OPEN ACCESS
IOP Publishing
Inverse Problems
Inverse Problems **41** (2025) 105010 (36pp)
https://doi.org/10.1088/1361-6420/ae115f

Accuracy improvement in ensemble Kalman inversion through data-informed ensemble selection

Ruben Harris* **and Claudia Schillings**

Fachbereich Mathematik und Informatik, Freie Universität Berlin, Arnimallee 6, 14195 Berlin, Germany

E-mail: r.harris@fu-berlin.de and c.schillings@fu-berlin.de

Received 14 March 2025; revised 20 August 2025
Accepted for publication 9 October 2025
Published 24 October 2025

Abstract

The ensemble Kalman inversion (EKI) method is widely used for solving inverse problems, leveraging ensemble-based techniques to iteratively refine parameter estimates. Despite its versatility, the accuracy of EKI is constrained by the subspace spanned by the initial ensemble, which may poorly represent the solution in cases of limited prior knowledge. This work addresses these limitations by optimising the subspace in which EKI operates, improving accuracy and computational efficiency. We derive a theoretical framework for constructing optimal subspaces in linear settings and extend these insights to nonlinear cases. A novel greedy strategy for selecting initial ensemble members is proposed, incorporating prior, data, and model information to enhance performance. Numerical experiments on both linear and nonlinear problems demonstrate the effectiveness of the approach, offering a significant advancement in the accuracy and scalability of EKI for high-dimensional and ill-posed problems.

Keywords: ensemble Kalman inversion, particle system, optimisation, Bayesian inverse problems

* Author to whom any correspondence should be addressed.

1. Introduction

Mathematical models are indispensable for understanding a wide array of systems of scientific interest. These models facilitate analysis and prediction of system behaviours, but their accuracy relies on calibration to observed data—a process referred to as 'inversion'. Inversion is the cornerstone of fields such as weather forecasting, medical imaging and machine learning that employ various techniques categorised as either variational/optimisation-based or Bayesian/statistical.

One notable inversion method is ensemble Kalman inversion (EKI), which integrates elements from both variational and Bayesian approaches. Introduced by Iglesias *et al* (2013), EKI iteratively employs an Ensemble Kalman–Bucy Filter to tackle inverse problems. In linear scenarios, this takes the form of a preconditioned gradient flow equation (see Schillings and Stuart 2017b).

The structure of the EKI update step results in the so-called subspace property (Iglesias *et al* 2013), whereby the EKI estimate remains confined to the subspace spanned by the initial ensemble. While this property acts as a natural regulariser, it can also lead to poor approximations if the subspace does not adequately capture the true solution. Traditionally, the initial ensemble is derived from prior knowledge and so, in the case of limited or uninformative priors, the chosen initial ensemble may result in a subspace poorly aligned with the true solution.

This work addresses these limitations by optimising the subspace within the EKI framework to improve both computational efficiency and solution accuracy. By constructing a subspace that takes into account the forward operator and the data set in addition to just the priors, we aim to maximise accuracy within a fixed computational budget (in terms of particle count). This refined approach leverages the structure of the inverse problem to make EKI more scalable and accurate, especially when dealing with limited or inaccurate prior information.

1.1. Literature overview

Since it was first developed by Evensen (2003), the Ensemble Kalman Filter (EnKF) has become widely used in both data assimilation and inverse problems. Its popularity is largely due to its straightforward implementation and effectiveness even with smaller ensemble sizes, as shown in various studies (see Bergemann and Reich 2009, 2010, Li and Reynolds 2009, Iglesias *et al* 2013, Iglesias 2014, 2016). This feature makes EnKF particularly well-suited to high-dimensional problems, where other inversion methods may face computational challenges. EKI was first introduced by Iglesias *et al* (2013) as an extension of EnKF to inverse problems. EKI sets up artificial dynamics in which the target data acts as the observed signal, which is repeatedly integrated into the model estimates. A key advantage of EKI over traditional inversion methods is its derivative-free formulation, which allows it to function as a computationally efficient black-box method. EKI also inherits methodological ideas from ensemble smoothing and randomised likelihood optimisation, which have influenced both its formulation and implementation. Early developments in ensemble smoothers (Evensen and van Leeuwen 2000) and ensemble-based optimisation methods for inverse problems (Chen *et al* 2009, Chen and Oliver 2012) have contributed valuable insights into the use of ensemble approximations for joint parameter and state estimation. These connections highlight the broader methodological lineage of EKI, linking it to data assimilation and iterative likelihood-based optimisation.

Research on EKI has increasingly focused on understanding convergence and stability. Much of this analysis was has been based on studying the continuous-time limit of EKI.

Stability studies have been initiated in Tong *et al* 2015, 2016 and convergence analysis to and in the continuous time limit can be found in Schillings and Stuart 2017a, 2017b, Blömker *et al* 2019, 2021, Bungert and Wacker 2021. These studies show that convergence in parameter space often requires regularisation; Tikhonov regularisation has been extensively examined in this context (see Iglesias 2016, Chada *et al* 2020, Parzer and Scherzer 2022). Recently, adaptive regularisation strategies were introduced to increase the robustness of stochastic EKI (see Iglesias and Yang 2021, Weissmann *et al* 2022). The mean-field limit of EKI has also been analysed, providing valuable insights into ensemble dynamics as the ensemble size grows large (see Ding and Li (2020), Calvello *et al* 2025).

To gain a foundational understanding of EKI, researchers have frequently studied its behaviour under linear forward operators. Iglesias *et al* (2013) demonstrated that—in the linear, Gaussian setting—EKI converges to the maximum *a posteriori* (MAP) estimate in the mean-field limit (or, equally, minimises a Tikhonov-regularised objective). In general, however, the ensemble is not an approximation of the posterior in the Bayesian setting (see Ernst *et al* 2015). Further analysis by Schillings and Stuart (2017b) showed that in the linear, noiseless regime, EKI experiences ensemble collapse, whereby particle positions converge to an estimate of the MAP within a specific subspace. Connections to other optimisation methods have been explored in Chada *et al* (2021), Weissmann (2022). The evolution of the EKI ensemble was further analysed by Bungert and Wacker (2021), who provided explicit expressions for the particle trajectories in the linear, noiseless regime. In the discrete setting, an analysis of the particle trajectories can be found in Qian and Beattie (2025).

In practical applications, the choice of initial ensemble is particularly important, especially since the number of particles in the ensemble is often far smaller than the dimension of the observation or state space. Iglesias *et al* (2013) showed that particles remain in the subspace spanned by the initial ensemble, thus regularising the inverse problem. This subspace is therefore a critical design parameter in EKI. When prior knowledge is represented by a Gaussian distribution, two strategies have been suggested for selecting initial ensemble members: random sampling from the Gaussian prior or utilising the terms in the Karhunen–Loéve (KL) expansion that correspond to the largest eigenvalues. In the linear setting, Ghattas and Sanz-Alonso (2022) provided theoretical support for the random sampling approach, showing that this technique approximates the posterior mean and covariance with error bounds that decay according to particle count and the 'effective dimension' of the prior covariance.

The subspace property is broken, however, by popular variations of EKI that have proven remarkably effective in practice. Techniques such as localisation and sample error correction (SEC) alter the empirical covariance in EKI—which can be highly sensitive to noise, especially for small ensembles—to reduce variance and to remove spurious covariances. Localisation techniques typically cause correlation between parameter components to decay with distance (when a natural notion of distance is available). More generally, SEC techniques make no such spatial assumption and are structure-agnostic approaches to reducing error in covariance estimation (see Vishny *et al* (2024) for a discussion of SEC techniques). By altering the empirical covariance, such techniques may break the subspace property at each step of EKI, potentially losing its regularising effects while allowing for greater parameter space exploration. By contrast, the proposed method in this paper can only break the subspace property at user-defined times, leading to a different trade-off between regularisation and parameter space exploration.

This paper is concerned with optimising the subspace within which the ensemble evolves and the choice of initial ensemble within this subspace. This refinement of the initial subspace promises to improve the accuracy of EKI and its scalability to high-dimensional inverse problems. A common strategy for initialising the ensemble in ensemble Kalman methods is

to sample from a truncated KL expansion of a Gaussian prior. This approach provides an efficient low-rank approximation of the prior, aligns with the subspace property of EKI, and ensures that the ensemble reflects the dominant modes of prior uncertainty (Evensen 2009, Stuart 2010, Iglesias *et al* 2013). It is particularly effective in high-dimensional or infinite-dimensional settings where the prior encodes smoothness through the decay of eigenvalues of the covariance operator. The motivation of our work is the exploration of more informed strategies for constructing the initial ensemble, going beyond standard KL-based sampling to incorporate problem-specific structure and data information. The continued development of subspace optimisation, alongside regularisation techniques, holds significant potential for improving EKI's performance in computationally challenging settings.

The idea of subspace optimisation in EKI is conceptually related to likelihood-informed subspace (LIS) methods in Bayesian inference, which identify low-dimensional subspaces where the likelihood is most informative relative to the prior. While LIS is typically used to approximate posterior distributions, our method focuses on accurate point estimation within a carefully chosen subspace. Moreover, standard LIS methods construct fixed subspaces off-line, often based on posterior samples or local Hessian information. In contrast, our approach constructs subspaces in a data-adaptive and sample-free manner, tailored to each realisation of the data. We note that data-dependent LIS variants have been proposed-see, e.g. Cui *et al* (2014), Spantini *et al* (2015), but these typically involve sampling.

1.2. Main ideas and contributions of the paper

In this work, we focus on improving the accuracy of EKI with small ensembles by optimising the subspace in which the algorithm operates. We aim to improve on the subspace arising from the standard ensemble initialisation, which might lead to low accuracy in the case of poor prior knowledge. By optimising the EKI subspace, we aim to maximise accuracy within a given computational budget (i.e. with a given number of particles). Our contributions are as follows:

- We derive closed-form expressions for the linear, noise-free particle dynamics. Though these new expressions are equivalent to those already derived by Bungert and Wacker (2021), their particular form facilitates the analysis of the choice of ensemble subspace.
- We generalise a standard approach for generating the initial ensemble by (1) expanding the choice of subspaces in which to situate the initial ensemble (2) allowing the initial ensemble to consist of any set of linearly independent vectors in a chosen subspace. Using our analysis of the dynamics of the particles in the linear, noiseless regime, we evaluate the 'goodness' of a given choice of subspace and linear combination.
- We propose a scalable, greedy strategy to select good subspaces. We characterise families of 'optimal' linear combinations given a choice of subspace and propose concrete methods for generating such linear combinations. These selection strategies incorporate prior, data and model knowledge.
- Based on insights from the linear, noiseless case, we propose a strategy for adaptively resampling the EKI ensemble that can be applied to general forward operators (i.e. not just linear forward operators).
- We numerically demonstrate the effectiveness of our approach by applying it to both linear and nonlinear example problems.

This paper is organised as follows. In sections 2 and 3, we introduce the problem setting and EKI methodology. Section 4 covers the analysis of the linear, noiseless particle dynamics, the characterisation of optimal initial ensembles in this setting, and strategies for constructing good ensembles in both the linear and nonlinear setting. Numerical experiments are presented in section 5 and conclusions are drawn in section 6. Auxiliary results and proofs of the main theorems can be found in the appendices.

1.3. Notation

For any $n \in \mathbb{N}^+$, we consider $\mathbb{R}^n$, equipped with either the standard inner product $\langle \cdot, \cdot \rangle$ and its norm $|\cdot|$ or the weighted inner product $\langle \cdot, \cdot \rangle_\Gamma := \langle \cdot, \Gamma^{-1} \cdot \rangle$ and its norm $|\cdot|_\Gamma$, where $\Gamma \in \mathbb{R}^{n \times n}$ is symmetric and positive definite. Further, $(\Omega, \mathcal{A}, \mathbb{P})$ denotes a probability space, X denotes a separable Hilbert space and $\mathcal{B}X := \mathcal{B}(X, |\cdot|)$ denotes the Borel σ-algebra on X (or $\mathcal{B}X := \mathcal{B}(X, |\cdot|_\Gamma)$). Introducing an additional space $\mathbb{R}^m$ with $m \in \mathbb{N}^+$, we denote the tensor product of vectors $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ by $x \otimes y := xy^\top$.

2. Inverse problem

In the inverse problem setting, our goal is to identify the unknown parameters $u \in X$ from noisy observations:

$$y = G(u) + \eta, \tag{2.1}$$

where $\eta \in Y$ represents additive observational noise, and $G : X \to Y$ is the operator that solves the underlying forward problem, with X being a separable Hilbert space and $Y = \mathbb{R}^m$ for some $m \in \mathbb{N}^+$.

Within the Bayesian framework, u and η are modelled as independent random variables—$u : \Omega \to X$ and $\eta : \Omega \to Y$—with the assumption that $u \perp \eta$. Assuming the noise η follows a Gaussian distribution $\eta \sim \mathrm{N}(0, \Gamma)$ and the operator G is continuous, the posterior distribution μ^y is given by

$$\mathrm{d}\mu^y(u) = \frac{1}{Z} \exp\left(-\frac{1}{2} |y - G(u)|_\Gamma^2 \right) \mathrm{d}\mu_0(u), \tag{2.2}$$

where μ_0 denotes the prior distribution on the unknown parameters, $l(u; y) := \frac{1}{2}|y - G(u)|_\Gamma^2$ represents the likelihood and $Z = \mathbb{E}_{\mu_0}\left[\exp\left(-\frac{1}{2}|y - G(u)|_\Gamma^2\right)\right]$ is the normalisation constant (which is assumed to be positive).

The primary focus of our study is the computation of the MAP estimate, which provides a point estimate for the unknown parameter. In a finite-dimensional context, this involves minimising the negative log density. Given a Gaussian prior on the unknown parameters $u \sim \mathcal{N}(\mu, R)$, the objective function that we seek to minimise is given by

$$\Phi(u) := \frac{1}{2}|y - G(u)|_\Gamma^2 + \frac{1}{2}|u - \mu|_R^2, \tag{2.3}$$

which means the MAP estimate is the point that maximises the posterior density. For a generalisation of MAP points to the infinite-dimensional setting, we refer to Dashti *et al* (2013), Klebanov and Wacker (2023).

We note that, under the transformations $\tilde{u} := u - \mu$, $\tilde{y} := \Gamma^{-\frac{1}{2}} y$, $\tilde{G}(u) := \Gamma^{-\frac{1}{2}} G(u + \mu)$, we can equivalently minimise the following objective function:

$$\frac{1}{2}|\tilde{y} - \tilde{G}(\tilde{u})|^2 + \frac{1}{2}|\tilde{u}|_R^2.$$

As such, we assume, without loss of generality, that $\Gamma = I$ and $\mu = 0$ in our analysis in section 4.

3. EKI

Our objective is to address the inverse problem specified in equation (2.1) using the EKI framework, with a particular focus on the continuous-time version of EKI that is described in Schillings and Stuart (2017b). We begin by outlining the derivation of EKI in discrete time and then discuss the transition to the continuous-time limit. Specifically, we employ a homotopy approach that approximates the posterior distribution μ^y defined in (2.2) through a sequence of probability measures μ_n^y for $n = 1, \ldots, N$:

$$\mu_n^y(\mathrm{d}u) \propto \exp\left(-nhl(u;y)\right)\mu_0(\mathrm{d}u), \quad h = N^{-1},$$

where the target measure $\mu_N^y = \mu^y$ represents the posterior distribution of u given y. The update from μ_n^y to μ_{n+1}^y is defined by

$$\mu_{n+1}^y(\mathrm{d}u) = \frac{1}{Z_n^y}\exp\left(-hl(u;y)\right)\mu_n^y(\mathrm{d}u).$$

With the normalisation constant Z_n^y computed as

$$Z_n^y = \int \exp\left(-hl(u;y)\right)\mu_n^y(\mathrm{d}u).$$

These intermediate measures can be interpreted as the repeated use of the observational data in a filtering context, where the observational noise is amplified by a factor of $1/h$. The measures are then approximated by an ensemble of J particles for some $J \in \mathbb{N}^+$. The initial ensemble is defined as $u_0 := \{u_0^{(i)}\}_{j=1}^J \in X^J$.

Applying the Ensemble Kalman Filter (EnKF) to this problem results in the following iterative update:

$$u_{n+1}^{(i)} = u_n^{(i)} + C^{up}(u_n)\left(C^{pp}(u_n) + h^{-1}\Gamma\right)^{-1}\left(y_{n+1}^{(i)} + \zeta_{n+1}^{(i)} - G\left(u_n^{(i)}\right)\right), \quad j = 1, \ldots, J, \tag{3.1}$$

where $\zeta_{n+1}^{(i)} \sim \mathcal{N}(0, \frac{1}{h}\Gamma)$ i.i.d. The operators C^{pp} and C^{up} are defined as

$$\begin{aligned}
C^{pp}(u) &= \frac{1}{J}\sum_{j=1}^{J}\left(G\left(u^{(j)}\right) - \overline{G}\right)\left(G\left(u^{(j)}\right) - \overline{G}\right)^T, \\
C^{up}(u) &= \frac{1}{J}\sum_{j=1}^{J}\left(u^{(j)} - \overline{u}\right)\left(G\left(u^{(j)}\right) - \overline{G}\right)^T, \\
\overline{u} &= \frac{1}{J}\sum_{j=1}^{J}u^{(j)}, \quad \overline{G} = \frac{1}{J}\sum_{j=1}^{J}G\left(u^{(j)}\right).
\end{aligned}$$

While this approach offers a practical solution to Bayesian inverse problems, it does not guarantee convergence to μ_n^y as $J \to \infty$ due to inherent approximation errors (see Ernst *et al* 2015). Our focus is on understanding the behaviour of the algorithm with a fixed number of particles, particularly when applied as a derivative-free optimiser to the minimisation problem (2.3).

In discussing EKI, it is important to highlight the invariant subspace property noted in Iglesias *et al* (2013):

Lemma 3.1. *Let $\mathcal{S}$ denote the affine space $\mathcal{S} = \bar{u}_0 + \text{span}\{u_0^{(i)} - \bar{u}_0, \quad i = 1,\dots,J\}$. Then $u_n^{(i)} \in \mathcal{S}$ for all $(n,i) \in \mathbb{N}^+ \times \{1,\dots,J\}$.*

We note that the subspace property allows us to reformulate the possibly infinite-dimensional problem as a finite-dimensional one by restricting the iterations to the finite-dimensional subspace spanned by the initial ensemble. In our analysis, we assume that the parameter space X is finite-dimensional, i.e. $X = \mathbb{R}^n$ for some $n \in \mathbb{N}^+$.

