LUT University
School of Engineering Science
Computational Engineering and Technical Physics
Technomathematics

Harrison Mansour

ENSEMBLE KALMAN SAMPLER

Master's Thesis

Examiners: Professor Tapio Helin
Professor Heikki Haario

Supervisors: Professor Tapio Helin

ABSTRACT

LUT University
School of Engineering Science
Computational Engineering and Technical Physics
Technomathematics

Harrison Mansour

Ensemble Kalman Sampler

Master's Thesis

2020

42 pages.

Examiners: Professor Tapio Helin
Professor Heikki Haario

Keywords: Inverse Problems, Bayesian Inverse Problems, Ensemble Kalman Filter, Ensemble Kalman Inversion

Many modern inverse problems are based upon a very complex forward model that is computationally expensive. In such a setting, numerical estimation of a gradient can be instable or infeasible and derivative-free solution methods are preferred. This thesis studies the Ensemble Kalman Sampler (EKS), a novel algorithm that samples from the posterior of a Bayesian inverse problem using no information of the gradient. EKS is based upon Langevin dynamics and is a noisy variation of Ensemble Kalman Inversion (EKI). The new noise structure causes EKS to effectively sample from the posterior, instead of collapsing to a single optimal point like EKI. Two numerical results are presented to demonstrate the algorithm's effectiveness.

PREFACE

I thank my supervisor Tapio Helin for proposing a very interesting study topic as well as providing support all along the way.

I thank the Fulbright Finland foundation and the U.S. Department of State for their generous support of my studies in Finland. It was the learning experience of a lifetime.

I thank my father, for always supporting my ambitions wherever they may go.

Finally I thank my mother, who bravely saw me off to study overseas, and whose recipe book was paramount for survival in the Finnish winter.

Harrison Mansour

CONTENTS

LIST OF ABBREVIATIONS

1D	one-dimensional
CM	conditional mean
d_H	Hellinger distance
EKI	Ensemble Kalman Inversion
EKS	Ensemble Kalman Sampler
EnKF	Ensemble Kalman Filter
I	Identity matrix
i.i.d	independent and identically distributed
LSQ	least squares
MAP	maximum a posteriori
pdf	probability density function
RWMH	Random Walk Metropolis Hastings

1 INTRODUCTION

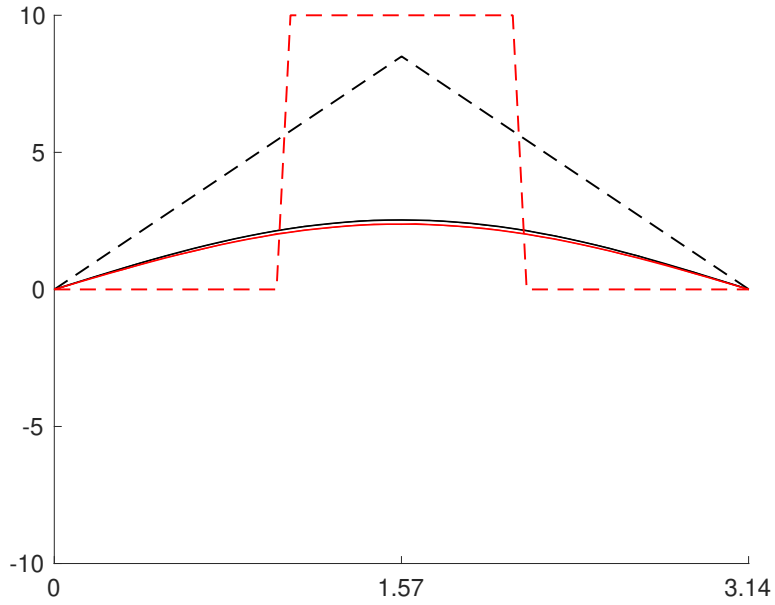
A typical mathematical model predicts the output of a system for a given set of parameters and initial condition. The goal of an inverse problem is to reverse, or *invert*, that process: given the observed (noisy) results, attempt to infer the initial parameters of the system.

A simple example of a forward model is the one-dimensional (1D) heat equation [1] given by

$$\frac{\partial u}{\partial t} = \frac{k\partial^2 u}{\partial x^2}, \tag{1}$$

where k is the thermal diffusivity constant. For specified initial condition $u_0(x)$ and boundary conditions, equation 1 uniquely determines the temperature distribution $u_T(x)$ along the rod at time T.

Contrast this with the inverse heat equation: given the temperature distribution $u_T(x)$ at time T, reconstruct the initial distribution $u_0(x)$. The boundary conditions are understood the same as in the forward problem.



The above figure shows two initial temperature distributions (dotted) with their corresponding final distribution at time T (solid). While the initial distributions are significantly different, the final distributions are almost indistinguishable, implying that the solution to the inverse heat equation is very sensitive to small changes in $u_T(x)$. This sen-

sitivity is only exacerbated by the real-world problem of measurement noise. This means that the inverse heat problem is ill-posed and we can not reliably solve it in the current state.

According to Hadamard [2], a problem is ill-posed if it does not satisfy any one of the following conditions:

1. Existence: A solution exists
2. Uniqueness: The solution is unique
3. Stability: The solution depends continuously on the input.

