partial x^2} &= f(x) && (t > 0, x \in (0,1))\\
u(0,x) &= 0 && (x \in (0,1))\\
u(t,0), u(t,1) &= 0 && (t \geqslant 0).
\end{aligned}$$

We aim to recover the forcing f given observations of the solution at specific grid points. For this we introduce the solution operator of the PDE $L = \frac{\mathrm{d}}{\mathrm{d}t} - \frac{\mathrm{d}^2}{\mathrm{d}^2 x}$ and an equidistant observation operator on $[0,1]\times\mathbb{R}_{\geqslant 0}$ defined as $\mathcal{O}: H_0^1([0,1],\mathbb{R}) \to \mathbb{R}^K, (p(\cdot)) \mapsto \mathcal{O}(p(\cdot)) = (p(x_1),\dots,p(x_K))^T$. Here $p \in H_0^1([0,1],\mathbb{R})$ denotes a solution operator of the PDE. Then the forward operator is given by $A = \mathcal{O}\circ L^{-1}$, and the inverse problem is respectively

$$u = Af + \eta,$$

where $\eta \sim \mathcal{N}(0,\Gamma)$, with $\Gamma = 0.1^2\mathrm{Id}_K$.

The forcing is assumed to follow a Gaussian distribution with zero mean and covariance given by $C_0 := C(s,t) = \sigma^2\exp\left(-\frac{|s-t|^2}{L_{sc}}\right)$, where $(s,t) \in [0,1]\times[0,1]$. In this expression, σ^2 is set to 10 and L_{sc} is set to 0.1. To simulate the random field, we employ a Karhunen-Loève (KL) expansion truncated after 12 terms. Specifically, we express $f(x,\omega)$ as follows:

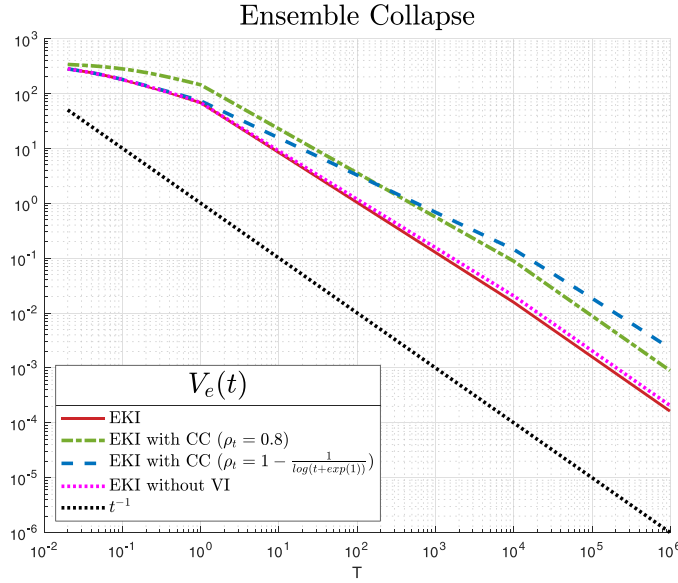$$f(x,\omega) = \sum_{i=1}^{8}\lambda_i^{1/2}e_i(x)\xi_i(\omega).$$

Figure 1. Ensemble collapse in the form of $V_e(t)$ of the particles for the EKI (red) and EKI with CC (green line represents fixed covariance inflation; blue line increasing covariance inflation). The black dotted line depicts the rate t^{-1}.

Here, λ_i represents the largest eigenvalues of $C(s,t)$, $e_i(x)$ denotes the corresponding eigenfunctions, and ξ_i are standard normal distributed random variables. Furthermore we draw our initial ensemble from the same random field by making $J = 10$ independent draws from $f(x,\omega)$. Note that the generated solution of the EKI will lie in $f_0^\perp + \mathcal{E}$, where $\mathcal{E}$ is the linear span of the centered initial ensemble and $f_0^\perp = \bar{f}(0) - P_{\mathcal{E}}\bar{f}(0)$. For the simulation, we use a spatial step size of $h = 0.01$, hence we have 99 interior points and therefore the state space dimension is $d = 99$, a time step size of $\Delta = 0.05$, and only solved for one step size into the future.

We draw the initial forcing $f^\dagger$ from $f(x,\omega)$ and compute our observations by solving the PDE given this forcing.

We impose convex constraints (CC) in the form of an upper bound on the norm of our solution, i.e. we set $h(u) = \frac{1}{2}\|u\|_{C_0}^2 - \frac{1}{2}\|f^\dagger\|_{C_0}$ and require $h(u_t^{(j)}) \leqslant 0$ for all $j = 1,\dots,J$ and $t \geqslant 0$.

We compare our solution to the optimum of (2.11), where we use, MATLABs optimisation tool `fmincon`. Additionally to the CC, we also impose the constraint of the subspace property, i.e. we search a solution in $f_0^\perp + \mathcal{E}$. For this to be satisfied, the computed solution u has to satisfy the equation $P_{\mathcal{E}}u = f_0^\perp$, where $P_{\mathcal{E}}$ is the projection onto $\mathcal{E}$. We solve the ODE (2.15) until $T = 10^6$ and use `ode45` as solver. We use $\tau = 10^4$ and a regularisation factor of $\beta = 0.01$.

Furthermore, we consider three different types of covariance inflation

- Increasing covariance inflation $\rho_t = 1 - \frac{1}{\log(t+\exp(1))}$.
- Constant covariance inflation $\rho_t = 0.8$ for all $t \geqslant 0$.
- No variance inflation, i.e. $\rho_t = 0$ for all $t \geqslant 0$.