3.1. Continuous-time limit

We investigate the continuous-time limit of EKI as the parameter h in the incremental Bayesian formulation approaches zero. For a detailed and rigorous derivation of these limits, we refer to Blömker *et al* (2018, 2021), Lange (2021). The EKI update step (3.1) can be interpreted as a time-stepping scheme. As $h \to 0$, this formulation converges to a tamed discretisation of the following coupled stochastic differential equations:

$$\frac{\mathrm{d}u^{(i)}}{\mathrm{d}t} = C^{up}(u)\Gamma^{-1}\left(y - G\left(u^{(i)}\right)\right) + C^{up}\left(u^{(i)}\right)\Gamma^{-1}\sqrt{\Gamma}\frac{\mathrm{d}W^{(i)}}{\mathrm{d}t}, \tag{3.2}$$

where the independent cylindrical Brownian motions $W^{(i)}$ on X result from the perturbed observations in the original formulation. From the pre-multiplication by $C^{up}(u)$ on the right-hand side of (3.2), it is clear that the subspace property remains intact in continuous time. This continuous-time formulation serves as the foundation for further analysis regarding the optimal choice of the initial ensemble.

4. Adaptive selection of the initial ensemble

Due to the subspace property (lemma 3.1), the EKI approximations are constrained to the subspace spanned by the initial ensemble. This section, therefore, focuses on strategies for optimising the selection of this subspace.

4.1. Analysis of particle dynamics

In this subsection, we analyse the dynamics of an ensemble moving under deterministic EKI with a linear forward operator. In this case, we are able to derive an explicit solution to equation (3.2) and characterise the behaviour of the ensemble as it evolves over time. In particular, we see how the choice of the initial ensemble affects the accuracy of the EKI in the long-term. This analysis is the starting point for understanding EKI methods in the nonlinear setting.

In addition to assuming that the forward operator, $A : X \to Y$, is linear, we also set the observation perturbations to zero and assume, without loss of generality, that $\Gamma = I$. The dynamics of the particle system $\{u^{(i)}(t)\}_{i=1}^J \in \mathbb{R}^n$ is then governed by the following set of coupled ordinary differential equations (ODEs):

$$\begin{aligned}\frac{\mathrm{d}u^{(i)}(t)}{dt} &= -C_t A^T\left(Au^{(i)}(t) - y\right),\\ C_t &= \frac{1}{J}\sum_{j=1}^{J}\left(u^{(j)} - \bar{u}\right)\left(u^{(j)} - \bar{u}\right)^T,\end{aligned} \tag{4.1}$$

where $C_t \in \mathbb{R}^{n\times n}$ is the empirical covariance matrix of the particle system. It is noteworthy that, in the linear case, EKI can be interpreted as a preconditioned gradient flow (see Schillings and Stuart 2017b).

Bungert and Wacker (2021) provide a comprehensive analysis of this system of ODEs, culminating in explicit expressions for the particle trajectories. Building on this foundation, we derive alternative expressions for the particle trajectories that we believe offer a different perspective on the system's dynamics. The key difference between these perspectives is that the solutions presented in Bungert and Wacker (2021) rely on the diagonal decomposition $C_0A^TA = SDS^{-1}$, whereas our solutions rely on the diagonal decomposition of the symmetric matrix $AC_0A^T =: \mathcal{U}\Sigma\mathcal{U}^T$ or the compact singular value decomposition $A\mathcal{E}_0 =: \sqrt{J}\mathcal{U}\Sigma^{\frac{1}{2}}\mathcal{V}^T$.

We begin by stating the necessary assumptions and notation used throughout our analysis.

Assumption 4.1.

(A1.1) The forward operator $A : X \to Y$ is linear , i.e. we identify the operator with a matrix $A \in \mathbb{R}^{m\times n}$.

(A1.2) The observation is not perturbed during EKI , i.e. there appears no diffusion term in equation (3.2).

(A1.3) Without any loss of generality we assume that $\Gamma = I$ in (3.2).

Notation.

(N1.1) We define the ensemble $U_t := [u^{(1)}(t), \dots, u^{(J)}(t)] \in \mathbb{R}^{n\times J}$, the deviations $\mathcal{E}_t := [u^{(1)}(t) - \bar{u}(t), \dots, u^{(J)}(t) - \bar{u}(t)] \in \mathbb{R}^{n\times J}$ and the empirical covariance $C_t := \frac{1}{J}\mathcal{E}_t\mathcal{E}_t^T = \frac{1}{J}\sum_{j=1}^{J}(u^{(j)}(t) - \bar{u}(t))(u^{(j)}(t) - \bar{u}(t))^T \in \mathbb{R}^{n\times n}$.

(N1.2) The matrix AC_0A^T is of rank $r \leqslant k$, $AC_0A^T = \mathcal{U}\Sigma\mathcal{U}^T$ is an eigen-decomposition and $A\mathcal{E}_0 = \sqrt{J}\mathcal{U}\Sigma^{\frac{1}{2}}\mathcal{V}^T$ is a corresponding compact singular value decomposition ($\mathcal{U} \in \mathbb{R}^{m\times r}$, $\Sigma \in \mathbb{R}^{r\times r}$, $\mathcal{V} \in \mathbb{R}^{J\times r}$, $\Sigma \succ 0$).

Theorem 4.2. *Under assumptions (A1.1)–(A1.3) and with notation (N1.1) and (N1.2), the initial value problem (IVP) for deterministic EKI*

$$\begin{aligned} \frac{\mathrm{d}u^{(i)}(t)}{\mathrm{d}t} &= -C_tA^T\left(Au^{(i)}(t) - y\right), \quad t > 0, \\ u^{(i)}(0) &= u_0^{(i)}, \end{aligned} \tag{4.2}$$

with $C_t := \frac{1}{J}\sum_{j=1}^{J}(u^{(j)}(t) - \bar{u}(t))(u^{(j)}(t) - \bar{u}(t))^T$ *and* $\bar{u}(t) = \frac{1}{J}\sum_{j=1}^{J}u^{(j)}(t)$ *for* $t \geqslant 0$ *has the unique solution given by*

$$\begin{aligned} u^{(i)}(t) &= u^{(i)}(0) + (AC_0)^T\mathcal{U}\Sigma^{-1}\left((I_r + 2\Sigma t)^{-\frac{1}{2}} - I_r\right)\mathcal{U}^T\left(Au^{(i)}(0) - y\right) \\ &= u^{(i)}(0) + \frac{1}{\sqrt{J}}U_0\mathcal{V}\Sigma^{-\frac{1}{2}}\left((I_r + 2\Sigma t)^{-\frac{1}{2}} - I_r\right)\mathcal{U}^T\left(Au^{(i)}(0) - y\right), \quad t \geqslant 0. \end{aligned}$$

Proof. From the ODE for $u^{(i)}(t)$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}C_t = -2C_tA^TAC_t, \quad t > 0.$$

The first equality follows directly from theorem A.1 with $\alpha = 2$. The second equality follows from lemma A.2, which shows that $(AC_0)^T\mathcal{U}\Sigma^{-1} = \frac{1}{\sqrt{J}}\mathcal{E}_0\mathcal{V}\Sigma^{-\frac{1}{2}} = \frac{1}{\sqrt{J}}U_0\mathcal{V}\Sigma^{-\frac{1}{2}}$. □

Theorem 4.2 allows us to quantify the residual in the image space as follows.

Corollary 4.3. *Under assumptions (A1.1)–(A1.3) and with notation (N1.1) and (N1.2), the residuals in the image space are given by*

$$Au^{(i)}(t) - y = \left(I_m - \mathcal{U}\mathcal{U}^T + \mathcal{U}\left(I_r + 2\Sigma t\right)^{-\frac{1}{2}}\mathcal{U}^T\right)\left(Au^{(i)}(0) - y\right),$$

and therefore

$$\lim_{t\to\infty} Au^{(i)}(t) - y = \mathrm{Proj}_{\ker(AC_0A^T)}\left(Au^{(i)}(0) - y\right).$$

The next result characterises the deviations of each particle from the empirical mean.

Corollary 4.4. *Under assumptions (A1.1)–(A1.3) and with notation (N1.1) and (N1.2), the deviations of the particles from their mean are given by*

$$\begin{aligned}\mathcal{E}_t &= \mathcal{E}_0 + (AC_0)^T\mathcal{U}\Sigma^{-1}\left(\left(I_r + 2\Sigma t\right)^{-\frac{1}{2}} - I_r\right)\mathcal{U}^TA\mathcal{E}_0\\ &= \mathcal{E}_0\left(I_J - \mathcal{V}\mathcal{V}^T + \mathcal{V}\left(I_r + 2\Sigma t\right)^{-\frac{1}{2}}\mathcal{V}^T\right),\end{aligned}$$

and so

$$\begin{aligned}\lim_{t\to\infty}\mathcal{E}_t &= \mathcal{E}_0 - (AC_0)^T\mathcal{U}\Sigma^{-1}\mathcal{U}^TA\mathcal{E}_0\\ &= \mathcal{E}_0\left(I_J - \mathcal{V}\mathcal{V}^T\right).\end{aligned}$$

The empirical covariance can then be quantified as follows.

Corollary 4.5. *Under assumptions (A1.1)–(A1.3) and with notation (N1.1) and (N1.2), and with $C_0 = C_LC_R$ being any desired decomposition with $C_L \in \mathbb{R}^{n\times p}$, $C_R \in \mathbb{R}^{p\times n}$ for some $p \in \mathbb{N}^+$, the empirical covariance is given by*

$$\begin{aligned}C_t &= C_0 + C_0A^T\mathcal{U}\Sigma^{-1}\left(\left(I_r + 2\Sigma t\right)^{-1} - I_r\right)\mathcal{U}^TAC_0\\ &= \frac{1}{J}\mathcal{E}_0\left(I_J - \mathcal{V}\mathcal{V}^T + \mathcal{V}\left(I_r + 2\Sigma t\right)^{-1}\mathcal{V}^T\right)\mathcal{E}_0^T\\ &= C_L\left(I_n + 2C_RA^TAC_Lt\right)^{-1}C_R,\end{aligned}$$

and so

$$\begin{aligned}\lim_{t\to\infty}C_t &= C_0 - C_0A^T\mathcal{U}\Sigma^{-1}\mathcal{U}^TAC_0\\ &= \frac{1}{J}\mathcal{E}_0\left(I_J - \mathcal{V}\mathcal{V}^T\right)\mathcal{E}_0^T.\end{aligned}$$

Proof. The first two of the above expressions for C_t are direct corollaries of theorem 4.2. The last expression for C_t is a direct corollary of theorem A.1 after noting that

$$\frac{d}{dt}C_t = -2C_tA^TAC_t.$$

□

Finally, we summarise the results concerning ensemble collapse.

Corollary 4.6 [Ensemble Collapse in State Space]. *Under assumptions (A1.1)–(A1.3) and with notation (N1.1) and (N1.2), and with $C_0 = C_L C_R$ being any desired decomposition with $C_L \in \mathbb{R}^{n\times p}$, $C_R \in \mathbb{R}^{p\times n}$ for some $p \in \mathbb{N}^+$, we have*

$$\begin{aligned} A\mathcal{E}_t &= \mathcal{U}\left(I_r + 2\Sigma t\right)^{-\frac{1}{2}} \mathcal{U}^T A\mathcal{E}_0 \\ &= \sqrt{J}\mathcal{U}\Sigma^{\frac{1}{2}}\left(I_r + 2\Sigma t\right)^{-\frac{1}{2}} \mathcal{V}^T, \end{aligned}$$

and thus

$$\lim_{t\to\infty} A\mathcal{E}_t = 0.$$

Similarly,

$$\begin{aligned} AC_t &= \mathcal{U}\left(I_r + 2\Sigma t\right)^{-1} \mathcal{U}^T AC_0 \\ &= \frac{1}{\sqrt{J}}\mathcal{U}\Sigma^{\frac{1}{2}}\left(I_r + 2\Sigma t\right)^{-1} \mathcal{V}^T \mathcal{E}_0^T \\ &= AC_L\left(I_n + 2C_R A^T AC_L t\right)^{-1} C_R, \end{aligned}$$

and thus

$$\lim_{t\to\infty} AC_t = 0.$$

4.2. Optimal initial ensemble

The subspace property (lemma 3.1) of EKI ensures that the particles remain within the span of the initial ensemble throughout the inversion process. This property can be used to regularise the inverse problem, especially when the problem is ill-posed and cannot feasibly be inverted across the entire space. The initial ensemble is therefore a critical design parameter that should be carefully chosen to ensure accurate and computationally feasible results. Typically, prior knowledge is used to inform the choice of initial ensemble.

We illustrate this concept using the linear Gaussian case, where the forward operator is linear, and both the prior and noise are Gaussian. The optimisation problem of finding the MAP estimate—i.e. choosing u to maximise the posterior probability—is equivalent to the following minimisation problem:

$$\min_{u\in\mathbb{R}^n} \frac{1}{2}|Au - y|^2 + \frac{1}{2}|u|_R^2 = \min_{u\in\mathbb{R}^n} \frac{1}{2}\left|\begin{pmatrix} Au \\ R^{-\frac{1}{2}}u \end{pmatrix} - \begin{pmatrix} y \\ 0 \end{pmatrix}\right|^2. \tag{4.3}$$

Typically, the dominant eigenvectors of a given prior are used to construct the smaller parameter subspace to which the inversion process is confined. Specifically, one selects the first J terms in the Karhunen–Loève expansion of the Gaussian prior. If the covariance matrix R of the prior has a diagonal decomposition $R := V\Lambda V^T$, where $\Lambda := \mathrm{diag}(\{\lambda_{(i)}\}_i) \in \mathbb{R}^{n\times n}$ contains the eigenvalues of R, and $V := [v_{(1)}, \ldots, v_{(n)}] \in \mathbb{R}^{n\times n}$ consists of the corresponding eigenvectors, then the truncated Karhunen-Loève expansion is given by

$$\sum_{i\in\mathcal{J}} \xi_i \sqrt{\lambda_{(i)}} v_{(i)} \sim \mathcal{N}(0, R),$$

where $\xi_i \sim \mathcal{N}(0,1)$ i.i.d. and $\mathcal{J} \subseteq \{1, \ldots, n\}$ indexes the J largest eigenvalues of R. The initial ensemble is then taken to be

$$U_0 = \left[u^{(1)}(0), \ldots, u^{(J)}(0)\right] := \left[v_{(\mathcal{J}_1)}, \ldots, v_{(\mathcal{J}_J)}\right] \mathrm{diag}\left(\left\{\sqrt{\lambda_{(i)}}\right\}_{i\in\mathcal{J}}\right) \mathrm{diag}\left(\{\xi_i\}_{i\leqslant J}\right) =: V_{\mathcal{J}} \Lambda_{\mathcal{J}}^{\frac{1}{2}} \Xi_{\mathcal{J}}.$$

In this approach, one then seeks to minimise the objective function in equation (4.3) by running EKI with forward operator $\begin{pmatrix}(Au)^T & (R^{-\frac{1}{2}}u)^T\end{pmatrix}^T$, data $\begin{pmatrix}y^T & 0^T\end{pmatrix}^T$ and initial ensemble $U_0 = V_{\mathcal{J}}\Lambda_{\mathcal{J}}^{\frac{1}{2}}\Xi_{\mathcal{J}}$.

We seek to generalise this approach by choosing among any desired set of J eigenvectors of R and by allowing the initial ensemble to consist of any desired linearly independent linear combination of these eigenvectors. Specifically, we define

$$U_{\mathcal{J},B}(0) = \left[u_{\mathcal{J},B}^{(1)}(0),\dots,u_{\mathcal{J},B}^{(J)}(0)\right] := [v_{\mathcal{J}_1},\dots,v_{\mathcal{J}_J}]B =: V_{\mathcal{J}}B,$$

where $\mathcal{J}\subseteq\{1,\dots,n\}$ is a subset of size J and $B\in\mathbb{R}^{J\times J}$ is an invertible matrix. Our objective is then to select $\mathcal{J}$ and B to optimise the long-term performance of EKI.

We know from corollary 4.3 that $\forall i : Au_{\mathcal{J},B}^{(i)}(t)$ converge to the same value. Our goal, therefore, is to choose $\mathcal{J}\subseteq\{1,\dots,n\}$ and B to minimise

$$\lim_{t\to\infty}\frac{1}{2}|A\bar{u}_{\mathcal{J},B}(t)-y|^2+\frac{1}{2}|\bar{u}_{\mathcal{J},B}(t)|_R^2.$$

In fact, to develop an EKI algorithm with adaptive resampling (see section 4.3), we later need to choose $\mathcal{J}\subseteq\{1,\dots,n\}$ and B to minimise the more general objective function

$$\lim_{t\to\infty}\frac{1}{2}|A\bar{u}_{\mathcal{J},B}(t)-y|^2+\frac{1}{2}|\bar{u}_{\mathcal{J},B}(t)-\mu|_R^2.$$

For any desired $\mu\in\mathbb{R}^n$. To this end we have the following theorem.

Theorem 4.7. *Suppose that assumptions (A1.1)–(A1.3) hold. Further, let $R\in\mathbb{R}^{n\times n}$ be symmetric and positive definite with an eigen-decomposition $R = V\Lambda V^T\in\mathbb{R}^{n\times n}$, where $\Lambda := diag(\{\lambda_{(i)}\}_i)\in\mathbb{R}^{n\times n}$ and $V := [v_1,\dots,v_n]\in\mathbb{R}^{n\times n}$. We denote by $\mathcal{J}\subseteq\{1,\dots,n\}$ an index set of size J and $\mathcal{J}^c := \{1,\dots,n\}\backslash\mathcal{J}$, and set $V_{\mathcal{J}} := [v_{\mathcal{J}_1},\dots,v_{\mathcal{J}_J}]\in\mathbb{R}^{n\times J}$ and $\Lambda_{\mathcal{J}} := diag(\{\lambda_{\mathcal{J}_1},\dots,\lambda_{\mathcal{J}_J}\})$.*

For an invertible matrix $B\in\mathbb{R}^{J\times J}$, we denote by $\{u_{\mathcal{J},B}^{(i)}(t)\}_{i=1}^J\in\mathbb{R}^n$ the particles moving under deterministic EKI with the linear forward operator $\begin{pmatrix}A^T & R^{-\frac{1}{2}}\end{pmatrix}^T\in\mathbb{R}^{(m+n)\times n}$ and data $\begin{pmatrix}y^T & (R^{-\frac{1}{2}}\mu)^T\end{pmatrix}^T\in\mathbb{R}^{m+n}$ and with starting position $[u_{\mathcal{J},B}^{(1)}(0),\dots,u_{\mathcal{J},B}^{(J)}(0)] := V_{\mathcal{J}}B\in\mathbb{R}^{n\times J}$, i.e. following the dynamics given by ODEs (4.2).

