Ensemble Kalman Inversion: Regularization, Acceleration and Localization

Xin T. Tong (NUS)

Andrew Stuart (Caltech) Neil Chada (KAUST)
Matthias Morzfeld (UCSD)

Roadmap

- Review on EKI.
- Tikhonov regularization.
- Acceleration with varying step size.
- Localization with EKI.

Inverse problems & EKI

Data assimilation problem

Estimate u_n using $y_1, \ldots, y_n$

$$u_n = F(u_{n-1}), \quad y_n = G(u_n) + \xi_n$$

Ensemble Kalman filter yields good result

- Both F and G can be nonlinear maps
- G are usually partial (linear) observations
- Particles size is small, around 20-100.
- Work well on high dimensional problems with localization

Math theory: need to relax more than one constraints.
Most difficult part: nonlinear forward F.

Inverse problem

Estimate u from noisy observations y in the form

$$y = G(u) + \eta, \quad \eta \sim \mathcal{N}(0, \Gamma).$$

- Examples: in medical imaging, geophysical sciences, numerical weather prediction, machine learning and oceanography.
- The forward map no longer exists.
- G is often complicated. Preferably treated as a black-box.
- Variational: $u^* := \text{argmin}_u l(u), l(u) = \frac{1}{2}\|y - G(u)\|_\Gamma^2$
  Notation: $\|y - G(u)\|_\Gamma^2 = (y - G(u))^T \Gamma^{-1} (y - G(u))$
- Bayesian: Sample the posterior $p(u|y) \propto p(y|u)p_0(u)$

Ensemble Kalman inversion (EKI)

Ensemble Kalman methods for inverse problems. [Iglesias, Law, Stuart 13].

Initialization: draw $u_0^{(j)}$ based on some rules.

Prediction step

$$\bar{u}_n = \frac{1}{J}\sum_{j=1}^{J} u_n^{(j)}, \quad \bar{G}_n = \frac{1}{J}\sum_{j=1}^{J} G(u_n^{(j)}),$$

$$C_n^{pp} = \frac{1}{J-1}\sum_{j=1}^{J}(G(u_n^{(j)}) - \bar{G}_n)(G(u_n^{(j)}) - \bar{G}_n)^T$$

$$C_n^{up} = \frac{1}{J-1}\sum_{j=1}^{J}(u_n^{(j)} - \bar{u})(G(u_n^{(j)}) - \bar{G}_n)^T.$$

Update step

$$u_{n+1}^{(j)} = u_n^{(j)} + C_n^{up}(C_n^{pp} + \Gamma)^{-1}(y + \eta_{n+1}^{(j)} - G(u_n^{(j)})).$$

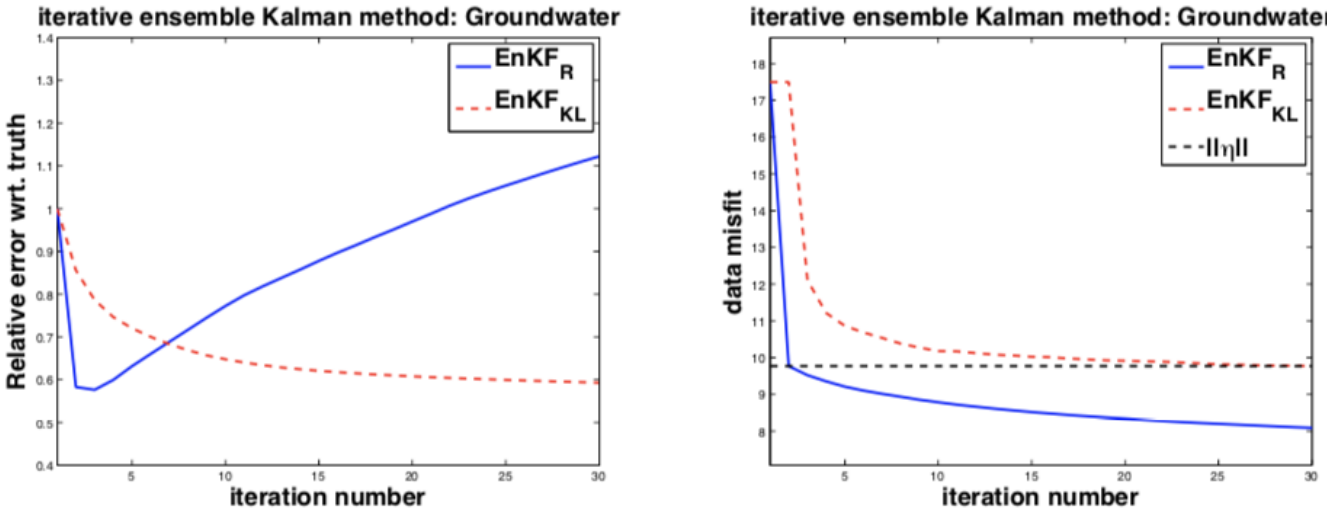
Iterate until ensemble collapse.