Figure 1 depicts the ensemble collapse of the particles in the form $V_e(t) = \frac{1}{J}\sum_{j=1}^J \frac{1}{2}\|e_t^{(j)}\|^2$. We can see that the ensemble collapse for all methods with convex constraints happen with the
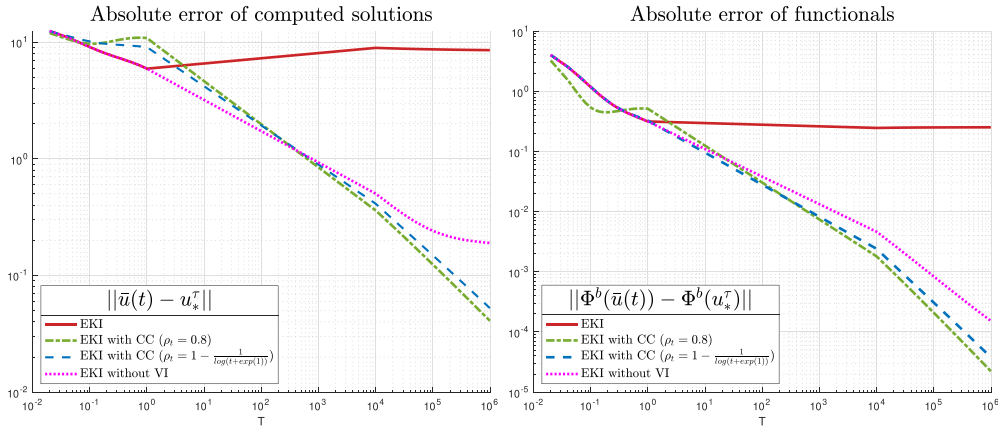
Figure 2. Absolute error of computed solutions in the parameter (left) and observation space (right). The red line illustrates the EKI, the green dash-dotted line EKI with CC and constant covariance inflation of $\rho_t = 0.8$, the blue dashed line considered an increasing covariance inflation of $\rho_t = 1 - \frac{1}{\log(t+\exp(1))}$ and the pink line represents no level of variance inflation, i.e. $\rho_t = 0$.

same rate t^{-1} as imposed by lemma 3.1. All three methods perform similarly well and only differ from the EKI due to constants.

In figure 2 we illustrate the error in the parameter space and observation space. The solutions computed by all methods with convex constraints converge to the KKT point u_*^τ, whereas the EKI without constraints does not converge, since it does not take the constraints into account. Furthermore, we note that the methods with a positive amount of covariance inflation perform similarly well. In regard to the performance of the algorithms this is an important observation since the ODE (2.15) get stiffer with higher value of covariance inflation and therefore also computationally slower. Since the experiment with $\rho_t = 0.8$ is performing similarly well as with $\rho_t = 1 - \frac{1}{\log(t+\exp(1)}$ one should preferably implement a fixed value of covariance inflation. Moreover, we note that the experiment without variance inflation, converges up until time $T = 1e4$ similarly fast as the experiments with variance inflation, but slows down afterwards.

In the left image of figure 3 we illustrate the computed solutions. We only illustrate the solution computed with $\rho = 0.8$, the dashed-green line, since it is very similar to the one for increasing covariance inflation and the one without variance inflation. We can see that that it is visually identical to the KKT-point u_*^τ, whereas the EKI solution, has more discrepancies. This can also be seen in the right image, where we illustrate the weighted norm of the solutions. The dotted black line, illustrates the upper bound, i.e. $\frac{1}{2}\|f^\dagger\|_C(s,t)$. We can see that the norms of all three solutions with CC (blue, green and pink line) increase over time but after reaching the upper bound they stay below this bound, whereas the EKI does not satisfy the bound, which explains why the error in figure 2 does not decrease.

5.3. 2D-Darcy flow

We consider the following elliptic PDE

$$
\begin{cases}
-\nabla \cdot (\exp(u)\,\nabla p) = f, & x \in D \\
\qquad\qquad\qquad p = 0, & x \in \partial D
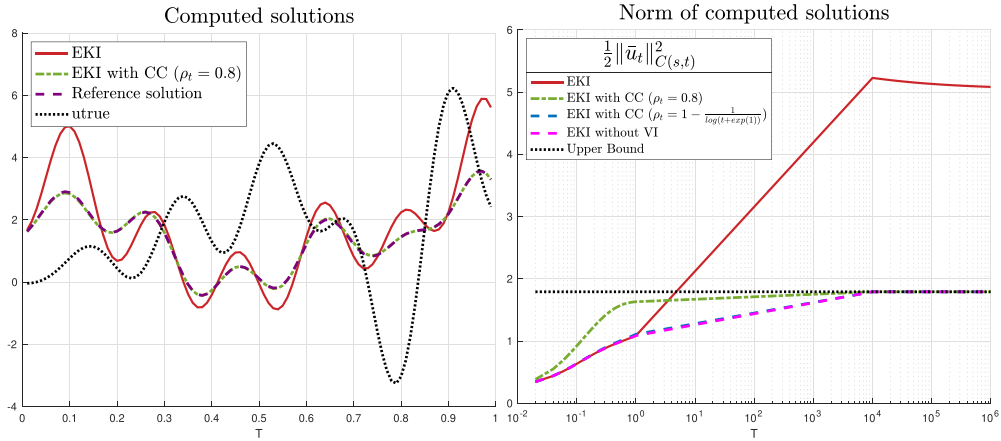\end{cases},
\tag{5.2}
$$

Figure 3. Estimated computed solutions (left) and weighted norm of the of the computed solutions (right).

where $D = (0,1)^2$ denotes the domain. Given observations of the solution $p \in H_0^1(D) \cap H^2(D) := \mathcal{V}$, our objective is to estimate the unknown diffusion coefficient $u^\dagger \in C^1(D) = X$. Additionally, we assume that the scalar field $f \in \mathbb{R}$ is known and define an observation operator for K randomly chosen points in X, i.e. $\mathcal{O}(p) = (p(x_1), \ldots, p(x_K))$.