Let $1_J\in\mathbb{R}^J$ be a vector of ones, $A_{\mathcal{J}} := AV_{\mathcal{J}}\in\mathbb{R}^{m\times J}$, $Q_{\mathcal{J}} := (I_m + A_{\mathcal{J}}\Lambda_{\mathcal{J}}A_{\mathcal{J}}^T)^{-1}\in\mathbb{R}^{m\times m}$, $M_{\mathcal{J}} := (\Lambda_{\mathcal{J}}^{-1}+A_{\mathcal{J}}^TA_{\mathcal{J}})^{-1}\in\mathbb{R}^{J\times J}$, $R_{\mathcal{J}}^\dagger := V_{\mathcal{J}}\Lambda_{\mathcal{J}}^{-1}V_{\mathcal{J}}$, $R_{\mathcal{J}^c}^\dagger := V_{\mathcal{J}^c}\Lambda_{\mathcal{J}^c}^{-1}V_{\mathcal{J}^c}\in\mathbb{R}^{m\times m}$ and $z_{\mathcal{J}} := y - AV_{\mathcal{J}}V_{\mathcal{J}}^T\mu\in\mathbb{R}^m$.

Then $u_{\mathcal{J},B}^{(i)}(t)$ converge for $i = 1,\dots,J$ and

$$\begin{aligned}
&\lim_{t\to\infty}|A\bar{u}_{\mathcal{J},B}(t)-y|^2+|\bar{u}_{\mathcal{J},B}(t)-\mu|_R^2\\
&\quad= z_{\mathcal{J}}^TQ_{\mathcal{J}}z_{\mathcal{J}}+\frac{\left(1-1_J^TB^{-1}\left(V_{\mathcal{J}}^T\mu+\Lambda_{\mathcal{J}}A_{\mathcal{J}}^TQ_{\mathcal{J}}z_{\mathcal{J}}\right)\right)^2}{1_J^TB^{-1}\Lambda_{\mathcal{J}}B^{-T}1_J-\left(A_{\mathcal{J}}\Lambda_{\mathcal{J}}B^{-T}1_J\right)^TQ_{\mathcal{J}}A_{\mathcal{J}}\Lambda_{\mathcal{J}}B^{-T}1_J}+\mu^TR_{\mathcal{J}^c}^\dagger\mu\\
&\quad= z_{\mathcal{J}}^Tz_{\mathcal{J}}-z_{\mathcal{J}}^TA_{\mathcal{J}}M_{\mathcal{J}}A_{\mathcal{J}}^Tz_{\mathcal{J}}+\frac{\left(1-1_J^TB^{-1}M_{\mathcal{J}}V_{\mathcal{J}}^T\left(A^Ty+R_{\mathcal{J}}^\dagger\mu\right)\right)^2}{1_J^TB^{-1}M_{\mathcal{J}}B^{-T}1_J}+\mu^TR_{\mathcal{J}^c}^\dagger\mu.
\end{aligned}$$

The proof of the theorem can be found in appendix B.

Corollary 4.8. *Under the assumptions of theorem 4.7, the following inequality holds:*

$$\lim_{t\to\infty}|A\bar{u}_{\mathcal{J},B}(t)-y|^2+|\bar{u}_{\mathcal{J},B}(t)-\mu|_R^2\geqslant z_{\mathcal{J}}^Tz_{\mathcal{J}}-z_{\mathcal{J}}^TA_{\mathcal{J}}M_{\mathcal{J}}A_{\mathcal{J}}^Tz_{\mathcal{J}}+\mu^TR_{\mathcal{J}^c}^\dagger\mu,$$

where the lower bound is attained if and only if

$$1_J^T B^{-1} M_{\mathcal{J}} V_{\mathcal{J}}^T \left(A^T y + R_{\mathcal{J}}^\dagger \mu \right) = 1. \tag{4.4}$$

This optimality condition for the matrix B allows us to characterise the EKI estimate as the best approximation within the given subspace.

Corollary 4.9. *Under the assumptions of theorem 4.7, and assuming that $B \in \mathbb{R}^{J\times J}$ satisfies the optimality condition given in corollary 4.8, the following holds:*

$$\lim_{t\to\infty} u^{(i)}(t) = \arg\min_{u\in \mathrm{Im}(V_{\mathcal{J}})} \frac{1}{2}|Au-y|^2 + \frac{1}{2}|u-\mu|_R^2 .$$

4.2.1. Optimal linear combination. For any invertible $D \in \mathbb{R}^{J\times J}$ such that $1_J^T D^{-1} M_{\mathcal{J}} V_{\mathcal{J}}^T (A^T y + R_{\mathcal{J}}^\dagger \mu) \neq 0$,

$B := (1_J^T D^{-1} M_{\mathcal{J}} V_{\mathcal{J}}^T (A^T y + R_{\mathcal{J}}^\dagger \mu))^{-1} D$ satisfies equation (4.4). As such, there is a lot of flexibility in constructing an optimal B.

Given this flexibility in choosing an optimal B—i.e. a B that minimises the final value of the objective function in equation (4.2) - we can restrict our attention to those optimal B that further minimise the initial value of the objective function.

Corollary 4.10. *Let $B^* \in \mathbb{R}^{J\times J}$ be such that*

$$B^* 1_J/J = M_{\mathcal{J}} V_{\mathcal{J}}^T \left(A^T y + R_{\mathcal{J}}^\dagger \mu \right).$$

Then B^* satisfies equation (4.4) and further

$$\forall t : |A\bar{u}_{\mathcal{J},B^*}(t) - y|^2 + |\bar{u}_{\mathcal{J},B^*}(t) - \mu|_R^2 = \min_{u\in Im(V_{\mathcal{J}})} |Au-y|^2 + |u-\mu|_R^2 .$$

The proof of this corollary can be found in appendix B.

Corollary 4.8 shows that, for any choice of B satisfying the optimality condition given by equation (4.4), $\bar{u}_{\mathcal{J},B}^{(i)}(t)$ converges to the unique minimiser of the Tikhonov-regularised objective function on the subspace $\mathrm{Im}(V_{\mathcal{J}})$. Corollary 4.10 shows that a B satisfying the optimality condition given by equation (4.4) can be chosen such that $\forall t : \bar{u}_{\mathcal{J},B}^{(i)}(t)$ is the unique minimiser of the Tikhonov-regularised objective function on the subspace $Im(V_{\mathcal{J}})$.

4.2.2. Optimal index set. Having established methods for generating an optimal B given a choice of eigenvectors $\mathcal{J}$, we now seek to choose a $\mathcal{J}$ that minimises the remaining term in theorem 4.7 , i.e. the lower bound in corollary 4.8:

$$z_{\mathcal{J}}^T z_{\mathcal{J}} - z_{\mathcal{J}}^T A_{\mathcal{J}} \mathcal{M}_{\mathcal{J}} A_{\mathcal{J}}^T z_{\mathcal{J}} + \mu^T R_{\mathcal{J}^c}^\dagger \mu = \min_{u\in \mathrm{Im}(V_{\mathcal{J}})} \frac{1}{2}|Au-y|^2 + \frac{1}{2}|u-\mu|_R^2. \tag{4.5}$$

Given that there are $\binom{n}{J}$ possible choices for $\mathcal{J}$, minimising the objective function (4.5) by brute force quickly becomes computationally intractable for large n. We instead seek to iteratively build up the index set $\mathcal{J}$, i.e. $\mathcal{J} := \mathcal{J}_J$ and for k < J : $\mathcal{J}_k \subseteq \{1,\dots,n\}$ with $|\mathcal{J}_k| = k$ and $\mathcal{J}_{k+1} := \mathcal{J}_k \cup \{j^*\}$ for some $j^* \in \{1,\dots,n\}\backslash\mathcal{J}_k$. Recalling the definitions of $z_{\mathcal{J}}$ and $|\mu|_{\mathcal{J}^c}^2$, it is clear that these are simple to calculate iteratively ($z_{\mathcal{J}_{k+1}} = z_{\mathcal{J}_k} - A v_{j^*} v_{j^*}^T \mu$ and $\mu^T R_{\mathcal{J}_{k+1}^c}^\dagger \mu = \mu^T R_{\mathcal{J}_k^c}^\dagger \mu - \frac{1}{\lambda_{j^*}} \langle \mu, v_{j^*} \rangle^2$).

Algorithm Greedy.. Greedy algorithm for selecting eigenvectors.

Initialise:
$k \leftarrow 1$
$p_i \leftarrow \lambda_i^{\frac{1}{2}} A v_i \in \mathbb{R}^m, \quad i = 1, \dots, n$
$\mathcal{J} \leftarrow \{\}$
$\mathcal{J}^c \leftarrow \{1, \dots, n\}$
$P \leftarrow []$
$\tilde{M} \leftarrow []$
$z \leftarrow y$
while k < J **do**
  $k \leftarrow k+1$
  min. val. $\leftarrow +\infty$
  for $j \in \mathcal{J}^c$ **do**
    $\tilde{z} = z - (\mu^T v_j) A v_j$
    val. $\leftarrow -2(\mu^T v_j)(z^T A v_j) + (\mu^T v_j)^2 (|Av_j|^2 - \frac{1}{\lambda_j}) - \tilde{z}^T P \tilde{M} P^T \tilde{z} - \frac{(\tilde{z}^T P \tilde{M} P^T p_i - \tilde{z}^T p_i)^2}{1 + p_i^T p_i - p_i^T P \tilde{M} P^T p_i}$
    if val. < min. val. **then**
      min. val. $\leftarrow$ val.
      $j^* \leftarrow j$
    end if
  end for
  $\mathcal{J} \leftarrow \mathcal{J} \cup \{j^*\}$
  $\mathcal{J}^c \leftarrow \mathcal{J}^c \backslash \{j^*\}$
  $\Delta \leftarrow 1 + p_{j^*}^T p_{j^*} - p_{j^*}^T P \tilde{M} P^T p_{j^*}$
  $\tilde{M} \leftarrow \frac{1}{\Delta} \begin{pmatrix} \Delta \tilde{M} + (\tilde{M} P^T p_{j^*})(\tilde{M} P^T p_{j^*})^T & -\tilde{M} P^T p_{j^*} \\ -p_{j^*}^T P \tilde{M} & 1 \end{pmatrix} \in \mathbb{R}^{k \times k}$
  $P \leftarrow \begin{bmatrix} P & p_{j^*} \end{bmatrix} \in \mathbb{R}^{m \times k}$
  $z \leftarrow z - \langle \mu, v_{j^*} \rangle A v_{j^*}$
end while
$M_{\mathcal{J}} \leftarrow \Lambda_{\mathcal{J}}^{\frac{1}{2}} \tilde{M} \Lambda_{\mathcal{J}}^{\frac{1}{2}}$

It remains only to iteratively calculate $z_{\mathcal{J}}^T A_{\mathcal{J}} \mathcal{M}_{\mathcal{J}} A_{\mathcal{J}}^T z_{\mathcal{J}}$. If we define $P_{\mathcal{J}} := A V_{\mathcal{J}} \Lambda_{\mathcal{J}}^{\frac{1}{2}}$ and $\tilde{M}_{\mathcal{J}} := (I + P_{\mathcal{J}}^T P_{\mathcal{J}})^{-1}$, then

$$\begin{aligned} A_{\mathcal{J}} M_{\mathcal{J}} A_{\mathcal{J}}^T &= A V_{\mathcal{J}} \left(\Lambda_{\mathcal{J}}^{-1} + V_{\mathcal{J}}^T A^T A V_{\mathcal{J}} \right)^{-1} V_{\mathcal{J}}^T A^T \\ &=: P_{\mathcal{J}} \tilde{M}_{\mathcal{J}} P_{\mathcal{J}}^T. \end{aligned}$$

In this form, we can make use of the following lemma.

Lemma 4.11. *Let $x \in \mathbb{R}^m$, $P_k := [p_1, \dots, p_k] \in \mathbb{R}^{m \times k}$, and $\tilde{M}_k := (I_k + P_k^T P_k)^{-1} \in \mathbb{R}^{k \times k}$. Define $\Delta_{k+1} := 1 + p_{k+1}^T p_{k+1} - p_{k+1}^T P_k \tilde{M}_k P_k^T p_{k+1}$. Then*

$$\tilde{M}_{k+1} = \frac{1}{\Delta_{k+1}} \begin{pmatrix} \Delta_{k+1} \tilde{M}_k + \left(\tilde{M}_k P_k^T p_{k+1} \right) \left(\tilde{M}_k P_k^T p_{k+1} \right)^T & -\tilde{M}_k P_k^T p_{k+1} \\ -p_{k+1} P_k \tilde{M}_k & 1 \end{pmatrix}.$$

$$x^T P_{k+1} \tilde{M}_{k+1} P_{k+1}^T x = x^T P_k \tilde{M}_k P_k^T x + \frac{1}{\Delta_{k+1}} \left(x^T P_k \tilde{M}_k P_k^T p_{k+1} - x^T p_{k+1} \right)^2.$$

The proof of this lemma can be found in appendix B.

Using this lemma, we can iteratively calculate the objective function as follows. Let $\mathcal{J}_k \subseteq \{1,\dots,n\}$ with $|\mathcal{J}_k| = k$ and $k < J$. Define $p_j := \lambda_j^{\frac{1}{2}} A v_j$, $P_k := [p_{\mathcal{J}_1},\dots,p_{\mathcal{J}_k}] \in \mathbb{R}^{m\times k}$, $\tilde{M}_k := (I_k + P_k^T P_k)^{-1} \in \mathbb{R}^{k\times k}$, $V_k = [v_{\mathcal{J}_1},\dots,v_{\mathcal{J}_k}] \in \mathbb{R}^{n\times k}$, $\Lambda_k = \operatorname{diag}(\lambda_{\mathcal{J}_1},\dots,\lambda_{\mathcal{J}_k}) \in \mathbb{R}^{k\times k}$. Then

$$\begin{aligned}
&\left|z_{\mathcal{J}_{k+1}}\right|^2 + \mu^T R^\dagger_{\mathcal{J}^c_{k+1}}\mu - z^T_{\mathcal{J}_{k+1}} A_{\mathcal{J}_{k+1}} M_{\mathcal{J}_{k+1}} A^T_{\mathcal{J}_{k+1}} z_{\mathcal{J}_{k+1}} \\
&\qquad = \left|z_{\mathcal{J}_{k+1}}\right|^2 + \mu^T R^\dagger_{\mathcal{J}^c_{k+1}}\mu - z^T_{\mathcal{J}_{k+1}} P_{\mathcal{J}_{k+1}} \tilde{M}_{\mathcal{J}_{k+1}} P^T_{\mathcal{J}_{k+1}} z_{\mathcal{J}_{k+1}} \\
&\qquad = \left|z_{\mathcal{J}_{k+1}}\right|^2 + \mu^T R^\dagger_{\mathcal{J}^c_{k+1}}\mu - z^T_{\mathcal{J}_{k+1}} P_{\mathcal{J}_k} \tilde{M}_{\mathcal{J}_k} P^T_{\mathcal{J}_k} z_{\mathcal{J}_{k+1}} \\
&\qquad\quad - \frac{\left(z^T_{\mathcal{J}_{k+1}} P_{\mathcal{J}_k} \tilde{M}_{\mathcal{J}_k} P^T_{\mathcal{J}_k} p_{\mathcal{J}_{k+1}} - z^T_{\mathcal{J}_{k+1}} p_{\mathcal{J}_{k+1}}\right)^2}{1 + p^T_{k+1} p_{k+1} - p^T_{k+1} P_{\mathcal{J}_k} \tilde{M}_{\mathcal{J}_k} P^T_{\mathcal{J}_k} p_{k+1}}.
\end{aligned}$$

This leads naturally to the algorithm Greedy, a greedy algorithm for iteratively selecting indices for the index set $\mathcal{J}$ as defined in theorem 4.7.

This algorithm does not, in general, select an optimal subset $\mathcal{J}$. We do, however, show a sufficient condition on the forward operator A that ensures that algorithm Greedy selects an optimal $\mathcal{J}$.

Proposition 4.12. *Under the definitions and assumptions of theorem 4.7, if $R = V\Lambda V^T$ and A has singular value decomposition given by $A = \mathcal{U}\Sigma V^T$, then algorithm Greedy finds an optimal subset $\mathcal{J}$, i.e. a $\mathcal{J}$ that minimises $z^T_{\mathcal{J}} z_{\mathcal{J}} - z^T_{\mathcal{J}} A_{\mathcal{J}} \mathcal{M}_{\mathcal{J}} A^T_{\mathcal{J}} z_{\mathcal{J}} + \mu^T R^\dagger_{\mathcal{J}^c}\mu$.*

The proof of the proposition can be found in appendix B.