Tikhonov regularization

Overfitting

EKI minimizes: $\frac{1}{2}\|y - G(u)\|_\Gamma^2$.

- *Iterative regularization*:
  Levenberg-Marquardt regularization [Hanke 94, Iglesias 16].
- *Hierarchical approaches*: [Chada 18].
- *Connection with sequential Monte Carlo*:
  Tempering and gradient flow structure [Iglesias 18, Schillings 17].



From [ILS13]

Aim: Solve for u^* where

$$u^* := \text{argmin}_u \frac{1}{2}\|y - G(u)\|_\Gamma^2 + R(u). \tag{1}$$

- *Aids with*: stability, preventing overfitting.
- *Various forms*: Tikhonov, Levenberg-Maruqardt, LASSO, Landweber [Burger 18].

Tikhonov regularization is associated with choosing

$$R(u) = \frac{1}{2}\|u\|_\Sigma^2, \quad E \subset X.$$

Question:How to incorporate this additional term?

Expanding the system

Artificial observation

$$y = G(u) + \eta,$$
$$0 = u + \zeta,$$

where $\eta \sim N(0, \Gamma), \zeta \sim N(0, \Sigma)$.

Introduce z, η and mapping H as follows:

$$z = \begin{bmatrix} y \\ 0 \end{bmatrix}, \quad H(u) = \begin{bmatrix} G(u) \\ u \end{bmatrix}, \quad \xi = \begin{bmatrix} \eta \\ \zeta \end{bmatrix},$$

and

$$\xi \sim N(0, \Gamma_+), \quad \Gamma_+ = \begin{bmatrix} \Gamma & 0 \\ 0 & \Sigma \end{bmatrix}.$$

Consider the inverse problem

$$z = H(u) + \xi.$$

Optimization, minimize

$$l(u) = \|H(u) - z\|^2_{\Gamma_+} = \|G(u) - y\|^2_{\Gamma} + \|u\|^2_{\Sigma}.$$

Initialization: draw $u_0^{(j)}$ based on some rules.

Prediction step

$$\bar{u}_n = \frac{1}{J}\sum_{j=1}^{J} u_n^{(j)}, \quad \bar{F}_n = \frac{1}{J}\sum_{j=1}^{J} F(u_n^{(j)}) = \begin{bmatrix} \bar{G}_n \\ \bar{u}_n \end{bmatrix}$$

$$B_n^{pp} = \begin{bmatrix} C_n^{pp} & C_n^{pu} \\ C_n^{up} & C_n^{uu} \end{bmatrix}, \quad B_n^{up} = \begin{bmatrix} C_n^{pu} \\ C_n^{uu} \end{bmatrix}.$$

$$C_n^{uu} = \mathrm{Cov}(u_n^{(\cdot)}), \quad C_n^{up} = \mathrm{Cov}(u_n^{(\cdot)}, G(u_n^{(\cdot)})), \quad C_n^{pp} = \mathrm{Cov}(G(u_n^{(\cdot)}))$$