Usually the third condition is the most difficult to overcome, and a major goal of inverse problem theory is to stabilize the reconstruction process, formulating a well-posed problem that we can hope to solve. The regularization approach (see section 2.1) seeks to approximate the solution by solving a similar but more stable problem instead. This is an optimization method and results in a single point estimate.

Bayesian inverse problems (section 2.3) instead treat the problem as one of statistical inference. Stabilization is done by injecting prior information in the form of a prior distribution μ_0. Then instead of the single point estimate our solution is a probability distribution $u|y$ that tells the likelihood of a parameter value u given the observation y. $u|y$ is called the posterior distribution. Solving inverse problems from a Bayesian perspective then gives the potential for extracting a richer set of information from the solution. We not only have a best guess but also levels of confidence in that guess.

The Ensemble Kalman Sampler (EKS) is an algorithm for solving Bayesian inverse problems introduced by Garbuno-Inigo et al. [3] and is the subject of this paper. It is a *derivative-free* method for generating approximate samples from the posterior distribution. This point is important because it allows EKS to solve problems where the forward model $\mathcal{G}$ has a derivative that is computationally expensive or even impossible.

To make sense of EKS, we first must review some mathematical background. In chapter 2 we first take a deeper look at linear inverse problems, why they are ill-posed, and how regularization methods can remedy that. Section 2.2 then presents prerequisite probability theory, which prepares us to study Bayesian inverse problems for the rest of the chapter.

The Ensemble Kalman Sampler is a data assimilation algorithm, which is a methodology

interested in combining a mathematical model with sequential data to obtain better estimates and is introduced in section 3.1. This section also introduces Ensemble Kalman Inversion, another data assimilation algorithm that EKS is closely related to. Section 3.2 introduces Langevin dynamics, the set of stochastic differential equations that EKS is based on. Finally section 3.3 builds on the previous sections and introduces the Ensemble Kalman Sampler.

Chapter 4 presents two numerical experiments meant to test if EKS effectively samples from the posterior distribution. We look at the 1D deconvolution problem, where we use the Normal equation solution as a baseline. We also look at the 1D elliptic boundary value problem, where we use the Random Walk Metropolis Hastings algorithm as a baseline. With this second problem, we also study how the ensemble size J and the number of iterations N affects runtime performance. Chapter 5 gives a conclusion.

1.1 Objectives and Delimitations

The objectives of this paper are to:

1. Define the Ensemble Kalman Sampler as introduced by Garbuno-Inigo et al [3] and show that it effectively generates approximate samples of the posterior distribution $u|y$ of a Bayesian inverse problem.
2. Describe Bayesian inverse problems. To understand this subject, we introduce inverse problems and basic probability theory up to Bayes' Theorem.
3. Define the mathematical foundation behind the EKS. It is based on the overdamped Langevin diffusion process and pulls heavily from the data assimilation field and the Ensemble Kalman Inversion algorithm.
4. Test the EKS algorithm on the 1D elliptic boundary value problem and compare the mean and covariance of the generated samples with results from Random Walk Metropolis Hastings. With this problem we also study the runtime performance of EKS for different ensemble sizes J and number of iterations N.
5. Test the EKS algorithm on the 1D deconvolution problem, a classic linear inverse problem, and compare results to those found through the Normal equation.

The Ensemble Kalman Sampler is designed as a derivative-free algorithm, capable of solving problems where the forward map $\mathcal{G}(u)$ contains very expensive or uncomputable

derivatives. However, we only study numerical examples in low dimension that are easy to compute. While these examples do not fully display the potential of the algorithm, they do effectively show its ability to sample from the posterior π and presents visual results that are easy to analyze.

2 Mathematical Preliminaries

2.1 About Inverse Problems

Consider the linear inverse problem of estimating the unknown parameters $u \in \mathbb{R}^d$ from known observations $y \in \mathbb{R}^k$:

$$y = Au \tag{2}$$

where $A \in \mathbb{R}^{k \times d}$ models the relationship between u and y. Real-world measurements are always perturbed by some noise, so really we are interested in recovering u from

$$y = Au + \eta \tag{3}$$

where $\eta \sim N(0, \Gamma)$ is unknown, but assumed to be generated by a centered Gaussian with known covariance $\Gamma \sim \mathbb{R}^{k \times k}$. The naive solution is to simply rearrange and invert A:

$$\hat{u} = A^{-1}(y - \eta)$$

however, this approach quickly runs into problems [4]:

1. Consider the case where $d < k$, an overdetermined system. Assume that $A : \mathbb{R}^d \to R(A) \subset \mathbb{R}^k$, where $R(A)$ is the range of A, and that A has a unique inverse $A^{-1} : R(A) \to \mathbb{R}^k$. Because the specific instance of η is unknown, it cannot be subtracted out and $y \notin R(A)$. Thus simply inverting A to solve is impossible.

2. Now consider $d > k$. This is an underdetermined system, so there are infinitely many solutions. This issue can be resolved, for example, by searching for the least squares solution.