The observations are represented by the equation

$$y = \mathcal{O}(p) + \eta.$$

The noise on our data, denoted by η is assumed to be Gaussian, representing a realisation of $\mathcal{N}(0,\Gamma)$, where $\Gamma = 0.1^2 \mathrm{Id}_K$.

Thus, our inverse problem is given by

$$y = \mathcal{G}(u) + \eta,$$

where $\mathcal{G} = \mathcal{O} \circ G$ and $G : X \to \mathbb{R}^K$ denotes the solution operator of the PDE (5.2). We solve the PDE on a uniform mesh of the size 32×32 using a FEM method with continuous, piece wise linear finite element basis functions. Consequently, the parameter space is of size $d = 1024$. As prior belief distribution we consider the random field

$$u(x,\omega) = \sum_{i=1}^{s} \lambda_i^{1/2} e_i(x) \xi_i(\omega),$$

where $\lambda_i = \left(\pi^2(k_j^2 + l_j^2) + \tau^2\right)^{-\alpha}$ and $e_i(x) = \cos(\pi x_1 k_j) \cos(\pi x_2 l_j)$ with $\tau = 0.01, \alpha = 2, s = 25, (k_j, l_j)_{j \in \{1,\ldots,s\}} \in \{1,\ldots,s\}^2$. The variables ξ_i are i.i.d standard normal variables.

We select our underlying groundtruth by drawing $u^\dagger$ from the prior distribution. Furthermore, we consider Box-constraints (BC) in the form of upper and lower bounds on the solution, i.e. the solutions have to satisfy $(u_t^{(j)})_i \in [a,b]$, where $a = \min u^\dagger + 0.3 |\min u^\dagger|$ and $b = \max u^\dagger - 0.3 |\max u^\dagger|$ for all $j = 1,\ldots,J, t \geqslant 0$ and components $i = 1,\ldots,d$.

Again we use MATLABs optimisation tool `fmincon` to compute the KKT point of (2.11) and solve the ODE (3.1) with `ode45`. We compare different levels of covariance inflation to illustrate the effect of adding the inflation, we choose
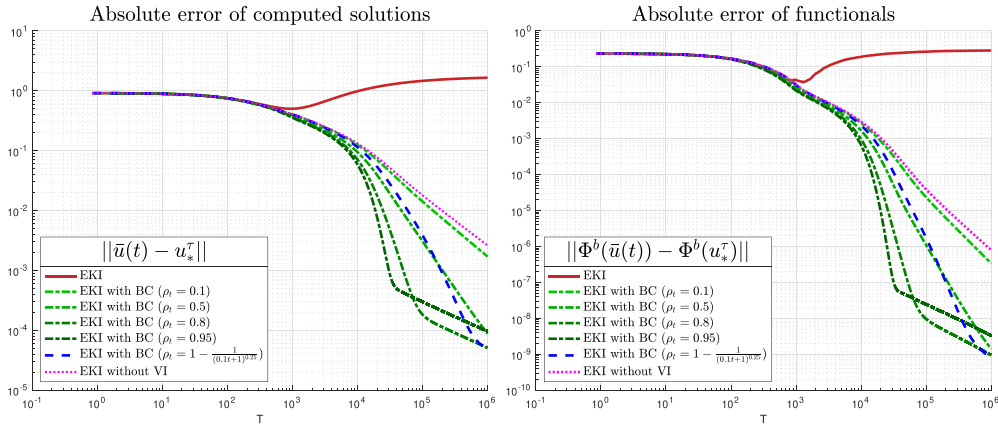
Figure 4. Absolute error of computed solutions in the parameter (left) and observation space (right). The red line illustrates the EKI, the green lines EKI with BC and different levels of constant covariance inflation, where the level increases for a darker shades of green, the blue line uses increasing variance inflation in the form of $\rho_t = 1 - \frac{1}{(0.1\,t+1)^{1/4}}$ and the pink line has no variance inflation.

 (i) Increasing covariance inflation $\rho_t = 1 - \frac{1}{(0.1\,t+1)^{1/4}}$.
 (ii) Constant covariance inflation $\rho_t = 0.1$ for all $t \geqslant 0$.
(iii) Constant covariance inflation $\rho_t = 0.5$ for all $t \geqslant 0$.
(iv) Constant covariance inflation $\rho_t = 0.8$ for all $t \geqslant 0$.
 (v) Constant covariance inflation $\rho_t = 0.95$ for all $t \geqslant 0$.
(vi) No variance inflation, i.e. $\rho_t = 0$ for all $t \geqslant 0$.

The particle size is set to $J = 10$ and therefore, the approach seeks a solution in the subspace $u_0^{\perp} + \operatorname{span}\{u_0^{(j)} - \bar{u}_0, j = 1, \ldots, J\}$.

We set a fixed penalty parameter $\tau = 10^4$, a regularisation factor of $\beta = 0.01$ and compute the solutions up until $T = 10^6$.

In figure 4 we can see similarly to the experiment in section 5.2 that the errors to the KKT point u_*^{τ} of (2.11) of the methods with BC perform similarly well and both converge to 0, whereas the original EKI does not converge. Furthermore, we can see that with an increasing amount of variance inflation the convergence rate also increases. However, the closer the fixed variance inflation factor $\rho_t = \rho$ is set to 1, the more involved is the computation and the ODE solver `ode45` takes a longer execution time. We emphasize that the adaptive choice $\rho_t \to 1$ can significantly reduces this execution time.

Figure 5 illustrates the estimated solutions of our methods. We can see that the EKI (bottom left) is outside the bounds, whereas our method with BC satisfies both upper and lower bound and looks very similar to the KKT-point obtained by MATLABs `fmincon` solver.