4.3. Nonlinear forward operator and resampling

We now generalise our approach to allow resampling of the ensemble and to handle nonlinear forward operators. Suppose we wish to use EKI to generate a set of particle trajectories $\{u_t^{(i)}\}_{i=1}^J$ to minimise the objective function

$$\frac{1}{2}\left|G(u) - y\right|^2 + \frac{1}{2}\left|u\right|^2_R. \tag{4.6}$$

Suppose at t_0 we wish to resample the J particles. We write $\{u^{(i)}_{t_0^-}\}_{i=1}^J$ for the particle positions at time t_0 before resampling and $\{u^{(i)}_{t_0^+}\}_{i=1}^J$ for the particle positions at time t_0 after resampling. One approach to resampling would be to linearise about $\bar{u}_{t_0^-}$ in order to apply our ensemble selection results for linear forward operators , i.e.

$$\begin{aligned}
\frac{1}{2}\left|G(\bar{u}_{t_0} + \tilde{u}) - y\right|^2 + \frac{1}{2}\left|\bar{u}_{t_0} + \tilde{u}\right|^2_R &\approx \frac{1}{2}\left|\nabla G\left(\bar{u}_{t_0^-}\right)\tilde{u} - \left(y - G\left(\bar{u}_{t_0^-}\right)\right)\right|^2 + \frac{1}{2}\left|\bar{u}_{t_0^-} + \tilde{u}\right|^2_R \\
&= \frac{1}{2}\left|A\tilde{u} - \tilde{y}\right|^2 + \frac{1}{2}\left|\tilde{u} - \mu\right|^2_R,
\end{aligned} \tag{4.7}$$

where $A := \nabla G(\bar{u}_{t_0^-})$, $\tilde{y} := y - G(\bar{u}_{t_0^-})$ and $\mu := -\bar{u}_{t_0^-}$. With these defined parameters, we can use algorithm Greedy and corollary 4.8 to generate $\mathcal{J}$, B and set $u^{(i)}_{t_0^+} := \bar{u}_{t_0^-} + \tilde{u}^{(i)}_{t_0} := \bar{u}_{t_0^-} + V_{\mathcal{J}} B e_i$ before continuing to run EKI.

This is a heuristic approach, which linearises a general forward operator in order to apply the linear results. Since we linearise around the point $\bar{u}_{t_0^-}$, we may want the resampled points

Algorithm AdaptSelect. Adaptive ensemble selection.

Initialise:
μ e.g. $\mu \leftarrow 0$
$0 = t_0 < \ldots < t_N < t_{N+1}$
for k = 1:N **do**
  $A \leftarrow \nabla G(-\mu)$
  $\tilde{y} \leftarrow y - G(-\mu)$
  $\mathcal{J}, M_{\mathcal{J}} \leftarrow$ *algorithm* Greedy $(A, \tilde{y}, \mu)$
  $\tilde{z} \leftarrow M_{\mathcal{J}} V_{\mathcal{J}}^T (A^T \tilde{y} + R_{\mathcal{J}}^\dagger \mu)$
  $B \leftarrow \sqrt{J}\,|\tilde{z}|_2\, Householder(1_J/\sqrt{J},\ \ \tilde{z}/|\tilde{z}|)$
  $[u^{(1)}, \ldots, u^{(J)}] \leftarrow -[\mu, \ldots, \mu] + V_{\mathcal{J}} B$
  $\{u^{(i)}\}_{i=1}^J \leftarrow EKI(t_{k+1} - t_k, \{u^{(i)}\}_{i=1}^J, G, R, y)$
  $\mu \leftarrow -\frac{1}{J}\sum_{i=1}^J u^{(i)}$
end for

$\{u_{t_0^+}^{(i)}\}_{i=1}^J$ to remain close, so that the linearisation is a good approximation. The total (squared) distance to $\bar{u}_{t_0^-}$ is given by

$$\sum_{i=1}^J \left| u_{t_0^+}^{(i)} - \bar{u}_{t_0^-}^{(i)} \right|^2 = \sum_{i=1}^J |V_{\mathcal{J}} B e_i|^2 = \text{trace}\left(B^T B\right).$$

As such, we seek a B that solves the following minimisation problem:

$$\min \text{trace}\left(B^T B\right) : 1 = 1_J^T B^{-1} M_{\mathcal{J}} V_{\mathcal{J}}^T \left(A^T \tilde{y} + R_{\mathcal{J}}^\dagger \mu\right).$$

To this end, lemma A.5 provides an approximate solution to this minimisation problem. Namely, let $\tilde{z} := M_{\mathcal{J}} V_{\mathcal{J}}^T (A^T \tilde{y} + R_{\mathcal{J}}^\dagger \mu)$ and let $U \in \mathbb{R}^{J \times J}$ be an orthogonal matrix such that $U 1_J/\sqrt{J} = \tilde{z}/|\tilde{z}|$. Then

$$B := \sqrt{J}\,|\tilde{z}|\, U.$$

In this case we have that

$$\text{trace}\left(B^T B\right) = J^2 |\tilde{z}|^2.$$

Alternatively, we can choose B to instead approximately minimise the total (squared) distance to $\bar{u}_{t_0^-}$ in projection space, i.e.

$$\sum_{i=1}^J \left| \begin{pmatrix} G\left(u_{t_0^+}^{(i)}\right) \\ R^{-\frac{1}{2}} u_{t_0^+}^{(i)} \end{pmatrix} - \begin{pmatrix} G\left(u_{t_0^-}^{(i)}\right) \\ R^{-\frac{1}{2}} u_{t_0^-}^{(i)} \end{pmatrix} \right|^2 \approx \sum_{i=1}^J \left| \begin{pmatrix} \nabla G\left(\bar{u}_{t_0}^{(i)}\right) V_{\mathcal{J}} B \\ R^{-\frac{1}{2}} V_{\mathcal{J}} B \end{pmatrix} \right|^2 = \text{trace}\left(B^T M_{\mathcal{J}}^{-1} B\right).$$

This leads to the following minimisation problem

$$\min \text{trace}\left(B^T \mathcal{M}_{\mathcal{J}}^{-1} B\right) : 1 = 1_J^T B^{-1} M_{\mathcal{J}} V_{\mathcal{J}}^T \left(A^T \tilde{y} + R_{\mathcal{J}}^\dagger \mu\right).$$

Again, lemma A.5 provides an approximate solution to this minimisation problem. Let $U \in \mathbb{R}^{J \times J}$ be an orthogonal matrix such that $U 1_J/\sqrt{J} = M_{\mathcal{J}}^{-\frac{1}{2}} \tilde{z} / \left| M_{\mathcal{J}}^{-\frac{1}{2}} \tilde{z} \right|$. Then

$$B := \sqrt{J} \left| M_{\mathcal{J}}^{-\frac{1}{2}} \tilde{z} \right| M_{\mathcal{J}}^{\frac{1}{2}} U.$$

In this case we have that

$$\text{trace}\left(B^T\mathcal{M}_{\mathcal{J}}^{-1}B\right) = J^2\left|M_{\mathcal{J}}^{\frac{1}{2}}\tilde{z}\right|^2.$$

In either cases, it is clear that these Bs chosen to minimise the various traces are optimal both in the sense of corollary 4.8 and the stronger sense of corollary 4.10.

These results lead to algorithm AdaptSelect for performing EKI for $t \in [0, t_{N+1}]$ with resampling at times $t_1 < \ldots < t_N$.

To summarise algorithm AdaptSelect, at each resample time t_k, the Tikhonov-regularised objective function (4.6) is linearised around the point $\bar{u}^{(i)}_{\mathcal{J},B}(t_k^-)$, the prior eigenvectors $\mathcal{J}_k$ are chosen by algorithm Greedy to approximately minimise the minimum possible value of the linearised objective function on the subspace $\text{Im}(V_{\mathcal{J}_k})$ and, finally, B is chosen such that $\bar{u}^{(i)}_{\mathcal{J},B}(t_k^+)$ achieves this minimum value. To stay in a region where the linearisation is a good approximation of the original objective function, B is chosen such that the resampled ensemble is close to the ensemble immediately before resampling.

Algorithm AdaptSelect is a largely derivative-free method, only requiring derivative calculations at initialisation and at resample times. For problems in which derivative calculations are expensive, algorithm AdaptSelect with minimal resampling could therefore be significantly more efficient computationally than standard optimisation methods that rely heavily on derivative calculations. In fact, if we were to use a derivative-free linearisation in equation (4.7), algorithm AdaptSelect itself would become a completely derivative-free method.

5. Numerical experiments

In this subsection, we numerically test the proposed EKI methods for linear and nonlinear problems.

5.1. Set-up

We run our experiments according to the following generalise scheme. We first generate the true solution u_{true} and define the observation data as $y := G(u_{\text{true}}) + \eta$, where η represents observational noise. Using the same data, we execute variants of algorithm AdaptSelect to minimise the objective function

$$\Phi(u) := \frac{1}{2}|G(u) - y|^2 + \frac{1}{2}|u|_R^2. \tag{5.1}$$

The variants of algorithm AdaptSelect differ in their (re-)sampling times $0 = t_0 < \ldots < t_N$, in their methods for selecting $\mathcal{J}$ and in their methods for constructing B. In our numerical experiments, we investigate the following variants:

- **Greedy_opt**: at $t = t_0$, $\mathcal{J}$ is selected according to algorithm Greedy and B is constructed according to corollary 4.10.
- **Dom_opt**: $\mathcal{J}$ consists of the indices of the dominant eigenvalues and B is constructed according to corollary 4.10.
- **Greedy_kl**: at $t = t_0$, $\mathcal{J}$ is selected according to algorithm Greedy and $B = \Lambda_{\mathcal{J}}^{\frac{1}{2}}\Xi_{\mathcal{J}} := \Lambda_{\mathcal{J}}^{\frac{1}{2}}\text{diag}(\{\xi_i\}_{i\leqslant J})$, where $\xi_i \sim \mathcal{N}(0,1)$.
- **Dom_kl**: at $t = t_0$, $\mathcal{J}$ is chosen to be the indices of the dominant eigenvalues and $B = \Lambda_{\mathcal{J}}^{\frac{1}{2}}\Xi_{\mathcal{J}} := \Lambda_{\mathcal{J}}^{\frac{1}{2}}\text{diag}(\{\xi_i\}_{i\leqslant J})$, where $\xi_i \sim \mathcal{N}(0,1)$, i.e. the standard Karhunen–Loéve initialisation.

- **Greedy_opt_r**: at times $t = t_0, \ldots, t_N$, $\mathcal{J}$ is selected according to algorithm Greedy and B is constructed according to corollary 4.10.
- **Dom_opt_r**: at times $t = t_0, \ldots, t_N$, $\mathcal{J}$ consists of the indices of the dominant eigenvalues, while B is constructed according to corollary 4.10. This variant is included to evaluate.
- **Greedy_kl_r**: times $t = t_0, \ldots, t_N$, $\mathcal{J}$ is selected according to algorithm Greedy and $B = \Lambda_{\mathcal{J}}^{\frac{1}{2}} \Xi_{\mathcal{J}} := \Lambda_{\mathcal{J}}^{\frac{1}{2}} \mathrm{diag}(\{\xi_i\}_{i \leqslant J})$, where $\xi_i \sim \mathcal{N}(0,1)$.
- **Random** (rand): at $t = t_0$, $\mathcal{J}$ is chosen randomly and B is constructed according to corollary 4.10. This variant is included as a baseline to evaluate, by way of comparison, the effect of different methods for selecting $\mathcal{J}$.
- **Best**: this is only included if both the forward operator is linear and $\binom{n}{J}$ is small. For every possible index set $\mathcal{J} \subseteq \{1, \ldots, n\}$ of size J, B is constructed according to corollary 4.10 and the resulting long-term value of the objective function is calculated. $\mathcal{J}$ and B are then taken to be the best performing pair of index set and linear combination. This variant is included to evaluate, by way of comparison, the performance of different methods for selecting $\mathcal{J}$.

We also investigate the performance of the different variants of the algorithm across various values for the following parameters:

- **J** : the number of particles.
- $\boldsymbol{\beta}$: a scalar that determines the relative weight of the prior/regularisation term in the objective function given in equation (5.1).

The value of the objective function in equation (5.1) at the final time t_{N+1} for each variant is denoted by an r with a superscript indicating the variant, e.g. $r^{\text{greedy\_opt}} := \Phi(\bar{u}_{t_{N+1}}^{\text{greedy\_opt}})$.

For comparison, we also define $r^{\min} := \min_u \Phi(u)$. In the case that the forward operator is linear, $r^{\min}$ is explicitly calculated; otherwise, it is be approximated using a numerical optimiser (more specifically, the fmincon MATLAB optimiser). We present our results as a ratio, e.g. $\frac{r^{\min}}{r^{\text{greedy\_opt}}}$.

For a given variant of the algorithm and a given parameter configuration, we evaluate its performance across a range of random data points, and potentially across a range of forward operators and/or covariance matrices. As such, for each parameter configuration, we run n_{exp} experiments, in each of which we generate a random data point, and potentially a random forward operator and/or covariance matrix, and run each variant of the EKI method on this given data point. The results of these experiments are indexed by $1 \leqslant i \leqslant n_{\text{exp}}$, e.g. $\frac{r_i^{\min}}{r_i^{\text{greedy\_opt}}}$ for $i = 1, \ldots, n_{\text{exp}}$. We then average across these experiments, e.g. $\frac{r^{\min}}{r^{\text{greedy\_opt}}} := \frac{1}{n_{\text{exp}}} \sum_{i=1}^{n_{\text{exp}}} \frac{r_i^{\min}}{r_i^{\text{greedy\_opt}}}$.

To assess the quality of a given index set $\mathcal{J}$—e.g. $\mathcal{J}^{\text{greedy\_opt}}$—for each of the n_{exp} experiments (indexed by i), we may also generate n_{rand} random subsets of $\{1, \ldots, n\}$ of size J, calculate the resulting values of the objective (indexed by j), and compute the percentage of $r_i^{\text{rand}_j}$ values greater than the objective function of a given variant, e.g. $r_i^{\text{greedy\_opt}}$. This is denoted as, for example, $\%\, r_i^{\text{rand}} \geq r_i^{\text{greedy\_opt}}$. Averaging over experiments, we define, for example, $\%\, r^{\text{rand}} \geq r^{\text{greedy\_opt}} := \frac{1}{n_{\text{exp}}} \sum_{i=1}^{n_{\text{exp}}} (\%\, r_i^{\text{rand}} \geqslant r_i^{\text{greedy\_opt}})$.

5.2. Linear forward operator

We focus first on the linear setting.

5.2.1. Experiment set-up. We conduct 100 numerical experiments ($n_{\text{exp}} = 100$). In each experiment, a minimisation problem is randomly generated with the following objective function:

$$\Phi_i(u) = \frac{1}{2}|A_i u - y_i|^2 + \frac{1}{2}|u|^2_{R_i} \quad \text{for }, i = 1, \ldots, 100. \tag{5.2}$$

The elements of the linear forward operators $A_i \in \mathbb{R}^{30\times 50}$ are independently and identically distributed (i.i.d.) according to a uniform distribution on $[0, 1]$.

To generate R_i, we first sample $P_i \in O(50)$ uniformly from the space of real orthogonal 50×50 matrices $O(50)$, and then define $R_i := \frac{1}{\beta} P_i \Sigma P_i^\top$, where Σ is a diagonal matrix with diagonal entries decaying according to

$$\sigma_k = (1+k)^{-2}, \quad k = 1, \ldots, 50.$$

To generate $y_i \in \mathbb{R}^{30}$, we first randomly generate $u_i \in \mathbb{R}^{50}$ by sampling i.i.d. from $\mathcal{N}(0, R_i)$. We then set $y_i := A_i u_i + 10^{-4}\eta_i$, where $\eta_i \sim \mathcal{N}(0, I_{30})$ are i.i.d.

Additionally, for each experiment, we generate 100 random subsets of size J ($n_{\text{rand}} = 100$), yielding $r_i^{\text{rand}_j}$ for $j = 1, \ldots, 100$. The average residual for these random subsets is defined as $r_i^{\text{rand}} := \frac{1}{100}\sum_{j=1}^{100} r_i^{\text{rand}_j}$.

To minimise the objective function given by (5.2), we apply the following variants of algorithm AdaptSelect: **greedy_opt**, **dom_opt**, **greedy_kl**, **dom_kl**, **rand** and—when J is relatively small - **best** (see section 5.1). For each configuration of parameters J and β, 100 numerical experiments are run ($n_{\text{exp}} = 100$). For each numerical experiment, a set of initial conditions are generated, and the resulting long-term value of the objective function is calculated for each of the above variants with these same initial conditions.

We run a further 100 experiments to test the relative performance of different ensemble selection methods when the prior is misspecified. For each experiment, we generate a random linear forward operator $A_i \in \mathbb{R}^{30\times 50}$ whose elements are i.i.d. according to a uniform distribution on [0, 1] and we generate $P_i, P_i' \in O(50)$ by sampling uniformly from the space of real orthogonal 50×50 matrices. We then define the correctly specified prior covariance matrix to be $R_i' := \frac{1}{\beta} P_i' \Sigma (P_i')^T$ and the misspecified prior covariance matrix to be $R_i := \frac{1}{\beta} P_i \Sigma (P_i)^T$. We then generate $u_i \sim \mathcal{N}(0, R_i')$ and set $y_i := A_i u_i + 10^{-4}\eta_i$, where $\eta_i \sim \mathcal{N}(0, I_{30})$ are i.i.d. We calculate $\frac{r^{\text{greedy\_opt}}}{r^{\text{dom\_opt}}} := \frac{1}{100}\sum_{i=1}^{100} \frac{r_i^{\text{greedy\_opt}}}{r_i^{\text{dom\_opt}}}$ and $\frac{r^{\text{greedy\_opt}}}{r^{\text{dom\_kl}}} := \frac{1}{100}\sum_{i=1}^{100} \frac{r_i^{\text{greedy\_opt}}}{r_i^{\text{dom\_kl}}}$, where the objective function having forward operator A_i, data y_i and (misspecified) prior covariance matrix R_i is used to calculate $r_i^{\text{greedy\_opt}}, r_i^{\text{dom\_opt}}$ and $r_i^{\text{dom\_kl}}$. For comparison, we run a further 100 experiments to calculate $\frac{r^{\text{greedy\_opt}}}{r^{\text{dom\_opt}}}$ and $\frac{r^{\text{greedy\_opt}}}{r^{\text{dom\_kl}}}$ with the exact same set-up except that $P_i = P_i'$, i.e. in each experiment, the prior covariance matrix R_i that is used to generate the data y_i is also used in the objective function.

5.2.2. Experiment results

Table 1. Results for increasing J, the number of particles, in the various EKI strategies with the regularised random linear forward operators described in section 5.2.1. $\beta = 2^{-6}$ for all experiments.