3. Finally consider $d = k$. Then $A^{-1} : \mathbb{R}^k \to \mathbb{R}^d$ exists. Assume the condition number $k = \lambda_1/\lambda_k$, where λ_1 and λ_k are respectively the biggest and smallest eigenvalues of A, is very large. This means that A is almost singular, and is called ill-conditioned. In this case, the naive solution $\hat{u} = A^{-1}y - A^{-1}\eta$ can be dominated by $A^{-1}\eta$, making it essentially meaningless. This is because in the worst case, the error $\|A^{-1}\eta\|_2 \approx \frac{\|\eta\|_2}{\lambda_k}$ can be arbitrarily large.

The third issue is certainly the most difficult to cope with. Stabilizing the reconstruction process is one of the key objectives in solving inverse problems, and there are two

major approaches to this problem. The classical approach utilizes regularization theory, where the idea is to approximate the solution by solving a similar, but much more stable problem, instead. The well-regarded Tikhonov regularization replaces equation 3 with the minimization problem [5]

$$\min_{u\in\mathbb{R}^d} \left(\|Au - y\|_2^2 + \alpha \|x\|_2^2 \right), \tag{4}$$

whose solution can be found through the Normal equation:

$$\hat{u}_\alpha = \left(A^T A + \alpha I \right)^{-1} A^T y =: R_\alpha y,$$

where I is the identity matrix, and R_α is called the reconstruction matrix. This problem is inherently more stable because $\lambda_{min}(R_\alpha) = \frac{\lambda_{min}}{\lambda_{min}^2+\alpha}$, giving a reconstruction error on the order of $\frac{1}{\lambda_{min}(R_\alpha)}$.

Although the worst-case error is significantly reduced, this approach introduces a second error term:

$$R_\alpha y = x + \left(R_\alpha A - I \right) x + R_\alpha \eta$$

where

$$\| \left(R_\alpha A - I \right) x \|_2 \leq \frac{\alpha}{\lambda_{min}^2 + \alpha} \|x\|_2,$$

and

$$\|R_\alpha \eta\|_2 \leq \frac{\lambda_{min}}{\lambda_{min}^2 + \alpha} \|\eta\|_2.$$

This result indicates that the choice of α is a delicate balancing act. Increasing α increases the first error term, while decreasing α increases the second error term.

The major alternative to regularization is the statistical, or Bayesian approach to inverse problems. This approach is introduced in section 2.3, after a review of the prerequisite probability theory.

2.2 Probability Prerequisites

The main references for this section are [5, 6].

A probability space is defined as a triplet $(\Omega, \mathcal{F}, \mathbb{P})$. The sample space $\Omega \neq \emptyset$ is a set containing all possible outcomes. The σ-algebra $\mathcal{F}$ on Ω is a set of all possible events

from an experiment, and follows the following axioms:

1. $\emptyset \in \mathcal{F}$
2. if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$, where A^c is the complement of A.
3. If $A_1, A_2, ... \in \mathcal{F}$ is a countable sequence of elements, then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$

The measure $\mathbb{P} : \mathcal{F} \to [0, 1]$ defines a probability to every event in $\mathcal{F}$ and satisfies:

1. $\mathbb{P}(A) \geq 0$, for all $A \in \mathcal{F}$
2. $\mathbb{P}(\Omega) = 1$
3. If $A_1, A_2, ... \in \mathcal{F}$ are pairwise disjoint (meaning that $A_i \cap A_j = 0$, for all $i \neq j$), then $\mathbb{P}(\cup_{i=1}^{\infty}(A_i)) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$. This property is called countable additivity.

Example 2.1. *Consider the simple experiment of rolling a six-sided die one time. The sample space is then* $\Omega = \{1, 2, 3, 4, 5, 6\}$*, and* $\mathcal{F}$ *contains all possible events, such as the die rolls 3:* $\{3\}$ *or the die rolls an odd number:* $\{1, 3, 5\}$.

For any discrete Ω, $\mathcal{F} = Pow(\Omega)$ is simply the power set of Ω. The measure $\mathbb{P}$ then defines a probability for each element of $\mathcal{F}$. In the above example, $\mathbb{P}(\{3\}) = \frac{1}{6}$, and $\mathbb{P}(\{1, 3, 5\}) = \frac{1}{2}$. Because $\mathbb{P}$ is a probability measure, $\mathbb{P}(\Omega) = 1$. More rigorously, the probability space must satisfy the following axioms:

1. $\emptyset \cup \Omega \in \mathcal{F}$
2. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$
3. If $A_i \in \mathcal{F}$, for $i \in \mathbb{N}$, then $\cup_{i=1}^{\infty}(A_i) \in \mathcal{F}$
4. $\mathbb{P}(\Omega) = 1$
5. σ-additivity of $\mathbb{P}$: If measurable sets A_i for $i \in \mathbb{N}$ are pairwise disjoint, then

$$\mathbb{P}\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

A measure is called σ-finite if Ω is a countable union of measurable sets with finite measure [4]. This paper restricts the discussion to the Euclidean space $\Omega = \mathbb{R}^n$, with Borel σ-algebra $\mathcal{F} = \mathcal{B}(\mathbb{R}^n)$, and with Lebesgue measure on $\mathbb{R}^n$.