5.3.1. Adaptive penalty parameter.   For our last experiment with adaptive penalty parameter we consider a mesh of the size 6×6, since we have to compute the solutions for several different values of τ, which is computationally expensive. Therefore, we have $d = 36$. We take the same upper and lower bounds as above. Furthermore, we compute the solution of (5.1) until time $T = 10^5$ and consider fixed covariance inflation of the magnitude $\rho_t = 0.7$. We consider
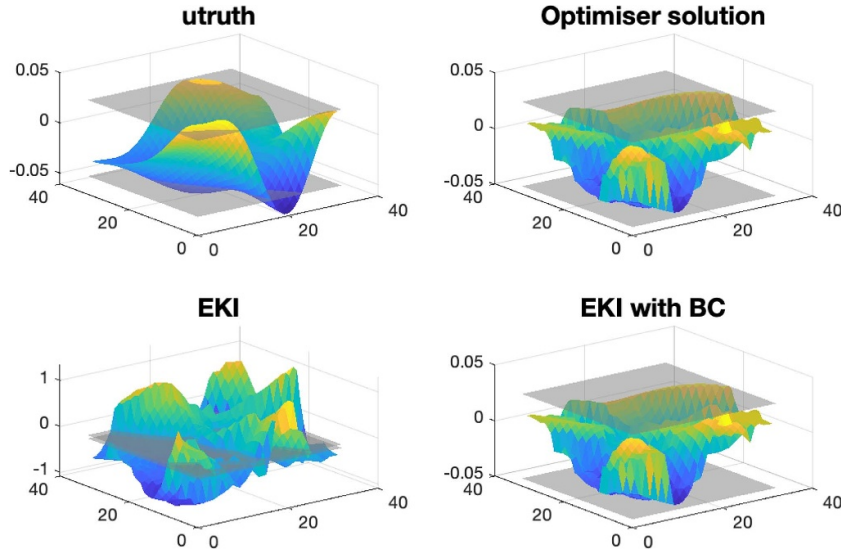
Figure 5. Estimated solutions of the EKI (bottom left) and EKI with BC, where $\rho_t = 1 - \frac{1}{(0.1\,t+1)^{1/4}}$ (bottom right) and optimiser (upper right). Upper left illustrates the true solution $u^\dagger$. The grey planes illustrate the upper and lower bound.
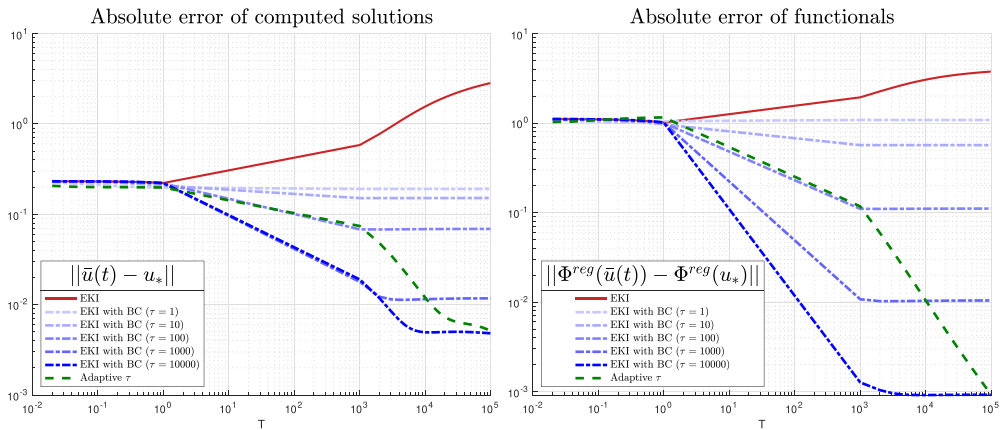

Figure 6. Absolute error of computed solutions in the parameter (left) and observation space (right). The red line depicts the EKI. The dotted/dashed blue lines illustrates the EKI with BC for different fixed values of $\tau \in \{1, 10, 100, 1000, 10000\}$. The dashed green line depicts the error of the the Box-constraints algorithm with adaptive increasing τ.

five different fixed penalty parameters $\tau \in \{1, 10, 100, 1000, 10000\}$ and one adaptive penalty parameter, $\tau(t) = t + 1$ to compare this solution with the one for the highest fixed penalty parameter, $\tau = 10\,000$. The remaining variables are the same as above.

In figure 6 we illustrate the errors in the parameter space and observation space to the KKT-point u_*. We can see that as the fixed penalty parameter τ increases, the error decreases as described by equation (2.12). Furthermore, for the adaptive choice of τ we can see that in

24

the beginning the error vanishes slower then for fixed $\tau = 10\,000$, but reaches the same error level at $T = 10^5$.

This results indicates that the adaptive choice of penalty parameter is a suitable alternative, since it results in a less stiff ODE that needs to be solved. The computational time of the adaptive choice was 4-times faster than the fixed penalty parameter.

6. Conclusion

Based on a log-barrier approach we introduced an adaptation of the EKI that allows the incorporation of convex inequality constraints for nonlinear forward problems. Through the inclusion of Tikhonov regularisation as well as covariance inflation we provided a convergence analysis for our method, which includes the quantification of the ensemble collapse as well as convergence as optimisation method. Our numerical experiments confirmed the results. In one experiment we consider a nonlinear forward problem given as linear problem with additional nonlinear perturbation. The magnitude of nonlinearity affects the amount of regularisation needed to obtain a strongly convex objective function and thus, satisfies our assumptions needed for theorem 4.2. In our second numerical experiment we considered the 2D-Darcy flow and showcased a method to address the problem of large penalty parameters that are needed in barrier-methods.

For future work we look to apply convex inequality constraints into subsampling approaches for the EKI [23]. Moreover, it would be interesting to incoporporate the proposed approach to particle based sampling methods such as the ensemble Kalman sampler [21].

Data availability statement

Acknowledgments

Appendix

A.1. Proofs of lemmas

Firstly, we show that strong convex functions fulfil the PL—inequality

Lemma A.1. *Let* $\Phi : \mathbb{R}^d \to \mathbb{R}$ *be strongly convex, then* Φ *satisfies the PL-inequality, i.e.*

$$\nu \|\nabla \Phi(x)\|^2 \geqslant \Phi(x) - \Phi(x^*),$$

for $\nu = \frac{1}{2\mu} > 0$ and all $x \in \mathbb{R}^d$, where x^* is a stationary point.