Update step

$$\begin{aligned} u_{n+1}^{(j)} &= u_n^{(j)} + B_n^{up}(B_n^{pp} + \Gamma_+)^{-1} \left(\begin{bmatrix} y \\ 0 \end{bmatrix} + \begin{bmatrix} \eta_{n+1}^{(j)} \\ \zeta_{n+1}^{(j)} \end{bmatrix} - \begin{bmatrix} G_n^{(j)} \\ u_n^{(j)} \end{bmatrix} \right). \\ &= u_n^{(j)} + B_n^{up}(B_n^{pp} + \Gamma_+)^{-1} \left(z + \xi_{n+1}^{(j)} - H_n^{(j)} \right). \end{aligned}$$

Stepsize: $u_{n+1}^{(j)} = u_n^{(j)} + B_n^{up}(B_n^{pp} + h^{-1}\Gamma_+)^{-1}\left(z + \xi_{n+1}^{(j)} - H_n^{(j)}\right)$.

Continuous time limit: $h \to 0$ and ignore noise term

$$\frac{d}{dt}u_t^{(j)} = B_t^{up}\Gamma_+^{-1}u_t^{(j)}(z - H(u_t^{(j)})).$$

Subspace [Chada, Stuart, T. 19]

Let $\mathcal{B} := \bar{u}_0 + \text{span}\{u_0^{(j)} - \bar{u}_0\}_{j=1}^J$. Then $u_t^{(j)}$ has a unique solution in $C^1([0,T),\mathcal{B})$ for some $T > 0$.

Interpretation: TEKI ensemble stay in a subspace.

Ensemble Collapse

Ensemble collapse [Chada, Stuart, T. 19]

The following upper bound holds:

$$\|C_t^{uu}\| \leq \frac{1}{\|C_0^{uu}\|^{-1} + 2\lambda_m t},$$

where

$$\lambda_m := \inf_{v \in \mathcal{B}, \|v\|^2 = 1} \|\Sigma^{-1/2} v\|^2.$$

Interpretation: TEKI ensemble collapses, with rate least of $O(1/t)$.

Long term analysis

- What does TEKI converges to?
- Best scenario

$$u_t^{(j)} \to u_{\mathcal{B}}^{\dagger} := \mathrm{argmin}_{u\in\mathcal{B}} \|u\|_{\Sigma}^2 + \|\Gamma^{-1/2}(G(u)-y)\|^2.$$

Theorem [Chada, Stuart, T. 19]

If obs are linear, $G(u) = Gu$, $O(\frac{1}{t})$ convergence

$$\|u_t^{(j)} - u_{\mathcal{B}}^{\dagger}\|_{\Gamma_+}^2 \leq \frac{m_0}{1+2m_0 t}\|u_0^{(j)} - u_{\mathcal{B}}^{\dagger}\|_{\Gamma_+}^2.$$

where

$$m_0 := \min_{u\in\mathcal{B},\|u\|=1} \langle u, C_0^{uu}\Gamma_+^{-1}u\rangle.$$

Inverse eikonal equation

Given a slowness or inverse velocity function $s(x) \in C^0(\bar{\mathcal{D}})$, characterizing the medium, and a source location $x_0 \in \mathcal{D}$, the forward eikonal equation is to solve for travel time $T(x) \in C^0(\bar{\mathcal{D}})$ satisfying

$$\begin{aligned} |\nabla T(x)| &= s(x), \quad x \in \mathcal{D} \setminus \{x_0\}, \\ T(x_0) &= 0, \\ \nabla T(x) \cdot \nu(x) &\geq 0, \quad x \in \partial\mathcal{D}. \end{aligned}$$

Data takes the form

$$y_j = l_j(T^\dagger) + \eta_j, \quad j = 1, \cdots, J,$$

Aim: The recovery of the slowness function $s(x)$

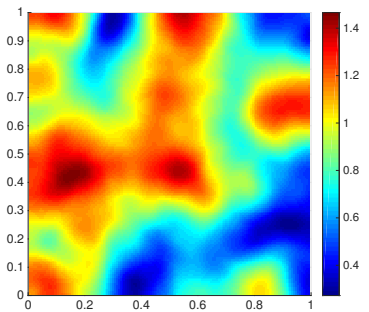


Figure: True random field: $\alpha = 3.2, a = 1$.