A random variable X is defined as a measurable map

$$X : (\Omega, \mathcal{F}) \to (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)),$$

which means that if $A \in \mathcal{B}(\mathbb{R}^n)$, then $X^{-1}(A) \in \mathcal{F}$. The probability distribution of X is then defined as a measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$:

$$\mu(A) := \mathbb{P}(X^{-1}(A)),$$

which is read as "the probability that $X \in A$". Shorthand for this is $X \sim \mu$.

Let μ and v be two measures on the same space. Measure μ is said to be absolutely continuous with respect to v if $v(A) = 0$ implies that $\mu(A) = 0$. This is written as $\mu \ll v$. If $\mu \ll v$ and $v \ll \mu$, then the two measures are equivalent. Measures μ and v are called mutually singular if they are supported on disjoint sets. [5]

Theorem 1 (Radon-Nikodym Theorem)**.** *Let μ and v be two measures on the same measure space $(\Omega, \mathcal{F})$. If v is σ-finite and $\mu \ll v$, then there exists a function $f \in L^1(\Omega, \mathcal{F}, v)$ such that*

$$\mu(A) = \int_A f(x) dv(x)$$

for all $A \in \mathcal{F}$,

and $f(x) = \frac{d\mu}{dv}(x)$ is called the Radon-Nikodym derivative of μ with respect to v. From theorem 1, the probability density function of a random variable X can be defined as follows

Definition 2.1 (Probability Density Function)**.** Suppose μ is a a probability measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, $\mathcal{L}_n$ is a Lebesgue measure on $\mathbb{R}^n$, and $\mu \ll \mathcal{L}_n$. Then there exists $\pi \in \mathcal{L}^1(\mathbb{R}^n)$ such that

$$\mu(A) = \int_A \pi(x) dx$$

for any $A \in \mathcal{B}(\mathbb{R}^n)$, and $\pi(x)$ is called the probability density function (pdf) of X.

Definition 2.2 (Joint Distribution Function)**.** Let X and Y be two random variables and

A and B be any corresponding measurable sets. Then the joint distribution of X and Y is

$$\mu_{X\times Y}(A \times B) = \mathbb{P}(X^{-1}(A) \cap Y^{-1}(B)).$$

Definition 2.3 (Marginal Distribution Function)**.** Suppose $Y : \Omega \to \mathbb{R}^n$, then the marginal distribution of X is defined as

$$\mu_X(A) = \mu_{X\times Y}(A \times \mathbb{R}^n).$$

This holds similarly for Y.

Definition 2.4 (Independence)**.** Two random variables X and Y are independent if

$$\mu_{X\times Y}(A \times B) = \mu_X(A)\mu_Y(B)$$

for any measurable sets A and B.

Suppose $\mathcal{G} \subset \mathcal{F}$ is a sub-σ-algebra. In contrast with $\mathcal{F}$, which represents all of the possible events from an experiment, $\mathcal{G}$ contains only partial information because it is generated from random variables. As an example, let $X : \Omega \to \mathbb{R}^n$. Then $\sigma(X)$ is defined as the smallest σ-algebra containing all sets $X^{-1}(A)$ where $A \in \mathcal{B}(\mathbb{R}^n)$. Knowing the value of X means we know whether $X \in A$ happened for each $A \in \mathcal{B}(\mathbb{R}^n)$. This does not mean that we know every possible outcome, only the ones that were observed, so $\sigma(x) \subset \mathcal{F}$. [5]

Definition 2.5 (Conditional Expectation)**.** A random variable $Y \in \mathcal{L}^1(\Omega, \mathcal{G}, \mathbb{P}; \mathbb{R}^n)$ is called the conditional expectation of $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R}^n)$ with respect to the partial-σ-algebra $\mathcal{G}$ if

$$\int_G X(\omega)d\mathbb{P}(\omega) = \int_G Y(\omega)d\mathbb{P}(\omega)$$

for all $G \in \mathcal{G}$. Then define the expectation $\mathbb{E}(X|\mathcal{G}) := Y$.

The conditional probability of an event $A \in \mathcal{F}$ given the sub-σ-algebra $\mathcal{G}$ is a conditional expectation of the form $\mathbb{E}(f(x), \mathcal{G})$ where

$$\mathbb{P}(A|\mathcal{G}) = \mathbb{E}(\mathbb{1}_A(X)|\mathcal{G}),$$

and

$$\mathbb{1}_A(X) := \begin{cases} 1 & if \quad x \in A \\ 0 & if \quad x \notin A \end{cases}$$

is the indicator function of a subset A of a set X.

Definition 2.6 (Regular Conditional Distribution)**.** A family of probability distributions $\mu(\cdot, \omega)_{\omega\in\Omega}$ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ is called a regular conditional distribution of x given $\mathcal{G} \subset \mathcal{F}$ if

$$\mu(A, \cdot) = \mathbb{E}(\mathbb{1}_A(x)|\mathcal{G}) \quad almost\ surely$$

for every $A \in \mathcal{B}(\mathbb{R}^n)$.