Mean Residual Ratio					
	J = 2	J = 4	J = 6	J = 8	J = 10
$\frac{r\text{min}}{r\text{greedy\_opt}}$	0.337	0.555	0.696	0.803	0.848
$\frac{r\text{min}}{r\text{dom\_opt}}$	0.239	0.402	0.539	0.661	0.732
$\frac{r\text{min}}{r\text{greedy\_kl}}$	0.194	0.380	0.569	0.629	0.728
$\frac{r\text{min}}{r\text{dom\_kl}}$	0.131	0.276	0.390	0.547	0.601
$\frac{r\text{min}}{r\text{rand}}$	0.0914	0.112	0.131	0.149	0.151
Mean Residual Percentile					
$\%_r\text{rand} \geqslant_r \text{greedy\_opt}$	99.87	99.99	100	100	100
$\%_r\text{rand} \geqslant_r \text{dom\_opt}$	89.84	97.79	99.62	99.93	99.98
$\%_r\text{rand} \geqslant_r \text{greedy\_kl}$	79.71	96.01	99.51	99.66	99.16
$\%_r\text{rand} \geqslant_r \text{dom\_kl}$	58.88	89.16	95.89	99.25	99.50

Table 2. Results for increasing β, the relative weight of the prior, in the various EKI strategies with the regularised random linear forward operators described in section 5.2.1. $J = 5$ for all experiments.

Mean residual ratio						
	$\beta = 2^{-10}$	$\beta = 2^{-8}$	$\beta = 2^{-6}$	$\beta = 2^{-4}$	$\beta = 2^{-2}$	$\beta = 2^{0}$
$\frac{r\text{min}}{r\text{best}}$	0.156	0.383	0.649	0.839	0.929	0.937
$\frac{r\text{min}}{r\text{greedy\_opt}}$	0.143	0.371	0.640	0.838	0.929	0.937
$\frac{r\text{min}}{r\text{dom\_opt}}$	0.0800	0.236	0.478	0.747	0.878	0.900
$\frac{r\text{min}}{r\text{greedy\_kl}}$	0.0843	0.227	0.461	0.674	0.758	0.822
$\frac{r\text{min}}{r\text{dom\_kl}}$	0.0513	0.152	0.375	0.605	0.743	0.797
$\frac{r\text{min}}{r\text{rand}}$	0.0199	0.0572	0.119	0.202	0.353	0.447
Mean residual percentile						
$\%_r\text{rand} \geqslant_r \text{greedy\_opt}$	99.99	100	100	100	100	100
$\%_r\text{rand} \geqslant_r \text{dom\_opt}$	97.83	97.81	98.08	99.55	99.56	99.86
$\%_r\text{rand} \geqslant_r \text{greedy\_kl}$	97.94	94.58	97.36	99.03	96.26	90.66
$\%_r\text{rand} \geqslant_r \text{dom\_kl}$	87.05	92.28	95.98	95.35	92.91	94.82

Table 3. Results for the relative performance of ensemble selection according algorithm Greedy vs. selection according to the dominant eigenvalues over a range of prior weights β, for both correctly and incorrectly specified the priors. $J = 5$ for all experiments.

Mean residual ratio—correctly specified prior						
	$\beta = 10^{-6}$	$\beta = 10^{-5}$	$\beta = 10^{-4}$	$\beta = 10^{-3}$	$\beta = 10^{-2}$	$\beta = 10^{-1}$
$\frac{r^{\text{greedy\_opt}}}{r^{\text{dom\_opt}}}$	0.524	0.504	0.584	0.548	0.723	0.891
$\frac{r^{\text{greedy\_opt}}}{r^{\text{dom\_kl}}}$	0.356	0.354	0.441	0.374	0.510	0.738
Mean residual ratio—misspecified prior						
$\frac{r^{\text{greedy\_opt}}}{r^{\text{dom\_opt}}}$	0.314	0.330	0.353	0.490	0.713	0.899
$\frac{r^{\text{greedy\_opt}}}{r^{\text{dom\_kl}}}$	0.290	0.298	0.318	0.446	0.652	0.819

5.2.3. Discussion of experiment results. Table 1 shows a monotonic increase in the ratios with the number of particles for all variants of the algorithm. That this should be the case is intuitively clear and can be mathematically confirmed using corollary 4.8.

Tables 1 and 2 show that, across a range of values for J and β, $r^{\text{greedy\_opt}} \leqslant r^{\text{greedy\_kl}}$ and $r^{\text{dom\_opt}} \leqslant r^{\text{dom\_kl}}$, i.e. the 'optimal B' linear combination outperforms the KL scaling. Furthermore, $r^{\text{greedy\_opt}} \leqslant r^{\text{dom\_opt}}$ and $r^{\text{greedy\_kl}} \leqslant r^{\text{dom\_kl}}$, i.e. algorithm Greedy yields a better choice of eigenvectors than the J dominant eigenvectors.

$r^{\text{greedy\_opt}} \leqslant r^{\text{dom\_opt}} \leqslant r^{\text{dom\_kl}}$ across all values of β. For all variants of EKI, the average ratios of the long-term values of the objective function increases monotonically as the value of β increases. $\frac{r^{\text{min}}}{r^{\text{greedy\_opt}}}$, $\frac{r^{\text{min}}}{r^{\text{dom\_opt}}}$ and $\frac{r^{\text{min}}}{\bar{r}}$ converge to 1, while $\frac{r^{\text{min}}}{r^{\text{dom\_kl}}}$ and $\frac{r^{\text{min}}}{r^{\text{greedy\_kl}}}$ converges to a value strictly less than 1. This is expected since $r^{\text{min}} \to \frac{1}{2}\|y\|^2$ as $\beta \to \infty$ and for all EKI with an 'optimal B', $M_{\mathcal{J}} \approx \frac{1}{\beta}\Sigma_{\mathcal{J}}$ and corollary 4.8 therefore shows that $r^{\text{greedy\_opt}}, r^{\text{dom\_opt}}, \bar{r} \to \frac{1}{2}\|y\|^2$ as $\beta \to \infty$. For those methods with the KL scaling, $B = \beta^{-\frac{1}{2}}\Sigma_{\mathcal{J}}^{\frac{1}{2}}\Xi_{\mathcal{J}}$, corollary 4.8 shows that $r^{\text{dom\_kl}}, r^{\text{greedy\_kl}} \to \frac{1}{2}\|y\|^2 + \frac{1}{2}(1_J^T\Xi_{\mathcal{J}}^{-2}1_J)^{-1}$ as $\beta \to \infty$.

Table 3 shows that misspecifying the prior covariance matrix in the objective function (5.2) increases the degree to which EKI initialised with algorithm Greedy and with 'optimal B' outperforms EKI initialised with dominant eigenvectors, both with 'optimal B' or with the standard scaling. This is as expected, since, unlike when initialising the ensemble with the dominant eigenvectors of a misspecified prior, in algorithm Greedy, the data generated by the true prior informs the ensemble selection. As is also to be expected, the smaller the β, the larger the improvement in the relative performance: decreasing β increases the importance of the misfit term and so the influence of the data in the Tikhonov-regularised objective function (5.2).

5.3. Nonlinear forward operator

In the following, we apply the proposed EKI methods to two inverse problems with nonlinear forward operators.

5.3.1. Nonlinear algebraic forward operator. We run 10 experiments ($n_{\text{exp}} = 10$). For each experiment we randomly generate a minimisation problem with the following objective function

$$\Phi_i(u) = \frac{1}{2}\left|G_i(u) - y_i\right|^2 + \frac{1}{2}\left|u\right|_{R_i}^2, \quad i = 1, \dots, 10. \tag{5.3}$$

To generate G_i, for each experiment, we randomly sample $W_i \in \mathbb{R}^{30\times 50}$, the elements of which are i.i.d. generated according to a uniform distribution on $[0,1]$. We then define

$$\left(G_i(u)\right)_j = 0.01 + \left(1 + \left(\exp\left(10 W_i u\right)_j\right)\right)^{-1}, \quad j = 1, \dots, 30. \tag{5.4}$$

To generate the prior covariance C_i, we first generate $P_i \in O(50)$ by sampling uniformly from the space of real orthogonal 50×50 matrices $O(50)$, and then we define $R_i := \frac{1}{\beta} P_i \Sigma P_i^T$, where Σ is a diagonal matrix whose diagonal entries decay according to

$$\sigma_k = (1 + 0.1k)^{-2}, \quad k = 1, \dots, 50.$$

To generate $y_i \in \mathbb{R}^{30}$ we first randomly generate $u_i \in \mathbb{R}^{50}$ by sampling i.i.d. uniformly from $\mathcal{N}(0, R_i)$ and then set $y_i := G_i(u_i) + 10^{-4}\eta_i$, where $\eta_i \sim \mathcal{N}(0, I_{30\times 30})$ i.i.d.

To minimise the objective function given by (5.3), we apply the following variants of algorithm AdaptSelect: **greedy_opt_r**, **greedy_opt**, **dom_opt_r**, **dom_opt**, **greedy_kl_r**, **greedy_kl** and **dom_kl** (see section 5.1). For each configuration of parameters J and β, 10 numerical experiments are run ($n_{\text{exp}} = 10$). For each numerical experiment, initial conditions are generated and each of the above variants is run with these same initial conditions. All variants of EKI are run for a total time of $T = 200$ and the resampling occurs at times $t = \frac{200}{3}, \frac{400}{3}$ for those variants that resample.

5.3.2. Experiment results

Table 4. Results for increasing J, the number of particles, in the various EKI strategies with the regularised algebraic forward operator given by equation (5.4). $\beta = 2^{-4}$ for all experiments.

	Mean residual ratio				
	$J = 2$	$J = 4$	$J = 6$	$J = 8$	$J = 10$
$\frac{r_{\min}}{r_{\text{greedy\_opt\_r}}}$	0.564	0.722	0.809	0.873	0.905
$\frac{r_{\min}}{r_{\text{greedy\_opt}}}$	0.253	0.365	0.461	0.495	0.517
$\frac{r_{\min}}{r_{\text{dom\_opt\_r}}}$	0.243	0.375	0.504	0.459	0.564
$\frac{r_{\min}}{r_{\text{dom\_opt}}}$	0.186	0.281	0.455	0.405	0.473
$\frac{r_{\min}}{r_{\text{greedy\_kl\_r}}}$	0.281	0.407	0.632	0.752	0.748
$\frac{r_{\min}}{r_{\text{greedy\_kl}}}$	0.283	0.339	0.457	0.440	0.533
$\frac{r_{\min}}{r_{\text{dom\_kl}}}$	0.126	0.280	0.435	0.395	0.441

Table 5. Results for increasing β, the relative weight of the prior, in the various EKI strategies with the regularised algebraic forward operator given by equation (5.4). $J = 5$ for all experiments.

	Mean residual ratio					
	$\boldsymbol{\beta = 2^{-10}}$	$\boldsymbol{\beta = 2^{-8}}$	$\boldsymbol{\beta = 2^{-6}}$	$\boldsymbol{\beta = 2^{-4}}$	$\boldsymbol{\beta = 2^{-2}}$	$\boldsymbol{\beta = 2^{0}}$
$\frac{r_{\min}}{r_{\text{greedy\_opt\_r}}}$	0.772	0.772	0.808	0.754	0.790	0.810
$\frac{r_{\min}}{r_{\text{greedy\_opt}}}$	0.444	0.440	0.467	0.451	0.447	0.499
$\frac{r_{\min}}{r_{\text{dom\_opt\_r}}}$	0.346	0.471	0.429	0.364	0.441	0.373
$\frac{r_{\min}}{r_{\text{dom\_opt}}}$	0.304	0.438	0.386	0.356	0.409	0.372
$\frac{r_{\min}}{r_{\text{greedy\_kl\_r}}}$	0.592	0.596	0.553	0.511	0.574	0.574
$\frac{r_{\min}}{r_{\text{greedy\_kl}}}$	0.331	0.381	0.355	0.351	0.419	0.364
$\frac{r_{\min}}{r_{\text{dom\_kl}}}$	0.302	0.391	0.327	0.376	0.331	0.199

5.3.3. Discussion of experiment results. Table 4 shows that, for all variants of the algorithm, there is the expected general decrease in the long-term value of the objective function as the number of particles increases.

Tables 4 and 5 show that, across a range of values of J and β, with few exceptions, the relative performances of the variants of EKI are as expected: ${}_r\text{greedy\_opt\_r} \leqslant {}_r\text{greedy\_opt}$, ${}_r\text{dom\_opt\_r} \leqslant {}_r\text{dom\_opt}$, ${}_r\text{greedy\_kl\_r} \leqslant {}_r\text{greedy\_kl}$, ${}_r\text{greedy\_opt\_r} \leqslant {}_r\text{dom\_opt\_r}$, ${}_r\text{greedy\_opt} \leqslant {}_r\text{dom\_opt}$ and ${}_r\text{dom\_opt} \leqslant {}_r\text{dom\_kl}$. In general, ${}_r\text{greedy\_kl} \leqslant {}_r\text{dom\_kl}$ as well, even though this is not necessarily to be expected, since algorithm Greedy only aims to minimise ${}_r\text{greedy\_opt}$ in the linear setting. These results suggest that resampling, selecting eigenvectors according to algorithm Greedy and selecting B 'optimally' - i.e. according to corollary 4.10—all improve performance.

5.3.4. 2D darcy flow. We consider the following elliptic PDE:

$$\begin{cases} -\nabla \cdot (\exp(u)\nabla p) = f, & \text{in } D \\ p = 0, & \text{on } \partial D. \end{cases} \tag{5.5}$$

With $D = (0,1)^2$. We seek to recover the unknown diffusion coefficient $u^\dagger \in C^1(D) = X$, given an observation of the solution $p \in H^2(D) \cap H_0^1(D) =: V$ (see Evans (2010), chapter 5 for details on the Sobolev spaces $H^2(D)$ and $H_0^1(D)$). Furthermore, we assume that the scalar field $f \in \mathbb{R}$ is known. The observations are given by

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

where $\mathcal{O}(p) : V \to \mathbb{R}^K$ is the observation operator, which considers K randomly chosen points in X, i.e. $\mathcal{O}(p) = (p(x_1), \ldots, p(x_K))$. Finally, $\eta \sim \mathcal{N}(0, 10^{-5} I_k)$ denotes the noise on our data. Then the inverse problem is given by

$$y = G(u) + \eta,$$

where $G = \mathcal{O} \circ \mathcal{G}$ and $\mathcal{G} : X \to \mathbb{R}^K$ denotes the solution operator of (5.5). We solve the PDE on a uniform mesh with N_{FEM}^2 points $[\frac{0}{N_{\text{FEM}}}, \frac{1}{N_{\text{FEM}}}, \ldots, \frac{N_{\text{FEM}}-1}{N_{\text{FEM}}}]^2$ using a finite element method (FEM) with continuous, piecewise linear finite element basis functions. We model the prior distribution as a truncated KL expansion with N_{KL}^2 terms:

$$u(x,\omega) = \mu + \sum_{k,l=1}^{N_{\text{KL}}} \lambda_{k,l}^{\frac{1}{2}} e_{k,l}(x)\, \xi_{k,l}(\omega),$$

where $\mu \in \mathbb{R}^{N_{\text{FEM}}^2}$, $\lambda_{k,l} = \frac{1}{\beta}(\pi^2(k^2+l^2)+\tau^2)^{-\alpha}$, $e_{k,l}(x) = \frac{2}{N_{\text{FEM}}} \cos(2\pi x_1 k)\cos(2\pi x_2 l) \in \mathbb{R}^{N_{\text{FEM}}^2}$, $\xi_{k,l} \sim \mathcal{N}(0,1)$ are i.i.d. and $(k,l) \in [1,\ldots,N_{\text{FEM}}]^2$. Due to the FEM discretisation, $u(\vec{x},\omega), e_{k,l}(\vec{x}), \vec{x} \in \mathbb{R}^{N_{\text{FEM}} \times N_{\text{FEM}}}$, where $\vec{x} \in \mathbb{R}^{N_{\text{FEM}} \times N_{\text{FEM}}}$ is a vector of the FEM mesh-points.

$$\begin{aligned} u(\vec{x},\omega) &= \mu + \left[e_{1,1}(\vec{x}), \ldots, e_{N_{\text{KL}},N_{\text{KL}}}(\vec{x})\right] \operatorname{diag}\left(\lambda_{1,1}^{\frac{1}{2}}, \ldots, \lambda_{N_{\text{KL}},N_{\text{KL}}}^{\frac{1}{2}}\right) \xi(\omega) \\ &=: \mu + E(\vec{x})\, \Lambda^{\frac{1}{2}} \xi(\omega), \end{aligned}$$

where $E(\vec{x}) \in \mathbb{R}^{N_{\text{FEM}}^2 \times N_{\text{KL}}^2}$, $\Lambda \in \mathbb{R}^{N_{\text{KL}}^2 \times N_{\text{KL}}^2}$ and $\xi \sim \mathcal{N}(0, I^2_{N_{\text{KL}}})$. (The columns of $E(\vec{x})$ and Λ are in lexicographic order with respect to the indices of $e_{k,l}(\vec{x})$ and $\lambda_{k,l}$ respectively.) We note that $E(\vec{x})^T E(\vec{x}) = I^2_{N_{\text{KL}}}$ and $R_u := Cov(u) = E\Lambda E^T$. We then define the un-whitened variable $w := E(\vec{x})^T(u-\mu) = \Lambda^{\frac{1}{2}}\xi \in \mathbb{R}^{N_{\text{KL}}^2}$ so that $w \sim \mathcal{N}(0,\Lambda)$. Further we define the forward operator $\tilde{G} : \mathbb{R}^{N_{\text{KL}}^2} \to \mathbb{R}^K$ such that $\tilde{G}(w) = G(\mu + Ew)$. Then

$$\frac{1}{2}\left|G(u)-y\right|^2 + \frac{\beta}{2}\left|u\right|^2_{R_u} = \frac{1}{2}\left|\tilde{G}(w)-y\right|^2 + \frac{1}{2}\left|w\right|^2_{\Lambda}. \tag{5.6}$$

To minimise this objective function, we apply EKI to the variable w rather than to u, i.e. we run EKI methods with forward operator $\left(\tilde{G}(w)^T \quad (\Lambda^{-\frac{1}{2}}w)^T\right)^T$ and data $\left(y^T \quad 0^T_{N_{\text{KL}}}\right)^T$.