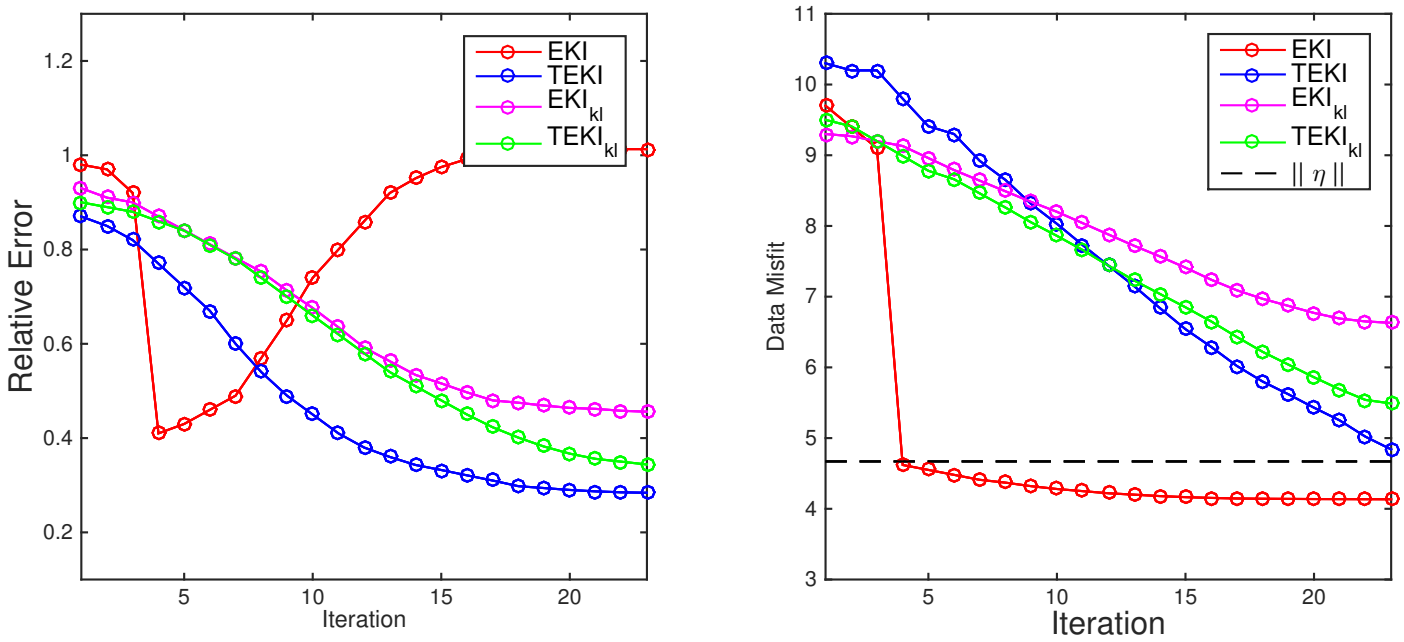


Figure: Relative errors and data misfits.

Acceleration

Temporal discretization

TEKI implementation:

$$u_{n+1}^{(j)} = u_n^{(j)} + B_n^{up}(B_n^{pp} + h^{-1}\Gamma_+)^{-1}\left(z + \xi_{n+1}^{(j)} - H_n^{(j)}\right).$$

- When $h \to 0$, we have the continuous time limit.
- Not much understanding for the temporal discretized version.
- Most analyses focused on the case where H or G is linear.