Theorem 2. *Given a random variable $X : (\Omega, \mathcal{F}) \to (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ and a sub-σ-algebra $\mathcal{G} \subset \mathcal{F}$, there exists a regular conditional distribution $\mu(\cdot, \omega)_{\omega\in\Omega}$ of X given $\mathcal{G}$.*

We can now define the posterior measure $\mu_{post}(A, y)$ from the regular conditional probability measure as

$$\mu_{post}(A, Y(\omega)) = \mathbb{E}(\mathbb{1}_E(X)|\sigma(Y))(\omega),$$

where $\sigma(Y)$ is the partial-σ-algebra generated by Y.

2.3 Bayesian Inverse Problems

The primary references for this section are [4, 7, 8].

In section 2.1 we introduced linear inverse problems. In this section, we take a look at the more general setting which is defined in terms of a known forward model $\mathcal{G} : \mathbb{R}^d \to \mathbb{R}^k$, that defines the relationship between unknown parameters $u \in \mathbb{R}^d$ and noisy observations $y \in \mathbb{R}^k$ according to

$$y = \mathcal{G}(u) + \eta. \tag{5}$$

In the Bayesian approach, u, y, and η are considered random variables. The unknown $u \in \mathbb{R}^d$ follows a prior distribution μ_0 with Lebesgue density $\rho_0(u)$. This prior density $\rho_0(u)$ contains any information that we already know about u; a good prior is key to formulating a well-posed inverse problem. The unknown measurement noise $\eta \sim N(0, \Gamma)$ is independent of u and is sampled from a centered Gaussian with known covariance $\Gamma \in \mathbb{R}^{k\times k}$. It has Lebesge density $\rho(\eta)$.

The known observations are $y \in \mathbb{R}^k$. The likelihood $y|u$ is defined by equation 5 and has Lebesgue density $\rho(y - \mathcal{G}(u))$. From definition 2.4, it follows that $(u, y) \in \mathbb{R}^d \times \mathbb{R}^k$ is a random variable with Lebesgue density $\rho(y - \mathcal{G}(u))\rho_0(u)$ [7].

The posterior distribution $u|y$ is then the solution to the inverse problem and is given by

Bayes' theorem:

$$\mathbb{P}(u|y) = \frac{1}{\mathbb{P}(y)}\mathbb{P}(y|u)\mathbb{P}(u). \tag{6}$$

Theorem 3 (Bayes' Theorem)**.** *Assume that*

$$Z := \int_{\mathbb{R}^n} \rho(y - \mathcal{G}(u))\rho_0(u) > 0.$$

Then $u|y$ is a random variable with Lebesgue density

$$\pi(u) = \frac{1}{Z}\rho(y - \mathcal{G}(u))\rho_0(u), \tag{7}$$

and Z is the probability of y and acts as a normalizing constant.

Proof. For random variables A and B, let $\mathbb{P}(A)$ be the pdf of A and $\mathbb{P}(A|B)$ be the conditional pdf of A given B.

From definition 2.2, we have

$$\mathbb{P}(u, y) = \mathbb{P}(u|y)\mathbb{P}(y) = \mathbb{P}(y|u)\mathbb{P}(u)$$

and from definition 2.3,

$$\mathbb{P}(y) = \int_{\mathbb{R}^d} \mathbb{P}(u, y)du \quad , \quad \mathbb{P}(u) = \int_{\mathbb{R}^k} \mathbb{P}(u, y)dy.$$

Assume that $\mathbb{P}(y) > 0$, then Bayes' theorem gives us

$$\mathbb{P}(u|y) = \frac{1}{\mathbb{P}(y)}\mathbb{P}(y|u)\mathbb{P}(u) = \frac{1}{Z}\rho(y - \mathcal{G}(u))\rho_0(u)du > 0$$

for $\mathbb{P}(u) \geq 0$. Recall that $\mathbb{P}(y) = Z$, where

$$Z = \int_{\mathbb{R}^d} Z\mathbb{P}(u|y)du = \int_{\mathbb{R}^d} \rho(y - \mathcal{G}(u))\rho_0(u)du > 0.$$

Therefore our assumption that $\mathbb{P}(y) > 0$ is justified and the proof is complete [9]. Note that the theorem only holds when $\mathbb{P}(y) = Z > 0$. It is assumed that Z is positive for the rest of the paper. □

If we define the negative log likelihood, or *potential*, as

$$\Phi(u; y) = -\log\, \rho(y - \mathcal{G}(u)), \tag{8}$$

then theorem 3 can be rewritten as

$$\frac{d\mu^y}{d\mu_0}(u) = \frac{1}{Z}\exp(-\Phi(u; y)) \tag{9}$$

where

$$Z = \int_{\mathbb{R}^d} \exp(-\Phi(u; y))\mu_0(du),$$

and μ^y and μ_0 are measures on $\mathbb{R}^k$ with densities π and ρ_0, respectively. This shows that the posterior is absolutely continuous with respect to the prior density, and the Radon-Nikodym derivative $\frac{d\mu^y}{d\mu_0}(u)$ is proportional to the likelihood. This generalizes theorem 3 to any infinite-dimensional case where μ^y has a Radon-Nikodym derivative with respect to μ_0.