Proof. Since Φ is strongly convex, there exists a $\mu > 0$ such that

$$\Phi(y) \geqslant \Phi(x) + (y-x)\nabla\Phi(x) + \frac{\mu}{2}\|y-x\|^2, \quad \forall x,y \in \mathbb{R}^d$$

which implies

$$\Phi(x) - \Phi(y) \leqslant \langle x-y, \nabla\Phi(x)\rangle - \frac{\mu}{2}\|x-y\|^2. \quad \forall x,y \in \mathbb{R}^d.$$

Note that $\|a-b\|^2 = \|a^2\| - 2\langle a,b\rangle + \|b\|^2$ for all $a,b \in \mathbb{R}^d$. Therefore, we have $\langle a,b\rangle - \frac{1}{2}\|a^2\| = \frac{1}{2}\|b^2\| - \frac{1}{2}\|a-b\|^2$. Setting $a = (x-y)\sqrt{\mu}$ and $b = \frac{1}{\sqrt{\mu}}\nabla\Phi(x)$, we obtain

$$\Phi(x) - \Phi(y) \leqslant \frac{1}{2}\|\frac{1}{\sqrt{\mu}}\nabla\Phi(x)\|^2 - \frac{1}{2}\|(x-y)\sqrt{\mu} - \frac{1}{\sqrt{\mu}}\nabla\Phi(x)\|^2$$
$$\leqslant \frac{1}{2\mu}\|\nabla\Phi(x)\|^2.$$

Hence, Φ satisfies the PL-inequality with constant $\frac{1}{2\mu}$.

$\square$

The next statement considers our applied inverse problem based on a linear forward model with nonlinear perturbation. It provides sufficient conditions for the Tikhonov regularized loss function Φ^{reg} to be strongly convex.

Lemma A.2. *Consider the potential (1.2) where $G(u) = Au + \varepsilon[\sin(u_1), \ldots, \sin(u_K)]^T$, where $A \in \mathbb{R}^{K \times K}$ and $\varepsilon > 0$. Further, assume that $\sup_{u \in \Omega}\|u\| \leqslant B$ and let $A_{\max} = \max_{i,j}|A_{i,j}|$, $y_{\max} = \max_i|y_i|$ and $\lambda > ((4+B)a_{\max} + y_{\max})\varepsilon + \varepsilon^2$, then the Tikhonov regularised loss function*

$$\Phi^{\text{reg}}(u) = \frac{1}{2}\|G(u) - y\|^2 + \frac{\lambda}{2}\|u\|^2$$

is strongly convex and L-smooth.

Proof. The Jacobian J_G of G is given by

$$J_G(x) = \left(A^T + \varepsilon\text{diag}(\cos(x_1), \ldots, \cos(x_K))\right).$$

Hence, the ith row and jth component of the row have the following structure

$$(J_G(x))_{i,:} = A^T_{i,:} + \varepsilon[0, \ldots, \cos(x_i), \ldots, 0]$$
$$(J_G(x))_{i,j} = A_{i,j} + \varepsilon\cos(x_i)\delta_{ij},$$

where $\delta_{i,j}$ denotes the Kronecker symbol. Hence, we have

$$\left(\frac{\partial\Phi}{\partial x}\right)_i = \sum_{j=1}^K (A_{j,i} + \varepsilon\cos(x_i)\delta_{ij})(A_{j,:}x + \varepsilon\sin(x_j) - y_j).$$

And therefore, we obtain for the entries of the Hessian

$$\frac{\partial^2 \Phi}{\partial x_i \partial x_l} = \sum_{j=1}^{K} \left(-\varepsilon \sin\left(x_i\right) \delta_{ij} \delta_{il}\right) \left(A_{j,:}x + \varepsilon \sin\left(x_j\right) - y_j\right)$$

$$+ \sum_{j=1}^{K} \left(A_{j,i} + \varepsilon \cos\left(x_i\right) \delta_{ij}\right) \left(A_{j,l} + \varepsilon \cos\left(x_j\right) \delta_{jl}\right) + \lambda$$

$$= \left(-\varepsilon \sin\left(x_i\right) \delta_{il}\right) \left(A_{i,:}x + \varepsilon \sin\left(x_i\right) - y_i\right)$$

$$+ \sum_{j=1}^{K} A_{j,i}A_{j,l} + \sum_{j=1}^{K} A_{j,i} \varepsilon \cos\left(x_j\right) \delta_{jl}$$

$$+ \sum_{j=1}^{K} \varepsilon \cos\left(x_i\right) \delta_{ij} A_{j,l} + \varepsilon^2 \cos\left(x_i\right)^2 \delta_{il} + \lambda.$$

Thus, the entries of the Hesse-Matrix are given by

$$\left(H_{\Phi^{\mathrm{reg}}}\right)_{i,l}$$
$$= \begin{cases} \sum_{j=1}^{K} A_{j,i}^2 - \varepsilon \sin(x_i) \left(A_{i,:}x + \varepsilon \sin(x_i) - y_i\right) + 2A_{i,i} \varepsilon \cos(x_i) + \varepsilon^2 \cos(x_i)^2 + \lambda, & \text{if } i = l, \\ \sum_{j=1}^{K} A_{j,i}A_{j,l} + A_{l,i} \varepsilon \cos(x_l) + A_{i,l} \varepsilon \cos(x_i), & \text{if } \neq l. \end{cases}$$