Non-contant stepsize and covariance inflation

Non-constant Stepsize/learning rate:
Machine learning: convergence rate [Robbins Monro 57, Nesterov 13]

$$u_{n+1}^{(j)} = u_n^{(j)} + B_n^{up}(B_n^{pp} + h_n^{-1}\Gamma_+)^{-1}\left(z + \xi_{n+1}^{(j)} - H_n^{(j)}\right).$$

Covariance inflation and square root filter: improving stability
Machine learning: escape local minimum in Langevin dynamics.

$$m_{n+1} = m_n + B_n^{up}(B_n^{pp} + h_n^{-1}\Gamma_+)^{-1}\left(z - \bar{H}_n\right)$$
$$C_{n+1}^{uu} = C_n^{uu} - B_n^{up}(h_n^{-1}\Gamma_+ + B_n^{pp})^{-1}B_n^{pu} + \alpha_n^2\Sigma$$

Parametric setup for step size and inflation:

$$h_n = h_0 n^{\beta}, \quad \alpha_n^2 = \alpha_0^2 h_0^{-1} n^{2\gamma-\beta-2}.$$

The range: $0 \le \gamma < 1$ and $\beta \le \gamma \le \beta + 1$.

Controllability and observability

Ensemble collapses [Chada, T. 19]

The following upper and lower bound holds:

$$\kappa_m n^{\gamma-\beta-1}\Sigma \preceq C_n^{uu} \preceq \kappa_M n^{\gamma-\beta-1}\Sigma$$

where κ_m and κ_M are both constants

Interpretation: TEKI ensemble collapses at well control rate. Controllability+observability.

Connection with Gauss-Newton

Gauss-Newton (GN): iterative Kalman filter [Bell Cathey 93]

GN for likelihood $\ell_{n+1}(u) = \|u - m_n\|^2_{C_n^{uu}} + \|H(u) - z\|^2_{\Gamma_+ h_n^{-1}}$

$m_{n+1} = m_n + G_n, J_n := \nabla H(m_n)$

$G_n = C_n^{uu} J_n^T (J_n C_n^{uu} J_n^T + h_n^{-1}\Gamma_+)^{-1}(z - H(m_n))$

TEKI $m_{n+1} = m_n + \Delta_n, \Delta_n = C_n^{up}(C_n^{pp} + h_n^{-1}\Gamma_+)^{-1}(z - H(m_n))$

Proposition [Chada, T. 19]

The difference between two schemes are of lower order

$\|G_n - \Delta_n\| \leq M_3 h_n K \|C_n^{uu}\|^{\frac{3}{2}} \|z - H(m_n)\| = O(n^{\frac{3}{2}\gamma - \frac{3}{2} - \frac{1}{2}\beta})$.

Strongly covex: converge to the global minimizer of ℓ

Theorem [Chada, T. 19]

Suppose there is a $\lambda_c > 0$ such that for any vectors x, y

$$\ell(x) - \ell(y) \geq \langle \nabla \ell(y), x - y \rangle + \lambda_c \|x - y\|^2.$$

Exist n_0 and D so for $N \geq n_0$:

1) If $\gamma = 0$, for any $\alpha < \min\{\frac{1}{2} + \frac{1}{2}\beta, h_0 \kappa_m \sigma_m \lambda_c\}$,

$$\lambda_c \|m_N - u^*\|^2 \leq \ell(m_N) - \ell(u^*) \leq \frac{D}{N^\alpha}.$$

2) If $1 > \gamma > 0$, for any $\alpha < \frac{1}{2} + \frac{1}{2}\beta - \frac{1}{2}\gamma$,

$$\lambda_c \|m_N - u^*\|^2 \leq \ell(m_N) - \ell(u^*) \leq \frac{D}{N^\alpha}.$$

Non-covex: hitting critical points of ℓ

Theorem [Chada, T. 19]

Suppose that the EKI mean sequence m_n is bounded and the parameter $\gamma \in [0, 1)$, then or any $\epsilon > 0$, then

$$\min_{n \leq N_\epsilon} \{\|\nabla \ell(m_n)\|\} \leq \epsilon.$$

The threshold iteration is given by

$$N_\epsilon = \begin{cases} \exp(D/\epsilon^2) & \text{if} \quad \gamma = 0, \\ (D/\epsilon^2)^{\frac{2}{\min\{2\gamma, \beta+1-\gamma+\delta\}}} & \text{if} \quad 1 > \gamma > 0. \end{cases}$$