Because the noise is additive Gaussian, we can define the Γ-weighted loss function

$$\ell(a, b) := \frac{1}{2}\|a - b\|_\Gamma^2, \tag{10}$$

where $\|a\|_\Gamma = \|\Gamma^{-\frac{1}{2}}a\|$. This leads to the nonlinear least squares (LSQ) function

$$\Phi(u) := \frac{1}{2}\|y - \mathcal{G}(u)\|_\Gamma^2. \tag{11}$$

If we define a regularization term

$$R(u) := \frac{1}{2}\|u\|_{\Gamma_0}^2, \tag{12}$$

and define the regularized loss function

$$\Phi_R(u) := \Phi(u) + R(u), \tag{13}$$

then we can rewrite the posterior density as

$$\pi(u) \propto \exp\left(-\Phi_R(u)\right). \tag{14}$$

2.3.1 Estimators for Bayesian Inverse Problems

The conditional mean estimator u_{CM} of u given y is analogous to the mean of the posterior distribution:

Definition 2.7 (Conditional Mean Estimator)**.**

$$u_{CM} = \int_{\mathbb{R}^d} u\pi(u)du.$$

The maximum a posteriori estimator u_{MAP} of u given y is analogous to the mode of the posterior distribution:

Definition 2.8 (Maximum A Posteriori Estimator)**.**

$$u_{MAP} = \arg\max_{u\in\mathbb{R}^d} \pi(u).$$

Dashti and Stuart [7] showed that u_{MAP} is equivalent to the solution of Tikhonov-Philips (equation 4), when regularized with respect to the Cameron-Martin space [10].

2.4 Well-Posedness of Bayesian Inverse Problems

The main reference for this section is [9].

Recall the three requirements of a well-posed problem:

1. A solution exists
2. The solution is unique
3. The solution depends continuously on the input.

In order to show that the Bayesian formulation of inverse problems is well-posed, we must show that the formulation satisfies condition 3. Often, the ideal forward map $\mathcal{G}$ is

uncomputable (for example, when $\mathcal{G}$ is a function on an infinite-dimensional space), so it must be approximated by some $\mathcal{G}_\delta$ (in this case, a discrete approximation of $\mathcal{G}$).

It is sufficient to show that, under certain assumptions on the likelihood, that small discrepancies in the forward error $|\mathcal{G} - \mathcal{G}_\delta|$ leads to a difference in the inverse error $d(\pi, \pi_\delta)$ on the same scale. That is, for some metric on probability densities $d(\cdot, \cdot)$ and sufficiently small $\delta > 0$, $|\mathcal{G} - \mathcal{G}_\delta| = O(\delta)$ implies that $d(\pi, \pi_\delta) = O(\delta)$. Here we will use the Hellinger distance $d_H(\pi, \pi')$.

Definition 2.9 (Hellinger Distance)**.** The Hellinger distance $d_H : M \times M \to \mathbb{R}_+$, where M is the set of all probability densities, is defined as

$$d_H(\pi, \pi') := \left(\frac{1}{2}\int |\sqrt{\pi(u)} - \sqrt{\pi'(u)}|^2 du\right)^{\frac{1}{2}} = \frac{1}{\sqrt{2}}\|\sqrt{\pi} - \sqrt{\pi'}\|_{L_2}.$$

The normalization constant of $\frac{1}{\sqrt{2}}$ is justified by the following lemma:

Lemma 2.1. *For any probability densities π and π',*

$$0 \leq d_H(\pi, \pi') \leq 1$$

Proof. $0 \leq d_H(\pi, \pi')$ follows directly from the definition. To show the upper bounds, we exploit the fact that $\int \pi(u)du = 1$ for any probability density function π, thus giving

$$\begin{aligned} d_H(\pi, \pi') &= \left(\frac{1}{2}\int |\sqrt{\pi(u)} - \sqrt{\pi'(u)}|^2 du\right)^{\frac{1}{2}} \\ &= \left(\frac{1}{2}\int (\pi(u) + \pi'(u) - 2\sqrt{\pi(u)\pi'(u)})du\right)^{\frac{1}{2}} \\ &\leq \left(\frac{1}{2}\int (\pi(u) + \pi'(u))du\right)^{\frac{1}{2}} \\ &= 1, \end{aligned}$$

□

completing the proof.

The next lemma is integral to the well-posedness proof. It shows that two densities that are close in Hellinger distance are also close in expectations computed with respect to that distance.

Lemma 2.2. *Let f be a function such that* $\mathbb{E}^{\pi}[|f|^2] + \mathbb{E}^{\pi'}[|f|^2] =: f_2^2 < \infty$, *then*

$$|\mathbb{E}^{\pi}[f] - \mathbb{E}^{\pi'}[f]| \leq 2 f_2 d_H(\pi, \pi').$$

Proof. By the definition of expected value,

$$\begin{aligned}
\left|\mathbb{E}^{\pi}[f] - \mathbb{E}^{\pi'}[f]\right| &= \left| \int_{\mathbb{R}^d} f(u)\pi(u) - f(u)\pi'(u) \right| \\
&= \left| \int_{\mathbb{R}^d} f(u)(\pi(u) - \pi'(u)) \right| \\
&= \left| \int_{\mathbb{R}^d} f(u) \left(\sqrt{\pi(u)} - \sqrt{\pi'(u)}\right) \left(\sqrt{\pi(u)} + \sqrt{\pi'(u)}\right) du \right| \\
&\leq \left(\frac{1}{2} \int \left| \sqrt{\pi(u)} - \sqrt{\pi'(u)} \right|^2 du \right)^{\frac{1}{2}} \left(2 \int |f(u)|^2 \left| \sqrt{\pi(u)} + \sqrt{\pi'(u)} \right| du \right)^{\frac{1}{2}}.
\end{aligned}$$