We can split the Hessian into two parts:

$$\left(H_{\Phi^{\mathrm{reg}}}\right) = A^T A + M_{\varepsilon,\lambda},$$

where

$$M_{\varepsilon,\lambda} = \begin{cases} -\varepsilon \sin\left(x_i\right) \left(A_{i,:}x + \varepsilon \sin\left(x_i\right) - y_i\right) + 2A_{i,i} \varepsilon \cos\left(x_i\right) + \varepsilon^2 \cos\left(x_i\right)^2 + \lambda, & \text{if } i = l, \\ \varepsilon \cos\left(x_l\right) + A_{i,l} \varepsilon \cos\left(x_i\right), & \text{if } i \neq l. \end{cases}$$

Since $A^T A$ is positive semi-definite we continue to show that $M_{\varepsilon,\lambda}$ is positive definite. We start by lower bounding the diagonal elements of $\left(M_{\varepsilon,\lambda}\right)$ by

$$\left(M_{\varepsilon,\lambda}\right)_{i,i} \geqslant -\varepsilon \left(A_{\max} B + \varepsilon + y_{\max}\right) - 2A_{i,i}\varepsilon + \lambda,$$

where $A_{\max} = \max_{i,j} |A_{i,j}|$, $y_{\max} = \max_i |y_i|$ and $B = \sup_{x \in \Omega} \|x\| < \infty$. On the other side, we can bound the off-diagonal elements by

$$\sum_{l=1}^{K} |\left(M_{\varepsilon,\lambda}\right)_{i,l}| \leqslant 2A_{\max}\varepsilon.$$

Hence, by assumption we have for all $i \in \{1, \dots, K\}$

$$\left(M_{\varepsilon,\lambda}\right)_{i,i} \geqslant -\varepsilon \left(A_{\max} B + \varepsilon + y_{\max}\right) - 2A_{i,i}\varepsilon + \lambda > 2A_{\max}\varepsilon \geqslant \sum_{i \neq j} |\left(M_{\varepsilon,\lambda}\right)_{i,j}|.$$

Therefore, $M_{\varepsilon,\lambda}$ is positive definite and by that also $H_{\Phi^{\mathrm{reg}}}$. Furthermore, we can upper bound the diagonal elements of $M_{\varepsilon,\lambda}$ by

$$\left(M_{\varepsilon,\lambda}\right)_{i,i} \leqslant \varepsilon \left(A_{\max} B + \varepsilon + y_{\max}\right) + 2A_{i,i}\varepsilon + \varepsilon^2 + \lambda.$$

Hence, all entries of $H_{\Phi^{\mathrm{reg}}}$ are bounded, implying the L-smothness of Φ^{reg}. $\qquad\square$

The next lemma provides the solution of the ODE needed to derive the lower bound on the ensemble collapse in lemma 3.2.

Lemma A.3. *Let* $a, b, c > 0$ *then the solution of the ODE*

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = -\frac{a}{t+b}x(t) - cx(t)^2 \tag{A.1}$$

is given by

$$x(t) = \frac{1-a}{c(t+b) + (1-a)k_1(t+b)^a}, \quad x(0) = x_0 > 0, \tag{A.2}$$

where k_1 *is chosen such that the initial condition is satisfied.*

Proof. The proof follows the ideas of lemma A.3 in [42]. We show that that (A.2) solves the ODE:

$$\begin{aligned}
\frac{\mathrm{d}x(t)}{\mathrm{d}t} &= -\frac{(1-a)\left(c + a(1-a)k_1(t+b)^{a-1}\right)}{\left(c(t+b) + (1-a)k_1(t+b)^a\right)^2} \\
&= \left(-\frac{c + a(1-a)k_1(t+b)^{a-1}}{c(t+b) + (1-a)k_1(t+b)^a}\right)\left(\frac{1-a}{c(t+b) + (1-a)k_1(t+b)^a}\right) \\
&= \left(-\frac{c + ca - ca + a(1-a)k_1(t+b)^{a-1}}{c(t+b) + (1-a)k_1(t+b)^a}\right)x(t) \\
&= \left(-\frac{ca + a(1-a)k_1(t+b)^{a-1} + c(1-a)}{c(t+b) + (1-a)k_1(t+b)^a}\right)x(t) \\
&= \left(-\frac{a(t+b)\left(c + (1-a)k_1(t+b)^{a-1}\right)}{(t+b)(c(t+b) + (1-a)k_1(t+b)^a)} - cx(t)\right)x(t) \\
&= \left(-\frac{a}{t+b} - cx(t)\right)x(t),
\end{aligned}$$

which is in the needed form of (A.1). $\qquad\square$

A.2. Proof of the feasibility and unique existence of solution

Proof of proposition 3.4. Without loss of generality we assume that $m = 1$, i.e. there is only one inequality constraint $h(u) \leqslant 0$. The evolution of $h(\bar{u}_t)$ is given by

$$\frac{\mathrm{d}h(\bar{u}_t)}{\mathrm{d}t} = \langle \nabla h(\bar{u}_t), \frac{\mathrm{d}\bar{u}_t}{\mathrm{d}t}\rangle = \langle \nabla h(\bar{u}_t), v(u_t)\rangle + \frac{1}{\tau}\frac{1}{h(\bar{u}_t)}\langle \nabla h(\bar{u}_t), \widehat{C}(u_t)\nabla h(\bar{u}_t)\rangle,$$

where v is defined in (2.17). Let $T > 0$, then for any solution $(\bar{u}_t)_{t \in [0,T]}$, $h(\bar{u}_T)$ can be represented by

$$h(\bar{u}_T) = h(\bar{u}_0) + \int_0^T \left(\langle \nabla h(\bar{u}_t), v(u_t)\rangle + \frac{1}{\tau}\frac{1}{h(\bar{u}_t)}\langle \nabla h(\bar{u}_t), \widehat{C}(u_t)\nabla h(\bar{u}_t)\rangle\right)\mathrm{d}t.$$