Localization

Large ensemble size

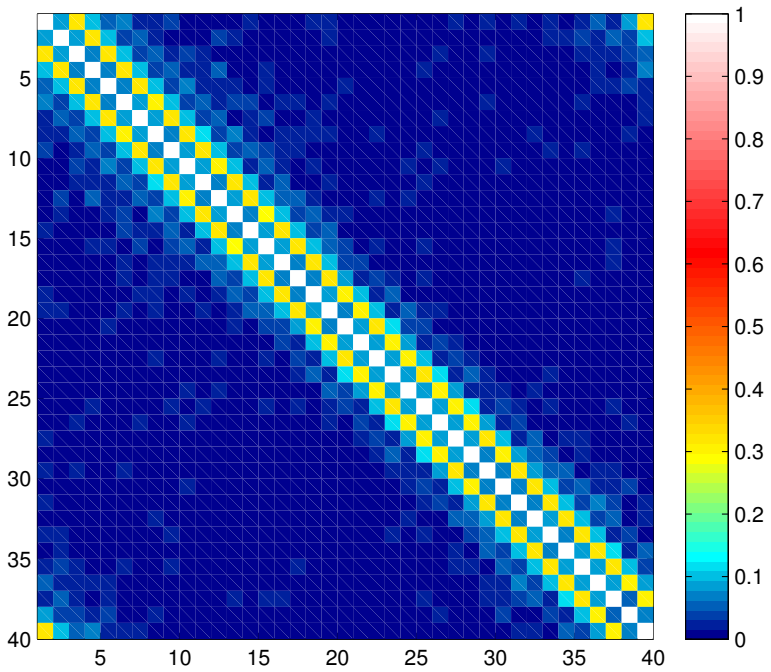
Subspace property

$$u_t^{(j)} \in \mathcal{B} = \text{span}\{u_0^{(j)}, j = 1, \ldots, J\}$$

- Can be exploited cleverly by starting with good smoothness
- To have $\mathcal{B}$ full space, we need $J > d$
- Not feasible in practice for high dimensional problems
- Consider the localization technique from EnKF
- Existing works: [Chen and Oliver (2017)].

Sparsity: local covariance

- Correlation decays quickly with the distance.
- Covariance is localized with a structure.



Correlation of Lorenz 96

- Implementation: Schur product with a mask

$$[C_n^{uu} \circ \mathbf{D}_L]_{i,j} = [C_n]_{i,j} \cdot [\mathbf{D}_L]_{i,j}$$

  Use $\widetilde{C}_n^{uu} = C_n^{uu} \circ \mathbf{D}_L$ to describe uncertainty
- $[\mathbf{D}_L]_{i,j} = \phi(|i-j|)$, with a radius L.
  Gaspari-Cohn matrix: $\phi(x) = \exp(-4x^2/L^2)\mathbf{1}_{|i-j|\le L}$.
- Also resolves rank deficiency, e.g.

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \circ \begin{bmatrix} 1 & 0.2 & 0 \\ 0.2 & 1 & 0.2 \\ 0 & 0.2 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.2 & 0 \\ 0.2 & 1 & 0.2 \\ 0 & 0.2 & 1 \end{bmatrix}.$$

Localizing the cross covariance

Continuous time limit

$$\frac{d}{dt}u_t^{(j)} = B_t^{up}\Gamma_+^{-1}u_t^{(j)}(z - H(u_t^{(j)})) + \lambda_t\xi^j(t).$$

- $B_t^{up} = \mathrm{cov}(u_t^{(j)}, H(u_t^{(j)}))$ sample covariance between state and observation
- There seems no explicit formula to localize B_t^{up}.
- Most schemes are described through heuristics

Two localization schemes

Suppose $H(u) = Hu$ for some matrix

- Naturally we use $\widetilde{B}_t^{up} = \widetilde{C}_t^{uu} H^T$.
- We call it linear localization scheme
- Nonlinear $H(u)$: use sensitivity or Jacobian matrix.