To minimise the objective function given by (5.6), we apply the following variants of algorithm AdaptSelect: **greedy_opt_r**, **greedy_opt**, **dom_opt_r**, **dom_opt**, **greedy_kl_r**, **greedy_kl** and **dom_kl** (see section 5.1). For each configuration of parameters J and β, 10 numerical experiments are run ($n_{\text{exp}} = 10$). For each numerical experiment, initial conditions are generated and each of the above variants is run with these same initial conditions. All variants of EKI are run for a total time of $T = 1$ and the resampling occurs at times $t = \frac{1}{3}, \frac{2}{3}$ for those variants that resample. We take $N_{\text{FEM}} = 64$, $N_{\text{KL}} = 7$, $K = 30$, $\mu = -8$, $\alpha = 1$, $\tau = 1$ and $f = 1$ in equation (5.5).

5.3.5. *Experiment results*

Table 6. Results for increasing the J, the number of particles, in the various EKI strategies, where the forward operator is the regularised observed solution operator to the Darcy flow problem given by PDE (5.5). In these experiments $\beta = 2^{-6}$.

	Mean residual ratio				
	$J = 2$	$J = 4$	J = 6	$J = 8$	$J = 10$
$\frac{r^{\text{min}}}{r^{\text{greedy\_opt\_r}}}$	0.810	0.929	0.962	0.977	0.984
$\frac{r^{\text{min}}}{r^{\text{greedy\_opt}}}$	0.499	0.732	0.824	0.872	0.893
$\frac{r^{\text{min}}}{r^{\text{dom\_opt\_r}}}$	0.373	0.589	0.671	0.816	0.826
$\frac{r^{\text{min}}}{r^{\text{dom\_opt}}}$	0.372	0.587	0.668	0.810	0.821
$\frac{r^{\text{min}}}{r^{\text{greedy\_kl\_r}}}$	0.270	0.350	0.356	0.357	0.367
$\frac{r^{\text{min}}}{r^{\text{greedy\_kl}}}$	0.223	0.289	0.302	0.297	0.312
$\frac{r^{\text{min}}}{r^{\text{dom\_kl}}}$	0.199	0.268	0.299	0.300	0.317

Table 7. Results for increasing β, the relative weight of the prior, in the various EKI strategies , where the forward operator is the regularised observed solution operator to the Darcy flow problem given by PDE (5.5). In these experiments $J = 5$.

	Mean residual ratio					
	$\beta = 2^{-10}$	$\beta = 2^{-8}$	$\beta = 2^{-6}$	$\beta = 2^{-4}$	$\beta = 2^{-2}$	$\beta = 2^{0}$
$\frac{r\text{min}}{r\text{greedy\_opt\_r}}$	0.588	0.848	0.956	0.982	0.998	0.998
$\frac{r\text{min}}{r\text{greedy\_opt}}$	0.289	0.581	0.813	0.925	0.988	0.990
$\frac{r\text{min}}{r\text{dom\_opt\_r}}$	0.169	0.412	0.732	0.890	0.978	0.985
$\frac{r\text{min}}{r\text{dom\_opt}}$	0.165	0.396	0.732	0.890	0.978	0.985
$\frac{r\text{min}}{r\text{greedy\_kl\_r}}$	0.456	0.373	0.383	0.535	0.481	0.539
$\frac{r\text{min}}{r\text{greedy\_kl}}$	0.221	0.223	0.256	0.275	0.463	0.598
$\frac{r\text{min}}{r\text{dom\_kl}}$	0.0873	0.164	0.318	0.447	0.451	0.481

5.3.6. Discussion of experiment results. Table 6 shows that, for all variants of the algorithm, there is the expected general decrease in the long-term value of the objective function as number of particles increases.

Table 7 shows that, for all variants of the algorithm, there is the expected general decrease in the long-term value of the objective function as β increases. This is as expected, since the forward operator G is approximately linear, a larger β can be interpreted as making the Tikhonov-regularised forward operator 'more linear' (by putting more weight on a linear component of the forward operator) and a general decrease in the long-term value of the objective function is mathematically expected in the linear setting (see section 5.2.3).

Tables 6 and 7 show that, across a range of values of J and β, with few exceptions, the relative performances of the variants of EKI are as expected: $r^{\text{greedy\_opt\_r}} \leqslant r^{\text{greedy\_opt}}$, $r^{\text{dom\_opt\_r}} \leqslant r^{\text{dom\_opt}}$, $r^{\text{greedy\_kl\_r}} \leqslant r^{\text{greedy\_kl}}$, $r^{\text{greedy\_opt\_r}} \leqslant r^{\text{dom\_opt\_r}}$, $r^{\text{greedy\_opt}} \leqslant r^{\text{dom\_opt}}$ and $r^{\text{dom\_opt}} \leqslant r^{\text{dom\_kl}}$. In general, $r^{\text{greedy\_kl}} \leqslant r^{\text{dom\_kl}}$ as well, even though this is not necessarily to be expected, since algorithm Greedy only aims to minimise $r^{\text{greedy\_opt}}$ in the linear setting. These results suggest that resampling, selecting eigenvectors according to algorithm Greedy and selecting B 'optimally' - i.e. according to corollary 4.10—all improve performance.

6. Conclusion and outlook

This work represents a comprehensive study on optimising the subspace within the EKI framework to address challenges associated with poor prior knowledge and small ensemble sizes. By explicitly solving the EKI particle dynamics in the linear setting, we develop a novel greedy strategy for selecting a solution subspace and strategies for selecting the initial ensemble within

this subspace. We extend these results to develop an algorithm for performing EKI with adaptive resampling that is suitable for both linear and nonlinear problems.

Numerical experiments validate the theoretical findings, showcasing the effectiveness of the proposed methods in diverse scenarios, including high-dimensional problems and ill-posed settings. These experiments show that, in terms of accuracy, the proposed method significantly outperforms standard EKI methods and performs well even compared to more computationally expensive optimisation methods. Preliminary experiments also suggest that, by using data, forward operator and prior information, the proposed method is more robust to prior misspecification than standard EKI methods. Despite small ensemble sizes and a limited number of derivative calculations, the proposed methods effectively leverage the structure of inverse problems to yield impressive results.

Future work could further develop the adaptive resampling in the proposed algorithm to improve its performance in highly nonlinear settings. More specifically, one could systematically investigate how to effectively choose certain parameters in the algorithm, such as ensemble sizes and resample times. Furthermore, it would be interesting to investigate alternative linearisations of the forward operator in order to make the proposed algorithm entirely derivative-free. Another direction of future research would be to combine our proposed method with machine learning techniques that use suitable embeddings to transform nonlinear problems into linear ones.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Acknowledgment

The work of R H has been funded by Deutsche Forschungsgemeinschaft (DFG) through the Grant CRC 1114 'Scaling Cascades in Complex Systems' (Project Number 235221301, Project A02 - Multiscale data and asymptotic model assimilation for atmospheric flows). C S acknowledges support from MATH+ Project EF1-19: Machine Learning Enhanced Filtering Methods for Inverse Problems and EF1-20: Uncertainty Quantification and Design of Experiment for Data-Driven Control, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy-The Berlin Mathematics Research Center MATH+ (EXC-2046/1, Project ID: 390685689).

Appendix A. Technical results

Several key quantities in the analysis of EKI for linear problems evolve according to the dynamics described in the following theorem. As noted by Bungert and Wacker (2021), the mean-field dynamics of EKI evolve according to equations below with $\alpha = 1$, while the leading term in the average sample equation for stochastic linear EKI with J particles evolves according to the equations below with $\alpha = \frac{J+1}{J}$. In this paper, however, we focus solely on the case $\alpha = 2$, which characterises the particle dynamics of deterministic linear EKI.

Theorem A.1. *Let $\alpha > 0$, $A \in \mathbb{R}^{k\times n}$, $x_0 \in \mathbb{R}^n$, $C_0 \in \mathbb{R}^{n\times n}$ be symmetric and positive semi-definite. Further let $C_0 = C_L C_R \in \mathbb{R}^{n\times n}$ be any desired decomposition, where $C_L \in \mathbb{R}^{n\times p}$,*

$C_R \in \mathbb{R}^{p\times n}$ for some $p \in \mathbb{N}^+$. Let AC_0A^T be of rank r, let $AC_0A^T = \mathcal{U}\Sigma\mathcal{U}^T$ be any desired eigen-decomposition with $\mathcal{U} \in \mathbb{R}^{n\times r}$ and $\Sigma \in \mathbb{R}^{r\times r}$. Then the IVP

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}C_t = -\alpha C_t A^T A C_t, \quad t > 0,\\
&\frac{d}{dt}x_t = -C_t A^T\left(Ax_t - y\right), \quad t > 0,\\
&C(0) = C_0, \quad u(0) = u_0.
\end{aligned} \tag{A.1}$$

Has unique solution

$$\begin{aligned}
C_t &= C_L\left(I_p + \alpha C_R A^T A C_L t\right)^{-1} C_R,\\
x_t &= x_0 + (AC_0)^T \mathcal{U}\Sigma^{-1}\left(\left(I_r + \alpha\Sigma t\right)^{-\frac{1}{\alpha}} - I_r\right)\mathcal{U}^T\left(Ax_0 - y\right).
\end{aligned}$$

Proof. It can be easily verified that the given expression for C_t satisfies the IVP (A.1). Perhaps the trickiest part in doing so is to prove that this expression is well defined, i.e. that $(I + (\alpha t)C_L A^T A C_R)$ is invertible.

$$\left(I + (\alpha t)\, C_R A^T A C_L\right) x = 0,$$

implies

$$0 = AC_L\left(I + (\alpha t)\, C_R A^T A C_L\right) x = \left(I + (\alpha t) AC_0 A^T\right) AC_L x,$$

which leads to

$$AC_L x = 0,$$

and thus

$$x = -\left(\alpha t\right) C_R A^T A C_L x = 0.$$

Likewise, it can be easily verified that the given expression for x_t—in combinations with the expression for C_t—satisfies the IVP (A.1). However, we presently include a derivation of this expression, since this might prove instructive for solving similar ODEs arising from EKI. To begin with, from the first ODE in the IVP (A.1), we derive the following ODE for AC_tA^T:

$$\frac{\mathrm{d}}{\mathrm{d}t}AC_tA^T = -\alpha\left(AC_tA^T\right)\left(AC_tA^T\right). \tag{A.2}$$

If $AC_0A^T = \mathcal{U}\Sigma\mathcal{U}^T$ is a diagonal decomposition, then the solution to (A.2) is given by

$$\begin{aligned}
AC_tA^T &= \mathcal{U}\Sigma_t\mathcal{U}^T,\\
\Sigma_t &= \Sigma\left(I + \alpha\Sigma t\right)^{-1}.
\end{aligned} \tag{A.3}$$

Next, from the first ODE in the IVP (A.1), we derive the ODE

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(AC_t\right) = -\alpha\left(AC_tA^T\right)\left(AC_t\right),$$

Which we solve using our expression for AC_tA^T given by equation (A.3)

$$\begin{aligned} AC_t &= \exp\left(-\alpha \int_0^t \mathcal{U}\Sigma_s \mathcal{U}^T ds\right) AC_0 \\ &= \left(I - \mathcal{U}\mathcal{U}^T + \mathcal{U}\left(I + \alpha\Sigma t\right)^{-\frac{1}{\alpha}} \mathcal{U}^T\right) AC_0. \end{aligned}$$

Similarly, from the second ODE in the IVP (A.1), we derive the expression

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(Ax_t - y\right) = -\left(AC_tA^T\right)\left(Ax_t - y\right),$$

which we solve using our expression for AC_tA^T given by equation (A.3)

$$\begin{aligned} Ax_t - y &= \exp\left(-\int_0^t AC_sA^T \mathrm{d}s\right)\left(Ax_0 - y\right) \\ &= \left(I - \mathcal{U}\mathcal{U}^T + \mathcal{U}\left(I + \alpha\Sigma t\right)^{-1}\mathcal{U}^T\right)\left(Ax_0 - y\right). \end{aligned}$$

Finally,

$$\begin{aligned} x_t &= x_0 + \int_0^t \frac{\mathrm{d}}{\mathrm{d}s} x_s \mathrm{d}s \\ &= x_0 - \int_0^t \left(AC_s\right)^T \left(Ax_s - y\right) \mathrm{d}s \\ &= x_0 + \left(AC_0\right)^T \mathcal{U}\Sigma^{-1}\left(\left(I + \alpha\Sigma t\right)^{-\frac{1}{\alpha}} - I\right)\mathcal{U}^T\left(Ax_0 - y\right). \end{aligned}$$

□

Lemma A.2. *Recall that* $\mathcal{E}_0 = [u_0^{(1)} - \bar{u}(0), \ldots, u_0^{(J)} - \bar{u}(0)] = U_0(I_J - \frac{1}{J}1_J1_J^T)$*, where* $1_J \in \mathbb{R}^J$ *is a vector of ones. Let* $A\mathcal{E}_0 = \sqrt{J}\mathcal{U}\Sigma V^T$ *be a compact singular value decomposition (*$\mathcal{U} \in \mathbb{R}^{m\times r}$*,* $\Sigma \in \mathbb{R}^{r\times r}$*,* $V \in \mathbb{R}^{J\times r}$*, where* $\Sigma \succ 0$*). Then* $\mathcal{E}_0V = U_0V$*.*

Proof. There holds

$$\begin{aligned} V^T 1_J &= \frac{1}{\sqrt{J}}\Sigma^{-1}\mathcal{U}^T\left(\sqrt{J}\mathcal{U}\Sigma V^T\right)1_J \\ &= \frac{1}{\sqrt{J}}\Sigma^{-1}\mathcal{U}^T A\mathcal{E}_0 1_J \\ &= \frac{1}{\sqrt{J}}\Sigma^{-1}\mathcal{U}^T A U_0\left(I_J - \frac{1}{J}1_J1_J^T\right)1_J \\ &= 0. \end{aligned}$$

Then

$$\begin{aligned} \mathcal{E}_0V &= U_0\left(I_J - \frac{1}{J}1_J1_J^T\right)V \\ &= U_0V. \end{aligned}$$

□

Lemma A.3. *Let $R \in \mathbb{R}^{n\times n}$ have eigen-decomposition $R = V\Lambda V^T$, where $V, \Lambda \in \mathbb{R}^{n\times n}$. Let $\mathcal{J} \subset \{1,\dots,n\}$ be an index set of size $J \leqslant n$. If $V = [v_1,\dots,v_n]$, define $V_{\mathcal{J}} := [v_{\mathcal{J}_1},\dots,v_{\mathcal{J}_J}]$, $V_{\mathcal{J}^c} := [v_{\mathcal{J}_1^c},\dots,v_{\mathcal{J}_J^c}]$, $\Lambda_{\mathcal{J}} := diag(\{\lambda_{\mathcal{J}_1},\dots,\lambda_{\mathcal{J}_J}\})$, $R_{\mathcal{J}}^{\frac{1}{2}} := V_{\mathcal{J}}\Lambda_{\mathcal{J}}^{\frac{1}{2}}V_{\mathcal{J}}^T$ and $(R_{\mathcal{J}}^{\dagger})^{\frac{1}{2}} := V_{\mathcal{J}}\Lambda_{\mathcal{J}}^{-\frac{1}{2}}V_{\mathcal{J}}^T$. $U_0 := V_{\mathcal{J}}B$, where $B \in \mathbb{R}^{J\times J}$ is invertible. $C_0 := \frac{1}{J}U_0(I_J - \frac{1}{J}1_J1_J^T)U_0^T$, where $1_J \in \mathbb{R}^J$ is a vector of ones. $A \in \mathbb{R}^{m\times n}$ and $\tilde{A} := \begin{pmatrix} A \\ (R_{\mathcal{J}}^{\dagger})^{\frac{1}{2}} \end{pmatrix} \in \mathbb{R}^{(m+n)\times n}$. Then*

$$ker\left(\tilde{A}C_0\tilde{A}^T\right) = Im\begin{pmatrix} I_{m\times m} & 0_{m\times 1} & 0_{m\times(n-J)} \\ -R_{\mathcal{J}}^{\frac{1}{2}}A^T & R_{\mathcal{J}}^{\frac{1}{2}}V_{\mathcal{J}}B^{-T}1_J & V_{\mathcal{J}^c} \end{pmatrix}.$$

Furthermore, the columns of the matrix on the right-hand side are linearly independent.

Proof. Take

$$p \in \ker\left(\tilde{A}C_0\tilde{A}^T\right).$$

Then

$$0 = \left(I_J - \frac{1}{J}1_J1_J^T\right)B^TV_{\mathcal{J}}^T\tilde{A}^Tp,$$

and thus

$$\tilde{A}^Tp = \gamma V_{\mathcal{J}}B^{-T}1_J + x,$$

for some $\gamma \in \mathbb{R}$ and $x \in \mathrm{Im}(V_{\mathcal{J}^c})$.