Suppose there exists $T_1 > 0$ such that $h(\bar{u}_{T_1}) > 0$. Since we initialize with $h(\bar{u}_0) < 0$ by continuity of solutions and continuity of h, it follows that there exists $t_{\min} \in (0, T_1)$ such that

$h(\bar{u}_{t_{\min}}) = 0$ and $h(\bar{u}_t) < 0$ for all $t < t_{\min}$. Moreover, again by continuity it follows that there must exists $\delta > 0$ such that

$$\langle \nabla h(\bar{u}_t), v(u_t) \rangle + \frac{1}{\tau} \frac{1}{h(\bar{u}_t)} \langle \nabla h(\bar{u}_t), \widehat{C}(u_t) \nabla h(\bar{u}_t) \rangle > 0 \tag{A.3}$$

for all $t \in (t_{\min} - \delta, t_{\min})$ and therefore, also $\nabla h(\bar{u}_t) \neq 0$ for all $t \in (t_{\min} - \delta, t_{\min})$. Remember, by assumption 2.2 we also have that $\nabla h(\bar{u}_{t_{\min}}) \neq 0$, see remark 2.3. Next, for $s \in [0, t_{\min})$ we consider

$$v(u_s) \leqslant \|\widehat{C}_s^{u,G}\|_F \|\Gamma^{-1}\|_F (\|\bar{G}_s\| + \|y\|) + \|\widehat{C}(u_s)\|_F \|C_0^{-1}\|_F \|\bar{u}_s\|.$$

We can approximate the covariance $\widehat{C}_s^{u,G}$ by the following

$$\|\widehat{C}_s^{u,G}\|_F^2 \leqslant \left(\frac{1}{J} \sum_{j=1}^{J} \|e_s^{(j)}\|^2 \right) \left(\frac{1}{J} \sum_{j=1}^{J} \|G\left(u_s^{(j)}\right) - \bar{G}_s\|^2 \right)$$

using a same argumentation as in proof of lemma 3.2. We approximate the first term by

$$\frac{1}{J} \sum_{j=1}^{J} \|e_s^{(j)}\|^2 = V_e(s) \leqslant V_e(0)$$

and the second factor by

$$\frac{1}{J} \sum_{j=1}^{J} \|G\left(u_s^{(j)}\right) - \bar{G}\|^2 \leqslant c_{\text{lip}}^2 V_e(s) \leqslant c_{\text{lip}}^2 V_e(0),$$

where we used lemma 3.1. Hence, we have $\|v(u_s)\| \leqslant B$ for all $s \in [0, t_{\min}]$ and some $B > 0$. Similarly, using that $\|\nabla h(u)\| \leqslant C(R)$, we have that $|\langle \nabla h(\bar{u}_s), v(u_s) \rangle| \leqslant B$ for $s \in [0, t_{\min}]$ and some constant B. Finally, we observe that

$$\frac{1}{\tau} \frac{1}{h(\bar{u}_s)} \langle \nabla h(\bar{u}_s), \widehat{C}(u_s) \nabla h(\bar{u}_s) \rangle < 0$$

for all $s \in [0, t_{\min})$, since $\langle \nabla h(\bar{u}_s), \widehat{C}(u_s) \nabla h(\bar{u}_s) \rangle > 0$ using lemma 3.2. Therefore, using

$$\lim_{\varepsilon \to 0} \frac{1}{h(\bar{u}_{t_{\min} - \varepsilon})} = -\infty$$

since $h(\bar{u}_{t_{\min} - \varepsilon}) \nearrow 0$ for $\varepsilon \to 0$, and $\nabla h(\bar{u}_t) \neq 0$ for all $t \in (t_{\min} - \delta, t_{\min}]$, we obtain

$$\lim_{\varepsilon \to 0} \langle \nabla h(\bar{u}_{t_{\min} - \varepsilon}), v(u_{t_{\min} - \varepsilon}) \rangle + \frac{1}{\tau} \frac{1}{h(\bar{u}_{t_{\min} - \varepsilon})} \langle \nabla h(\bar{u}_{t_{\min} - \varepsilon}), \widehat{C}(u_{t_{\min} - \varepsilon}) \nabla h(\bar{u}_{t_{\min} - \varepsilon}) \rangle = -\infty$$

which is in contradiction to (A.3). For the second part, we know by local Lipschitz-continuity that locally there exists a unique solution of (2.15). Hence, as the local solution remains in Ω it is also a global solution. $\qquad\square$

ORCID iDs

Matei Hanu ⬤ https://orcid.org/0000-0002-6494-8971
Simon Weissmann ⬤ https://orcid.org/0000-0002-5111-6658