From definition 2.9,

$$\begin{aligned}
&\leq d_H(\pi, \pi') \left(4 \int |f(u)|^2 (\pi(u) + \pi'(u)) du \right)^{\frac{1}{2}} \\
&= 2 f_2 d_H(\pi, \pi'),
\end{aligned}$$

completing the proof. □

Now we are ready to present the well-posedness theorem for Bayesian inverse problems.

Definition 2.10. Define $g(u)$ and $g_\delta(u)$ as the likelihood and approximate likelihood associated with $\mathcal{G}(u)$ and $\mathcal{G}_\delta(u)$, respectively

$$g(u) = \rho(y - \mathcal{G}(u))$$

$$g_\delta(u) = \rho(y - \mathcal{G}_\delta(u))$$

such that

$$\pi(u) = \frac{1}{Z} g(u)\rho(u)$$

$$\pi_\delta(u) = \frac{1}{Z_\delta} g_\delta(u)\rho(u)$$

where

$$Z = \int g(u)\rho(u) du > 0$$

$$Z_\delta = \int g_\delta(u)\rho(u)du > 0$$

are the respective normalization constants.

Theorem 4 (Well-posedness of Bayesian posterior)**.** *Assume that there exists* $\delta^+ > 0$ *and* $K_1, K_2 < \infty$ *such that for all* $\delta \in (0, \delta^+)$*,*

1. $\left|\sqrt{g(u)} - \sqrt{g_\delta(u)}\right| \leq \phi(u)\delta$*, for some* $\phi(u)$ *such that* $\mathbb{E}^{\rho_0}[\phi^2(u)] \leq K_1$
2. $sup_{u\in\mathbb{R}^d}\left(\left|\sqrt{g(u)} - \sqrt{g_\delta(u)}\right|\right) \leq K_2$.

Then

$$d_H(\pi, \pi_\delta) \leq c\delta, \quad \delta \in (0, \tilde{\delta})^+,$$

for some $\tilde{\delta}^+$ *and some* $c \in (0, \infty) \perp \delta$*.*

The proof of theorem 4 is presented after this helpful lemma.

Lemma 2.3. *Under theorem 4's assumptions, there exists* $\tilde{\delta}^+$ *and* $c_1, c_2 \in (0, \infty)$ *such that, for* $\delta \in (0, \tilde{\delta}^+)$*,*

$$|Z - Z_\delta| \leq c_1\delta.$$

Proof. From definition 2.10 we have, for $\delta \in (0, \delta^+)$,

$$\begin{aligned}
|Z - Z_\delta| &= \int \left(g(u) - g_\delta(u)\right)\rho(u)du \\
&\leq \left(\int \left|\sqrt{g(u)} - \sqrt{g_\delta(u)}\right|^2 \rho(u)du\right)^{\frac{1}{2}} \left(\int \left|\sqrt{g(u)} + \sqrt{g_\delta(u)}\right|^2 \rho(u)du\right)^{\frac{1}{2}} \\
&\leq \left(\delta^2\phi(u)^2 du\right)^{\frac{1}{2}} \left(\int K_2^2\phi(u)du\right)^{\frac{1}{2}} \\
&\leq \sqrt{K_1}K_2\delta.
\end{aligned}$$

When $\delta \leq \tilde{\delta}^+ := \min\left(\frac{Z}{2\sqrt{K_1}K_2}, 2\right)$,

$$Z_\delta \geq Z - |Z - Z_\delta| \geq \frac{Z}{2}.$$

We finish the proof by choosing $c_1 = \sqrt{K_1}K_2$ and $c_2 = \frac{Z}{2}$. □

Proof of Theorem 4. We have

$$\begin{aligned} d_H(\pi, \pi_\delta) &= \frac{1}{\sqrt{2}} \left\| \sqrt{\pi} - \sqrt{\pi_\delta} \right\|_{L^2} \\ &= \frac{1}{2} \left\| \sqrt{\frac{g\rho}{Z}} - \sqrt{\frac{g\rho}{Z_\delta}} + \sqrt{\frac{g\rho}{Z_\delta}} - \sqrt{\frac{g_\delta \rho}{Z_\delta}} \right\|_{L^2} \\ &\leq \frac{1}{\sqrt{2}} \left\| \sqrt{\frac{g\rho}{Z}} - \sqrt{\frac{g\rho}{Z_\delta}} \right\|_{L^2} + \frac{1}{\sqrt{2}} \left\| \sqrt{\frac{g\rho}{Z_\delta}} - \sqrt{\frac{g_\delta \rho}{Z_\delta}} \right\|_{L^2}, \end{aligned}$$