Writing $p = \begin{pmatrix} r \\ s \end{pmatrix}$ for $r \in \mathbb{R}^m$ and $s \in \mathbb{R}^n$, we get

$$A^Tr + \left(R_{\mathcal{J}}^{\dagger}\right)^{\frac{1}{2}}s = \tilde{A}^Tp = \gamma V_{\mathcal{J}}B^{-T}1_J + x.$$

Rearranging,

$$V_{\mathcal{J}}V_{\mathcal{J}}^Ts = -R_{\mathcal{J}}^{\frac{1}{2}}A^Tr + \gamma V_{\mathcal{J}}\Lambda_{\mathcal{J}}^{\frac{1}{2}}B^{-T}1_J,$$

and therefore

$$p = \begin{pmatrix} I_m \\ -R_{\mathcal{J}}^{\frac{1}{2}}A^T \end{pmatrix}r + \gamma\begin{pmatrix} 0_{m\times 1} \\ V_{\mathcal{J}}\Lambda_{\mathcal{J}}^{\frac{1}{2}}B^{-T}1_J \end{pmatrix} + \begin{pmatrix} 0_{m\times 1} \\ V_{\mathcal{J}^c}V_{\mathcal{J}^c}^Ts \end{pmatrix}.$$

It follows that $\ker(\tilde{A}C_0\tilde{A}^T)$ is spanned by the columns of the following matrix:

$$\begin{pmatrix} I_m & 0_{m\times 1} & 0_{m\times(n-J)} \\ -R_{\mathcal{J}}^{\frac{1}{2}}A^T & R_{\mathcal{J}}^{\frac{1}{2}}V_{\mathcal{J}}B^{-T}1_J & V_{\mathcal{J}^c} \end{pmatrix} \in \mathbb{R}^{(m+n)\times(m+n-J+1)}.$$

It is clear that the columns of this matrix are linearly independent. □

Lemma A.4. *Let $x, y \in R^n$, $\Sigma \in \mathbb{R}^{n\times n}$ be diagonal with $\Sigma \succ 0$ and $1 = x^T\Sigma^{-1}y$. Define $\alpha \in R^n$ by $\alpha_i := x_iy_i$. Then*

$$trace\left(\Sigma^2\right) \geqslant \left(|\alpha|_{\frac{2}{3}}\right)^2.$$

Furthermore, this lower bound is achieved iff

$$\forall k : \sigma_k = \alpha_k^{\frac{1}{3}} \left(|\alpha|_{\frac{2}{3}} \right)^{\frac{2}{3}}.$$

Proof. There holds

$$\left\{ \Sigma : \Sigma \succ 0, \quad x^T \Sigma^{-1} y = 1 \right\} = \left\{ \left(x^T \Sigma^{-1} y \right) \Sigma : \Sigma \succ 0, \quad x^T \Sigma^{-1} y \neq 0 \right\}.$$

With $\tilde{\Sigma} := (x^T \Sigma^{-1} y)\Sigma$ we get

$$\begin{aligned} \frac{\partial}{\partial \sigma_k} \text{trace} \left(\tilde{\Sigma}^T \tilde{\Sigma} \right) &= \frac{\partial}{\partial \sigma_k} \left(\sum_i \frac{\alpha_i}{\sigma_i} \right)^2 \sum_j \sigma_j^2 \\ &= -2 \frac{\alpha_k}{\sigma_k^2} \left(\sum_i \frac{\alpha_i}{\sigma_i} \right) \sum_j \sigma_j^2 + 2\sigma_k \left(\sum_i \frac{\alpha_i}{\sigma_i} \right)^2. \end{aligned}$$

Setting

$$\frac{\partial}{\partial \sigma_k} \text{trace} \left(\tilde{\Sigma}^T \tilde{\Sigma} \right) = 0,$$

and re-arranging, we get

$$\sigma_k = \gamma \alpha_k^{\frac{1}{3}},$$

for some $\gamma \neq 0$. This yields

$$\tilde{\sigma}_k = \left(\sum_i \frac{\alpha_i}{\sigma_i} \right) \sigma_k = \left(\sum_i \frac{\alpha_i}{\gamma \alpha_i^{\frac{1}{3}}} \right) \gamma \alpha_k^{\frac{1}{3}} = \alpha_k^{\frac{1}{3}} \sum_i \alpha_i^{\frac{2}{3}}.$$

□

Lemma A.5. *Let $x, y \in \mathbb{R}^J$ and $U, V \in \mathbb{R}^{J \times J}$ be orthogonal matrices such that*

$$\begin{aligned} V^T x / |x|_2 &= 1_J / \sqrt{J}, \\ U^T y / |y|_2 &= 1_J / \sqrt{J}, \end{aligned}$$

where $1_J \in \mathbb{R}^J$ is a vector of ones. Define

$$B := |x|_2 |y|_2 UV^T.$$

Then

$$\begin{aligned} 1 &= x^T B^{-1} y, \\ \text{trace} \left(B^T B \right) &= J |x|_2^2 |y|_2^2. \end{aligned}$$

Proof. Let $\tilde{B} \in \mathbb{R}^{J \times J}$ be invertible with singular value decomposition $\tilde{U} \tilde{\Sigma} \tilde{V}^T$ such that $1 = x^T \tilde{B}^{-1} y = (\tilde{V}^T x)^T \tilde{\Sigma}^{-1} (\tilde{U}^T y)$. Suppose we are choosing $\tilde{\Sigma}$ to minimise $\text{trace}(\tilde{B}^T \tilde{B}) = \text{trace}(\tilde{\Sigma}^T \tilde{\Sigma})$. According to lemma A.4, we should set

$$\tilde{\sigma}_j = \left(\tilde{V}^T x \right)_i^{\frac{1}{3}} \left(\tilde{U}^T y \right)_i^{\frac{1}{3}} \left(\left| \left(\tilde{V}^T x \right) \left(\tilde{U}^T y \right) \right|_{\frac{2}{3}} \right)^{\frac{3}{2}}.$$

In this case

$$\begin{aligned}\operatorname{trace}\left(\tilde{B}^T\tilde{B}\right) &= \left(\left|\left(\tilde{V}^T x\right)\left(\tilde{U}^T y\right)\right|_{\frac{2}{3}}\right)^2 \\ &\leqslant \left(\left|\left(\tilde{V}^T x\right)\right|_{\frac{4}{3}}\right)^2 \left(\left|\left(\tilde{U}^T y\right)\right|_{\frac{4}{3}}\right)^2.\end{aligned}$$

Setting $\tilde{U} = U$ and $\tilde{V} = V$—where U and V are defined above—minimises the right-hand side of this inequality. In this case

$$\begin{aligned}\tilde{\sigma}_j &= |x|_2 |y|_2, \\ \tilde{B} &= |x|_2 |y|_2 UV^T.\end{aligned}$$

□

Appendix B. Proofs of the main results

Proof of theorem 4.7. Let $u \in \operatorname{span}\{v_{\mathcal{J}_i}\}_{i\leqslant J}$. Then

$$\begin{aligned}|Au-y|^2 + |u-\mu|_R^2 &= |Au-y|^2 + |u-\mu|_{R_{\mathcal{J}}}^2 + |u-\mu|_{R_{\mathcal{J}^c}}^2 \\ &= \left|\begin{pmatrix} A \\ \left(R_{\mathcal{J}}^{\dagger}\right)^{\frac{1}{2}} \end{pmatrix} u - \begin{pmatrix} y \\ \left(R_{\mathcal{J}}^{\dagger}\right)^{\frac{1}{2}} \mu \end{pmatrix}\right|^2 + |\mu|_{R_{\mathcal{J}^c}}^2 \\ &= \left|\tilde{A}u - \tilde{y}\right|^2 + |\mu|_{R_{\mathcal{J}^c}}^2,\end{aligned}$$

where $\tilde{A} := \begin{pmatrix} A \\ (R_{\mathcal{J}}^{\dagger})^{\frac{1}{2}} \end{pmatrix} \in \mathbb{R}^{(m+n)\times n}$ and $\tilde{y} := \begin{pmatrix} y \\ (R_{\mathcal{J}}^{\dagger})^{\frac{1}{2}}\mu \end{pmatrix} \in \mathbb{R}^{m+n}$.

Since $[u_{\mathcal{J},B}^{(1)}(0),\ldots,u_{\mathcal{J},B}^{(J)}(0)] = V_{\mathcal{J}}B$, the subspace property (lemma 3.1) means that $u_{\mathcal{J},B}^{(i)}(t) \in \operatorname{span}\{v_{\mathcal{J}_i}\}_{i\leqslant J}$ for all $t \geqslant 0$. Furthermore, Since $\operatorname{span}\{v_{\mathcal{J}_i}\}_{i\leqslant J} \subseteq \mathbb{R}^n$ is a closed subspace, $\lim_{t\to\infty} u_{\mathcal{J},B}^{(i)}(t) \in \operatorname{span}\{v_{\mathcal{J}_i}\}_{i\leqslant J}$ as well. Due to corollary 4.6, all particles converge to the same point in projection space, i.e. $\lim_{t\to\infty} \tilde{A}u_{\mathcal{J},B}^{(i)}(t) = \lim_{t\to\infty} \tilde{A}\bar{u}_{\mathcal{J},B}^{(i)}(t)$. As such, we are interested in calculating

$$\lim_{t\to\infty} |A\bar{u}_{\mathcal{J},B}(t) - y|^2 + |\bar{u}_{\mathcal{J},B}(t) - \mu|_R^2 = \lim_{t\to\infty} \left|\tilde{A}\bar{u}_{\mathcal{J},B}(t) - \tilde{y}\right|^2 + |\mu|_{R_{\mathcal{J}^c}^{\dagger}}^2.$$

According to corollary 4.3, $\lim_{t\to\infty} \tilde{A}\bar{u}_{\mathcal{J},B}(t) - \tilde{y} = Proj_{\ker(\tilde{A}C_0\tilde{A}^T)}(\tilde{A}\bar{u}_{\mathcal{J},B}(0) - \tilde{y})$. From lemma A.3 we have that $\ker(\tilde{A}C_0\tilde{A}^T) = \operatorname{Im}(P)$, where $P = \begin{pmatrix} I_m & 0_{m\times 1} & 0_{m\times(n-J)} \\ -R_{\mathcal{J}}^{\frac{1}{2}}A^T & \beta & V_{\mathcal{J}^c} \end{pmatrix}$ and $\beta := R_{\mathcal{J}}^{\frac{1}{2}} V_{\mathcal{J}} B^{-T} 1_J = V_{\mathcal{J}} \Lambda_{\mathcal{J}}^{\frac{1}{2}} B^{-T} 1_J$. Since the columns of P are linearly independent, P^TP is invertible and $Proj_{\ker(\tilde{A}C_0\tilde{A}^T)} = P(P^TP)^{-1}P^T$. As such

$$\left|Proj_{\ker\left(\tilde{A}C_0\tilde{A}^T\right)}\left(\tilde{A}\bar{u}_{\mathcal{J},B}(0) - \tilde{y}\right)\right|^2 = \left(P^T\left(\tilde{A}\bar{u}_{\mathcal{J},B}(0) - \tilde{y}\right)\right)^T \left(P^TP\right)^{-1} P^T\left(\tilde{A}\bar{u}_{\mathcal{J},B}(0) - \tilde{y}\right).$$

We first calculate

$$
\begin{aligned}
P^T\left(\tilde{A}\bar{u}_{\mathcal{J},B}(0)-\tilde{y}\right) &= P^T\left(\frac{1}{J}\begin{pmatrix} A \\ \left(R^\dagger_{\mathcal{J}}\right)^{\frac{1}{2}}\end{pmatrix} V_{\mathcal{J}}B1_J - \begin{pmatrix} y \\ \left(R^\dagger_{\mathcal{J}}\right)^{\frac{1}{2}}\mu\end{pmatrix}\right)\\
&= \begin{pmatrix} I_m & -AR_{\mathcal{J}}^{\frac{1}{2}} \\ 0_{1\times m} & \beta^T \\ 0_{(n-J)\times m} & V_{\mathcal{J}^c}^T\end{pmatrix}\left(\frac{1}{J}\begin{pmatrix} AV_{\mathcal{J}}B1_J \\ \left(R^\dagger_{\mathcal{J}}\right)^{\frac{1}{2}}V_{\mathcal{J}}B1_J\end{pmatrix} - \begin{pmatrix} y \\ \left(R^\dagger_{\mathcal{J}}\right)^{\frac{1}{2}}\mu\end{pmatrix}\right)\\
&= \begin{pmatrix} 0_{m\times 1} \\ \frac{1}{J}\beta^T\left(R^\dagger_{\mathcal{J}}\right)^{\frac{1}{2}}V_{\mathcal{J}}B1_J \\ 0_{(n-J)\times 1}\end{pmatrix} - \begin{pmatrix} y-AV_{\mathcal{J}}V_{\mathcal{J}}^T\mu \\ \beta^T\left(R^\dagger_{\mathcal{J}}\right)^{\frac{1}{2}}\mu \\ 0_{(n-J)\times 1}\end{pmatrix} = \begin{pmatrix} 0_{m\times 1} \\ 1 \\ 0_{(n-J)\times 1}\end{pmatrix} - \begin{pmatrix} y-AV_{\mathcal{J}}V_{\mathcal{J}}^T\mu \\ 1_J^TB^{-1}V_{\mathcal{J}}^T\mu \\ 0_{(n-J)\times 1}\end{pmatrix}\\
&= \begin{pmatrix} -\left(y-AV_{\mathcal{J}}V_{\mathcal{J}}^T\mu\right) \\ 1-1_J^TB^{-1}V_{\mathcal{J}}^T\mu \\ 0_{(n-J)\times 1}\end{pmatrix} =: \begin{pmatrix} -z_{\mathcal{J}} \\ \gamma \\ 0_{(n-J)\times 1}\end{pmatrix}.
\end{aligned}
$$

Now we turn our attention to $(P^TP)^{-1}$:

$$
\begin{aligned}
\left(P^TP\right)^{-1} &= \begin{pmatrix} I_m + AR_{\mathcal{J}}A^T & -AR_{\mathcal{J}}^{\frac{1}{2}}\beta & 0_{m\times(n-J)} \\ -\beta^TR_{\mathcal{J}}^{\frac{1}{2}}A^T & \beta^T\beta & 0_{1\times(n-J)} \\ 0_{(n-J)\times m} & 0_{(n-J)\times 1} & I_{(n-J)}\end{pmatrix}^{-1}\\
&=: \begin{pmatrix} Q_{\mathcal{J}}^{-1} & -\alpha & 0_{m\times(n-J)} \\ -\alpha & \beta^T\beta & 0_{1\times(n-J)} \\ 0_{(n-J)\times m} & 0_{(n-J)\times 1} & I_{(n-J)}\end{pmatrix}^{-1}\\
&= \frac{1}{\Delta}\begin{pmatrix} \Delta Q_{\mathcal{J}} + Q_{\mathcal{J}}\alpha\left(Q_{\mathcal{J}}\alpha\right)^T & Q_{\mathcal{J}}\alpha & 0_{m\times(n-J)} \\ \left(Q_{\mathcal{J}}\alpha\right)^T & 1 & 0_{1\times(n-J)} \\ 0_{(n-J)\times m} & 0_{(n-J)\times 1} & \Delta I_{(n-J)}\end{pmatrix},
\end{aligned}
$$

where $\Delta := \beta^T\beta - \alpha^TQ_{\mathcal{J}}\alpha$. Therefore

$$
\begin{aligned}
&|Proj_{\ker\left(\tilde{A}C_0\tilde{A}^T\right)}\left(\tilde{A}\bar{u}_{\mathcal{J},B}(0)-\tilde{y}\right)|^2 = \left(P^T\left(\tilde{A}\bar{u}_{\mathcal{J},B}(0)-\tilde{y}\right)\right)^T\left(P^TP\right)^{-1}P^T\left(\tilde{A}\bar{u}_{\mathcal{J},B}(0)-\tilde{y}\right)\\
&= \frac{1}{\Delta}\begin{pmatrix} -z_{\mathcal{J}}^T & \gamma & 0_{1\times(n-J)}\end{pmatrix}\begin{pmatrix} \Delta Q_{\mathcal{J}} + Q_{\mathcal{J}}\alpha\left(Q_{\mathcal{J}}\alpha\right)^T & Q_{\mathcal{J}}\alpha & 0_{m\times(n-J)} \\ \left(Q_{\mathcal{J}}\alpha\right)^T & 1 & 0_{1\times(n-J)} \\ 0_{(n-J)\times m} & 0_{(n-J)\times 1} & \Delta I_{(n-J)}\end{pmatrix}\begin{pmatrix} -z_{\mathcal{J}} \\ \gamma \\ 0_{(n-J)\times 1}\end{pmatrix}\\
&= z_{\mathcal{J}}^TQ_{\mathcal{J}}z_{\mathcal{J}} + \frac{1}{\Delta}\left(\alpha^TQ_{\mathcal{J}}z_{\mathcal{J}}-\gamma\right)^2.
\end{aligned}
$$

Recall $\beta := V_{\mathcal{J}}\Lambda_{\mathcal{J}}^{\frac{1}{2}}B^{-T}1_J$, $\alpha := AR_{\mathcal{J}}^{\frac{1}{2}}\beta = AV_{\mathcal{J}}\Lambda_{\mathcal{J}}B^{-T}1_J$, $z_{\mathcal{J}} := y - AV_{\mathcal{J}}V_{\mathcal{J}}^T\mu$, $\gamma := 1 - 1_J^TB^{-1}V_{\mathcal{J}}^T\mu$ and $\Delta := \beta^T\beta - \alpha^TQ_{\mathcal{J}}\alpha$. Defining $A_{\mathcal{J}} := AV_{\mathcal{J}}$ for convenience, we see that

$$
\begin{aligned}
&\lim_{t\to\infty}\left(|A\bar{u}_{\mathcal{J},B}(t)-y|^2 + |\bar{u}_{\mathcal{J},B}(t)-\mu|_R^2\right) = \lim_{t\to\infty}\left|\tilde{A}\bar{u}_{\mathcal{J},B}(t)-\tilde{y}\right|^2 + |\mu|^2_{R^\dagger_{\mathcal{J}^c}}\\
&= \left|Proj_{\ker\left(\tilde{A}C_0\tilde{A}^T\right)}\left(\tilde{A}\bar{u}_{\mathcal{J},B}(0)-\tilde{y}\right)\right|^2 + |\mu|^2_{R^\dagger_{\mathcal{J}^c}}
\end{aligned}
$$

$$= z_{\mathcal{J}}^T Q_{\mathcal{J}} z_{\mathcal{J}} + \frac{1}{\Delta}\left(\alpha^T Q_{\mathcal{J}} z_{\mathcal{J}} - \gamma\right)^2 + |\mu|^2_{R^\dagger_{\mathcal{J}^c}}$$
$$= z_{\mathcal{J}}^T Q_{\mathcal{J}} z_{\mathcal{J}} + \frac{\left(1 - 1_J^T B^{-1}\left(V_{\mathcal{J}}^T \mu + \Lambda_{\mathcal{J}} A_{\mathcal{J}}^T Q_{\mathcal{J}} z_{\mathcal{J}}\right)\right)^2}{1_J^T B^{-1}\Lambda_{\mathcal{J}} B^{-T} 1_J - \left(A_{\mathcal{J}}\Lambda_{\mathcal{J}} B^{-T} 1_J\right)^T Q_{\mathcal{J}} A_{\mathcal{J}}\Lambda_{\mathcal{J}} B^{-T} 1_J} + |\mu|^2_{R^\dagger_{\mathcal{J}^c}}.$$