Suppose $H_i(u)$ concerns only u_k

- Naturally we localize $[\widetilde{B}_t^{up}]_{m,i} = \phi(|m-k|)\mathrm{cov}(u_m^{(j)}, H_i(u_k^{(j)}))$
- We call it centralized localization scheme
- More general: $H_i(u)$ concerns u_m near u_k.

Intuitive with DA background
Not enough for general inverse problems.

Theoretical results

Theorem [Morzfeld, Stuart, T. 21+]

Under mild conditions, the continuous time localized EKI flow satisfies

1) $\widetilde{C}_t^{uu} = \Omega(1/t)$ in both maximum and minimum eigenvalues.
2) If the loss function $l(u) = \|H(u) - z\|_{\Gamma^+}^2$ is strongly convex enough, $\bar{u}_t$ converges to the global min at rate $O(1/t)$ in l_2
3) If $l(u) = \sum l_i(u)$ with $l_i(u) = \|H_i(u) - y_i\|^2$ being strongly convex enough, $\bar{u}_t$ converges to the global min at rate $O(1/t)$ at all components.

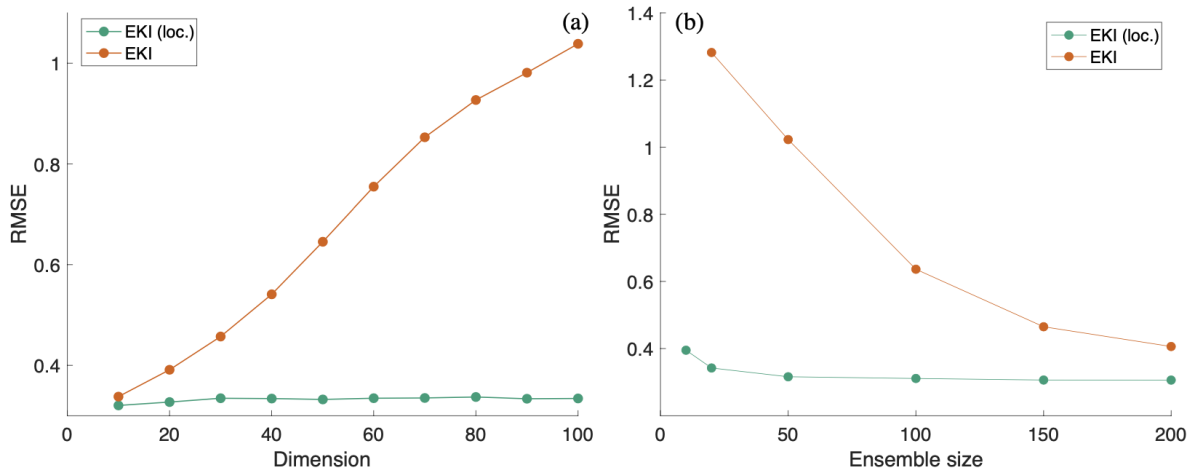
Main point: sample size J can be a fixed constant independent of d.

Nonlinear example

A nonlinear model for which $d_u = d_y = N$:

$$y_i = u_i - \sqrt{3}\hat{u}_i^2 + \hat{u}_i^3, \quad i = 1, \ldots, N,$$

where $\hat{u}_i = \frac{1}{10}\sum_{j=-5}^{5} u_{i-j}, 5 < i < N - 5$.
EKI and LEKI using centralized scheme with $J = 50/N = 100$.



Reference

- N. K. Chada, A. M. Stuart, X. T. Tong. Tikhonov regularization within ensemble Kalman inversion. arXiv:1901.10382, 2019
- N. K. Chada, X. T. Tong. Ensemble Kalman inversion as a derivative-free optimizer. (In preparation)

Slides can be found at sites.google.com/view/xintongthomson.

Thank you!