References

[1] 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
[2] Amor N, Rasool G and Bouaynaya N C 2018 Constrained state estimation - a review (arXiv: 1807.03463)
[3] Anderson J L 2007 An adaptive covariance inflation error correction algorithm for ensemble filters *Tellus* A **59** 210–24
[4] Bergemann K and Reich S 2010 A localization technique for ensemble Kalman filters *Q. J. R. Meteorol. Soc.* **136** 701–7
[5] Bergemann K and Reich S 2010 A mollified ensemble Kalman filter *Q. J. R. Meteorol. Soc.* **136** 1636–43
[6] Bertsekas D P 1982 Projected newton methods for optimization problems with simple constraints *SIAM J. Control Optim.* **20** 221–46
[7] Bertsekas D P 1997 Nonlinear programming *J. Oper. Res. Soc.* **48** 334–334
[8] Blömker D, Schillings C, Wacker P and Weissmann S 2022 Continuous time limit of the stochastic ensemble Kalman inversion: strong convergence analysis *SIAM J. Numer. Anal.* **60** 3181–215
[9] Blömker D, Schillings C, Wacker P and Weissmann S 2019 Well posedness and convergence analysis of the ensemble Kalman inversion *Inverse Problems* **35** 085007
[10] Boyd S and Vandenberghe L 2004 *Convex Optimization* (Cambridge University Press) (https://doi.org/10.1017/CBO9780511804441)
[11] Bungert L and Wacker P 2023 Complete deterministic dynamics and spectral decomposition of the linear ensemble Kalman inversion *SIAM/ASA J. Uncertain. Quantification* **11** 320–57
[12] Calvello E, Reich S and Stuart A M 2022 Ensemble Kalman methods: a mean field perspective (arXiv:2209.11371)
[13] Carrillo J A, Hoffmann F, Stuart A M and Vaes U 2022 Consensus-based sampling *Stud. Appl. Math.* **148** 1069–140
[14] Chada N K, Schillings C and Weissmann S 2019 On the incorporation of box-constraints for ensemble Kalman inversion *Found. Data Sci.* **1** 433–56
[15] Chada N K, Stuart A M and Tong X T 2020 Tikhonov regularization within ensemble Kalman inversion *SIAM J. Numer. Anal.* **58** 1263–94
[16] Chada N K and Tong X T 2019 Convergence acceleration of ensemble Kalman inversion in nonlinear settings *Math. Comput.* **91** 1247–80
[17] Crane R and Roosta F 2021 Invexifying regularization of non-linear least-squares problems (https://doi.org/https://doi.org/10.48550/arXiv.2111.11027)
[18] Ding Z and Li Q 2021 Ensemble Kalman inversion: mean-field limit and convergence analysis *Stat. Comput.* **31** 9
[19] Eigel M, Gruhlke R and Sommer D 2022 SIAM Journal on Applied Dynamical Systems Less interaction with forward models in langevin dynamics (https://doi.org/10.1137/23M1546841)
[20] Evensen G 2003 The ensemble Kalman filter: theoretical formulation and practical implementation *Ocean Dyn.* **53** 343–67
[21] Garbuno-Inigo A, Hoffmann F, Li W and Stuart A M 2020 Interacting Langevin diffusions: Gradient structure and ensemble Kalman sampler *SIAM J. Appl. Dyn. Syst.* **19** 412–41
[22] Garbuno-Inigo A, Nüsken N and Reich S 2020 Affine invariant interacting langevin dynamics for bayesian inference *SIAM J. Appl. Dyn. Syst.* **19** 1633–58
[23] Hanu M, Latz J and Schillings C 2023 Subsampling in ensemble Kalman inversion *Inverse Problems* **39** 094002
[24] Herty M and Visconti G 2020 Continuous limits for constrained ensemble Kalman filter *Inverse Problems* **36** 075006
[25] Iglesias M A 2014 Iterative regularization for ensemble data assimilation in reservoir models *Comput. Geosci.* **19** 177–212

[26] Iglesias M A 2016 A regularizing iterative ensemble Kalman method for PDE-constrained inverse problems *Inverse Problems* **32** 025002

[27] Iglesias M A, Law K J H and Stuart A M 2013 Ensemble Kalman methods for inverse problems *Inverse Problems* **29** 045001

[28] Kandepu R, Imsland L and Foss B A 2008 Constrained state estimation using the unscented Kalman filter *2008 16th Mediterranean Conf. on Control and Automation* pp 1453–8

[29] Karimi H, Nutini J and Schmidt M 2016 Linear convergence of gradient and proximal-gradient methods under the Polyak-ŁOjasiewicz condition *Machine Learning and Knowledge Discovery in Databases* ed P Frasconi, N Landwehr, G Manco and J Vreeken (Springer) pp 795–811

[30] Kokurin M Y 2010 Convexity of the Tikhonov functional and iteratively regularized methods for solving irregular nonlinear operator equations *Comput. Math. Math. Phys.* **50** 620–32

[31] Li G and Reynolds A 2009 Iterative ensemble Kalman filters for data assimilation *SPE J.* **14** 496–505

[32] Liu S, Reich S and Tong X T 2023 Dropout ensemble Kalman inversion for high dimensional inverse problems (https://doi.org/10.48550/arXiv.2308.16784)

[33] Pathiraja S and Reich S 2019 Discrete gradients for computational bayesian inference *J. Comput. Dyn.* **6** 385–400

[34] Reich S and Weissmann S 2021 Fokker–planck particle systems for bayesian inference: computational approaches *SIAM/ASA J. Uncertain. Quantification* **9** 446–82

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

[36] Schillings C and Stuart A M 2017 Convergence analysis of ensemble Kalman inversion: the linear, noisy case *Appl. Anal.* **97** 107–23

[37] Schmidt M, Kim D and Sra S 2011 Projected newton-type methods in machine learning *Optimization for Machine Learning* (The MIT Press) (https://doi.org/10.7551/mitpress/8996.003.0013)

[38] Shikhman V and Stein O 2009 Constrained optimization: projected gradient flows *J. Optim. Theory Appl.* **140** 117–30

[39] Simon D 2010 Kalman filtering with state constraints: a survey of linear and nonlinear algorithms *Control Theory Appl. IET* **4** 1303–18

[40] Tong X T, Majda A J and Kelly D 2015 Nonlinear stability of the ensemble Kalman filter with adaptive covariance inflation *Commun. Math. Sci.* **14** 1283–313

[41] Tong X T, Majda A J and Kelly D 2016 Nonlinear stability and ergodicity of ensemble based Kalman filters *Nonlinearity* **29** 657–91

[42] Tong X T and Morzfeld M 2023 Localized ensemble Kalman inversion *Inverse Problems* **39** 064002

[43] Wang D, Chen Y and Cai X 2009 State and parameter estimation of hydrologic models using the constrained ensemble Kalman filter *Water Resour. Res.* **45** 11

[44] Weissmann S 2022 Gradient flow structure and convergence analysis of the ensemble Kalman inversion for nonlinear forward models *Inverse Problems* **38** 105011

[45] Weissmann S, Chada N K, Schillings C and Tong X T 2022 Adaptive Tikhonov strategies for stochastic ensemble Kalman inversion *Inverse Problems* **38** 045009