To obtain the second equality of the theorem, we apply the Woodbury matrix identity:

$$Q_{\mathcal{J}} = \left(I_m + A_{\mathcal{J}}\Lambda_{\mathcal{J}} A_{\mathcal{J}}^T\right)^{-1} = I_m - A_{\mathcal{J}} M_{\mathcal{J}} A_{\mathcal{J}}^T,$$

where $M_{\mathcal{J}} := (\Lambda_{\mathcal{J}}^{-1} + A_{\mathcal{J}}^T A_{\mathcal{J}})^{-1}$. It can easily be shown that

$$Q_{\mathcal{J}} A_{\mathcal{J}}\Lambda_{\mathcal{J}} = A_{\mathcal{J}} M_{\mathcal{J}},$$
$$\left(A_{\mathcal{J}}\Lambda_{\mathcal{J}}\right)^T Q_{\mathcal{J}} A_{\mathcal{J}}\Lambda_{\mathcal{J}} = \Lambda_{\mathcal{J}}^{-1} - M_{\mathcal{J}}.$$

As such

$$1_J^T B^{-1}\Lambda_{\mathcal{J}} B^{-T} 1_J - \left(A_{\mathcal{J}}\Lambda_{\mathcal{J}} B^{-T} 1_J\right)^T Q_{\mathcal{J}} A_{\mathcal{J}}\Lambda_{\mathcal{J}} B^{-T} 1_J = 1_J^T B^{-1} M_{\mathcal{J}} B^{-T} 1_J.$$

Then

$$z_{\mathcal{J}}^T Q_{\mathcal{J}} z_{\mathcal{J}} + \frac{\left(1 - 1_J^T B^{-1}\left(V_{\mathcal{J}}^T \mu + \Lambda_{\mathcal{J}} A_{\mathcal{J}}^T Q_{\mathcal{J}} z_{\mathcal{J}}\right)\right)^2}{1_J^T B^{-1}\Lambda_{\mathcal{J}} B^{-T} 1_J - \left(A_{\mathcal{J}}\Lambda_{\mathcal{J}} B^{-T} 1_J\right)^T Q_{\mathcal{J}} A_{\mathcal{J}}\Lambda_{\mathcal{J}} B^{-T} 1_J}$$
$$= z_{\mathcal{J}}^T z_{\mathcal{J}} - z_{\mathcal{J}}^T A_{\mathcal{J}} M_{\mathcal{J}} A_{\mathcal{J}}^T z_{\mathcal{J}} + \frac{\left(1 - 1_J^T B^{-1}\left(V_{\mathcal{J}}^T \mu + M_{\mathcal{J}} A_{\mathcal{J}}^T z_{\mathcal{J}}\right)\right)^2}{1_J^T B^{-1} M_{\mathcal{J}} B^{-T} 1_J}$$
$$= z_{\mathcal{J}}^T z_{\mathcal{J}} - z_{\mathcal{J}}^T A_{\mathcal{J}} M_{\mathcal{J}} A_{\mathcal{J}}^T z_{\mathcal{J}} + \frac{\left(1 - 1_J^T B^{-1} M_{\mathcal{J}} V_{\mathcal{J}}^T\left(A^T y + R^\dagger_{\mathcal{J}}\mu\right)\right)^2}{1_J^T B^{-1} M_{\mathcal{J}} B^{-T} 1_J},$$

where, for the last line, we have used

$$V_{\mathcal{J}}^T \mu + M_{\mathcal{J}} A_{\mathcal{J}}^T z_{\mathcal{J}} = V_{\mathcal{J}}^T \mu + M_{\mathcal{J}} A_{\mathcal{J}}^T y - M_{\mathcal{J}} A_{\mathcal{J}}^T A_{\mathcal{J}}^T{}_{\mathcal{J}}^T \mu$$
$$= V_{\mathcal{J}}^T \mu + M_{\mathcal{J}} A_{\mathcal{J}}^T y - \left(I - M_{\mathcal{J}}\Lambda_{\mathcal{J}}^{-1}\right) V_{\mathcal{J}}^T \mu = M_{\mathcal{J}} V_{\mathcal{J}}^T\left(A^T y + R^\dagger_{\mathcal{J}}\mu\right).$$

□

Proof of corollary 4.10. The linear least squared problem

$$\min_{u \in Im(V_{\mathcal{J}})} |Au - y|^2 + |u - \mu|_R = \min_x |AV_{\mathcal{J}} x - y|^2 + |V_{\mathcal{J}} x - \mu|^2$$

has solution

$$x = M_{\mathcal{J}} V_{\mathcal{J}}^T\left(A^T y + R^\dagger_{\mathcal{J}}\mu\right).$$

As such, if $B^* \in \mathbb{R}^{J\times J}$ satisfies $B^* 1_J/J = M_{\mathcal{J}} V_{\mathcal{J}}^T(A^T y + R^\dagger_{\mathcal{J}}\mu)$, then it is clear that

$$|A\bar{u}_{\mathcal{J},\mathcal{B}^*}(0) - y|^2 + |\bar{u}_{\mathcal{J},\mathcal{B}^*}(0) - \mu|^2 = \min_{u \in Im(V_{\mathcal{J}})} |Au - y|^2 + |u - \mu|^2_R. \tag{B.1}$$

To get the final result, we observe from corollary 4.3 that $|A\bar{u}_{\mathcal{J},\mathcal{B}^*}(t) - y|^2 + |\bar{u}_{\mathcal{J},\mathcal{B}^*}(t) - \mu|^2$ is monotonic non-increasing and, by the subspace property, bounded below by the right-hand side of equation (B.1). □

Proof of lemma 4.11. Recall that $x \in \mathbb{R}^m$, $P_k := [p_1, \dots, p_k] \in \mathbb{R}^{m\times k}$, $\tilde{M}_k := (I_k + P_k^T P_k)^{-1} \in \mathbb{R}^{k\times k}$ and $\Delta_{k+1} := 1 + p_{k+1}^T p_{k+1} - p_{k+1}^T P_k \tilde{M}_k P_k^T p_{k+1}$. Then

$$\begin{aligned}
\tilde{M}_{k+1} &= \begin{pmatrix} I_k + P_k^T P_k & P_k^T p_{k+1} \\ p_{k+1}^T P_k & 1 + p_{k+1}^T p_{k+1} \end{pmatrix}^{-1} \\
&= \begin{pmatrix} \tilde{M}_k^{-1} & P_k^T p_{k+1} \\ p_{k+1}^T P_k & 1 + p_{k+1}^T p_{k+1} \end{pmatrix}^{-1} \\
&= \frac{1}{\Delta_{k+1}} \begin{pmatrix} \Delta_{k+1}\tilde{M}_k + \left(M_k P_k^T p_{k+1}\right)\left(\tilde{M}_k P_k^T p_{k+1}\right)^T & -\tilde{M}_k P_k^T p_{k+1} \\ -p_{k+1}^T P_k \tilde{M}_k & 1 \end{pmatrix},
\end{aligned}$$

where the last equality is due to block-wise matrix inversion. Then

$$\begin{aligned}
&x^T P_{k+1} M_{k+1} P_{k+1}^T x \\
&= \frac{1}{\Delta_{k+1}} \begin{pmatrix} P_k^T x \\ p_{k+1}^T x \end{pmatrix}^T \begin{pmatrix} \Delta_{k+1}\tilde{M}_k + \left(\tilde{M}_k P_k^T p_{k+1}\right)\left(\tilde{M}_k P_k^T p_{k+1}\right)^T & -\tilde{M}_k P_k^T p_{k+1} \\ -p_{k+1}^T P_k \tilde{M}_k & 1 \end{pmatrix} \begin{pmatrix} P_k^T x \\ p_{k+1}^T x \end{pmatrix} \\
&= x^T P_k \tilde{M}_k P_k^T x + \frac{1}{\Delta_{k+1}} \left(x^T P_k \tilde{M}_k P_k^T p_{k+1} - x^T p_{k+1}\right)^2.
\end{aligned}$$

□

Proof of proposition 4.12. Recall that $R = V\Lambda V^T$, A has a singular value decomposition $A = \mathcal{U}\Sigma V^T$ and $V_{\mathcal{J}} := [v_{\mathcal{J}_1}, \dots, v_{\mathcal{J}_J}] \in \mathbb{R}^{n\times J}$. We further define $\mathcal{U}_{\mathcal{J}} := [u_{\mathcal{J}_1}, \dots, u_{\mathcal{J}_J}] \in \mathbb{R}^{n\times J}$ and $\Sigma_{\mathcal{J}} := \mathrm{diag}(\{\sigma_{\mathcal{J}_1}, \dots, \sigma_{\mathcal{J}_J}\})$. Then

$$\begin{aligned}
A_{\mathcal{J}} &= AV_{\mathcal{J}} = \mathcal{U}_{\mathcal{J}}\Sigma_{\mathcal{J}}, \\
z_{\mathcal{J}} &= y - AV_{\mathcal{J}}V_{\mathcal{J}}^T\mu = y - U_{\mathcal{J}}\Sigma_{\mathcal{J}}V_{\mathcal{J}}^T\mu, \\
A_{\mathcal{J}}^T z_{\mathcal{J}} &= \Sigma_{\mathcal{J}}\mathcal{U}_{\mathcal{J}}^T y - \Sigma_{\mathcal{J}}^2 V_{\mathcal{J}}^T\mu, \\
\mathcal{M}_{\mathcal{J}} &= \left(\Lambda_{\mathcal{J}}^{-1} + A_{\mathcal{J}}^T A_{\mathcal{J}}\right)^{-1} = \left(\Lambda_{\mathcal{J}}^{-1} + \Sigma_{\mathcal{J}}^2\right)^{-1}.
\end{aligned}$$

It is then clear that

$$\begin{aligned}
z_{\mathcal{J}}^T z_{\mathcal{J}} - z_{\mathcal{J}}^T A_{\mathcal{J}} \mathcal{M}_{\mathcal{J}} A_{\mathcal{J}}^T z_{\mathcal{J}} - |\mu|_{R_{\mathcal{J}}^\dagger}^2 =& y^T y + \sum_{k\in\mathcal{J}} \Bigg(-2\sigma_k \left(y^T u_k\right)\left(\mu^T v_k\right) + \sigma_k^2 \left(\mu^T v_k\right)^2 \\
&- \frac{\lambda_k \left(\sigma_k \left(y^T u_k\right) - \sigma_k^2 \left(\mu^T v_k\right)\right)^2}{1 + \lambda_k \sigma_k^2} - \frac{\left(\mu^T v_k\right)^2}{\lambda_k} \Bigg). \qquad \text{(B.2)}
\end{aligned}$$

Examining algorithm Greedy, we see that

$$p_k := \lambda_k^{\frac{1}{2}} A v_k = \lambda_k^{\frac{1}{2}} \mathcal{U}\Sigma V^T v_k = \lambda_k^{\frac{1}{2}} \sigma_k u_k.$$

Since the $\{u_k\}_k$ are orthogonal, $\{p_k\}_k$ are also orthogonal. Since $j^* \in \mathcal{J}_k^c$, at every stage of the algorithm, we have that

$$\begin{aligned}
P^T p_j &= 0, \\
z^T u_j &= \left(y - U_{\mathcal{J}}\Sigma_{\mathcal{J}}V_{\mathcal{J}}^T e\mu\right)^T u_j = y^T u_j,
\end{aligned}$$

$$z^T A v_j = \sigma_j \left(z^T u_j\right) = \sigma_j \left(y^T u_j\right),$$
$$\tilde{z}^T p_j = \left(z - A v_j v_j^T \mu\right)^T \left(\lambda_j^{\frac{1}{2}} \sigma_j u_j\right) = \lambda_j^{\frac{1}{2}} \sigma_j \left(y^T u_j - \sigma_j \left(\mu^T v_j\right)\right).$$

This yields

$$-2\left(\mu^T v_j\right)\left(z^T A v_j\right) + \left(\mu^T v_j\right)^2 \left(|A v_j|^2 - \frac{1}{\lambda_j}\right) - \frac{\left(\tilde{z}^T P \tilde{M} P^T p_j - \tilde{z}^T p_j\right)^2}{1 + |p_j|^2 - p_j^T P \tilde{M} P^T p_j} =$$
$$-2\sigma_j \left(\mu^T v_j\right)\left(y^T u_j\right) + \left(\mu^T v_j\right)^2 \left(\sigma_j^2 - \frac{1}{\lambda_j}\right) - \frac{\lambda_j \sigma_j^2 \left(y^T u_j - \sigma_j \left(\mu^T v_j\right)\right)^2}{1 + \lambda_j \sigma_j^2}. \tag{B.3}$$

The RHS of this equation coincides with the summand in equation (B.2). As such, by choosing j at every stage of algorithm Greedy to maximise the RHS of equation (B.3), we end up maximising the objective function given by the LHS of equation (B.2).

□

ORCID iDs

Ruben Harris 0009-0004-9327-9369
Claudia Schillings 0009-0003-7121-2457

References

Bergemann K and Reich S 2009 A localization technique for ensemble Kalman filters *Q. J. E. Meteorol. Soc.* **136** 701–7
Bergemann K and Reich S 2010 A mollified ensemble Kalman filter *Q. J. R. Meteorol. Soc.* **136** 1636–43
Blömker D, Schillings C and Wacker P 2018 A strongly convergent numerical scheme from ensemble Kalman inversion *SIAM J. Numer. Anal.* **56** 2537–62
Blömker D, Schillings C, Wacker P and Weissmann S 2019 Well posedness and convergence analysis of the ensemble Kalman inversion *Inverse Problems* **35** 085007
Blömker D, Schillings C, Wacker P and Weissmann S 2021 Continuous time limit of the stochastic ensemble Kalman inversion: strong convergence analysis *SIAM J. Numer. Anal.* **60** 3181–215
Bungert L and Wacker P 2021 Complete deterministic dynamics and spectral decomposition of the ensemble kalman inversion (arXiv:2104.13281)
Calvello E, Reich S and Stuart A M 2025 Ensemble Kalman methods: a mean field perspective *Acta Numerica* **34** 123–291
Chada N K, Chen Y and Sanz-Alonso D 2021 Iterative ensemble Kalman methods: A unified perspective with some new variants *Found. Data Sci.* **3** 331–69
Chada N K, Stuart A M and Tong X T 2020 Tikhonov regularization within ensemble Kalman inversion *SIAM J. Numer. Anal.* **58** 1263–94
Chen Y and Oliver D S 2012 Ensemble randomized maximum likelihood method as an iterative ensemble smoother *Math. Geosci.* **44** 1–26
Chen Y, Oliver D S and Zhang D 2009 Efficient ensemble-based closed-loop production optimization *SPE J.* **14** 634–45
Cui T, Marzouk Y M and Willcox K E 2014 Data-driven model reduction for Bayesian inversion *SIAM J. Sci. Comput.* **36** A2323–56
Dashti M, Law K J H, Stuart A M and Voss J 2013 MAP estimators and their consistency in Bayesian nonparametric inverse problems *Inverse Problems* **29** 095017
Ding Z and Li Q 2020 Ensemble Kalman inversion: mean-field limit and convergence analysis *Staist. Comput.* **31** 9
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

Evans L C 2010 *Partial Differential Equations* (American Mathematical Society)
Evensen G 2003 The Ensemble Kalman Filter: theoretical formulation and practical implementation *Ocean Dyn.* **53** 343–67
Evensen G 2009 *Data Assimilation: The Ensemble Kalman Filter* 1st edn (Springer)
Evensen G and van Leeuwen P J 2000 An ensemble Kalman smoother for nonlinear dynamics *Mon. Weather Rev.* **128** 1852–67
Ghattas O A and Sanz-Alonso D 2022 Non-asymptotic analysis of ensemble kalman updates: effective dimension and localization (arXiv:2208.03246)
Iglesias M A 2014 Iterative regularization for ensemble data assimilation in reservoir models *Comput. Geosci.* **19** 177–212
Iglesias M A 2016a A regularizing iterative ensemble Kalman method for PDE-constrained inverse problems *Inverse Problems* **32** 025002
Iglesias M A, Law K J H and Stuart A M 2013 Ensemble Kalman methods for inverse problems *Inverse Problems* **29** 045001
Iglesias M and Yang Y 2021 Adaptive regularisation for ensemble Kalman inversion *Inverse Problems* **37** 025008
Klebanov I and Wacker P 2023 Maximum a posteriori estimators in ℓ^p are well-defined for diagonal Gaussian priors *Inverse Problems* **39** 065009
Lange T 2021 Derivation of ensemble Kalman–Bucy filters with unbounded nonlinear coefficients *Nonlinearity* **35** 1061
Li G and Reynolds A 2009 Iterative ensemble Kalman filters for data assimilation *SPE J.* **14** 496–505
Parzer F and Scherzer O 2022 On convergence rates of adaptive ensemble Kalman inversion for linear ill-posed problems *Numer. Math.* **152** 371–409
Qian E and Beattie C 2025 The fundamental subspaces of Ensemble Kalman inversion (arXiv:2409.08862v2)
Schillings C and Stuart A 2017a Convergence analysis of ensemble Kalman inversion: the linear, noisy case *Appl. Anal.* **97** 487
Schillings C and Stuart A M 2017b Analysis of the ensemble Kalman filter for inverse problems *SIAM J. Numer. Anal.* **55** 1264–90
Spantini A, Bigoni D and Marzouk Y M 2015 Optimal low-rank approximations of Bayesian linear inverse problems *SIAM J. Sci. Comput.* **37** A2451–87
Stuart A M 2010 Inverse problems: A Bayesian perspective *Acta Numer.* **19** 451–559
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
Tong X T, Majda A J and Kelly D 2016 Nonlinear stability and ergodicity of ensemble based Kalman filters *Nonlinearity* **29** 657–91
Vishny D, Morzfeld M, Gwirtz K, Bach E, Dunbar O R A and Hodyss D 2024 High-dimensional covariance estimation from a small number of samples *J. Adv. Model. Earth Syst.* **16** e2024MS004417
Weissmann S 2022 Gradient flow structure and convergence analysis of the ensemble Kalman inversion for nonlinear forward models *Inverse Problems* **38** 105011
Weissmann S, Chada N K, Schillings C and Tong X T 2022 Adaptive Tikhonov strategies for stochastic ensemble Kalman inversion *Inverse Problems* **38** 045009