breaking up the distance between pdf's into two L^2 error terms. The first is from the discrepancy between Z and Z_δ, and the second from the difference between g and g_δ. From Lemma 2.3, for $\delta \in (0, \tilde{\delta}^+)$, we can rewrite these error terms as

$$\begin{aligned} \left\| \sqrt{\frac{g\rho}{Z}} - \sqrt{\frac{g\rho}{Z_\delta}} \right\|_{L^2} &= \left| \frac{1}{Z} - \frac{1}{Z_\delta} \right| \left(\int g(u)\rho(u)du \right)^{\frac{1}{2}} \\ &= \frac{|Z - Z_\delta|}{(\sqrt{Z} - \sqrt{Z_\delta})\sqrt{Z_\delta}} \\ &\leq \frac{c_1}{2c_2}\delta, \end{aligned}$$

and

$$\begin{aligned} \left\| \sqrt{\frac{g\rho}{Z_\delta}} - \sqrt{\frac{g_\delta \rho}{Z_\delta}} \right\|_{L^2} &= \frac{1}{\sqrt{Z_\delta}} \left(\int \left| \sqrt{g(u)} - \sqrt{g_\delta(u)} \right|^2 \rho(u)du \right)^{\frac{1}{2}} \\ &\leq \sqrt{\frac{K_1}{c_2}}\delta. \end{aligned}$$

Therefore we have

$$d_H(\pi, \pi_\delta) \leq \frac{1}{\sqrt{2}} \frac{c_1}{2c_2}\delta + \frac{1}{\sqrt{2}} \sqrt{\frac{K_1}{c_2}}\delta = c\delta$$

where $c = \frac{1}{\sqrt{2}} \frac{c_1}{2c_2} + \frac{1}{\sqrt{2}} \sqrt{\frac{K_1}{c_2}} \perp \delta$. □

3 Ensemble Kalman Sampler

The Ensemble Kalman Sampler (EKS) is a data assimilation algorithm closely related to Ensemble Kalman Inversion (EKI), as introduced by Iglesias et al. in [11]. Both attempt to solve the inverse problem (equation 5) through ensemble Kalman methods. EKI takes an optimization approach, arriving at a single "best" estimate for parameters u. EKS, on the other hand, generates samples from the posterior distribution $\pi(u)$, from which we may analyze the statistical properties of u.

The ensemble members $U = \{u^{(j)}\}_{j=0}^{J}$ interact according to overdamped Langevin dynamics, which is introduced in section 3.2. Section 3.1 introduces data assimilation and its links to inverse problems along with the EKI algorithm. Finally, section 3.3 introduces the EKS algorithm.

3.1 Data Assimilation and Ensemble Kalman Inversion

The primary references for this section are [9, 11].

Data Assimilation seeks to *assimilate* sequential data into a theoretical model in order to obtain more realistic estimates, and is key to many fields including weather forecasting. Data assimilation techniques have long been co-opted by the Bayesian inference community, an early example being the Kalman Filter [12], which works in a two step prediction-analysis process. In the prediction stage, the current estimated state u_j and the current observed data y_j are used to generate the next state estimate u_{j+1}. In the analysis stage, we utilize the new observation y_{j+1} to see how good our estimate is, and to correct it.

Assuming linear Gaussian pdfs, the Kalman Filter produces exact results. It does this by maintaining the full covariance matrix at each step, which quickly becomes unfeasible as the dimension of the parameter space increases. The Ensemble Kalman Filter (EnKF) solves this problem by instead representing the state of the system as an ensemble of particles randomly sampled from said system. The covariance matrix is then replaced with the empirical covariance estimated from the ensemble. [13]

Ensemble Kalman Inversion seeks to apply the Ensemble Kalman Filter [12] to the problem of estimating parameters u from equation 5. To this end, we must reformulate the

inverse problem by introducing an artificial time dynamic and simulated data Y_j. The artificial time is given by

$$\begin{aligned} u_{j+1} &= u_j, \\ y_{j+1} &= \mathcal{G}(u_{j+1}) + \eta_{j+1}. \end{aligned}$$

EnKF functions on the assumption of a linear Gaussian model $\mathcal{G}$. Because linearity is not generally the case, we introduce variable w_j and rewrite the dynamics as

$$\begin{aligned} u_{j+1} &= u_j, \\ w_{j+1} &= \mathcal{G}(u_{j+1}), \\ y_{j+1} &= w_{j+1} + \eta_{j+1}. \end{aligned}$$

If we define the space $Z = X \times Y$, where $z = (u, w)^T$; the mapping $\Xi : Z \to Z$, where $\Xi(z) = (u, \mathcal{G}(u))^T$; and linear operators $H = [0, I]$, $H^\perp = [I, 0]$, the dynamic system becomes

$$\begin{aligned} z_{j+1} &= \Xi(z_j) \\ y_{j+1} &= H z_{j+1} + \eta_{j+1} \end{aligned}$$

As mentioned previously, inverse problems tend to be ill-posed and require regularization. In the case of EKI, this is done by restricting the domain where we search for a solution to the compact set $\mathcal{A} \subset \mathbb{R}^d$ that spans the initial ensemble U_0, which is sampled from the prior density $\rho_0(u)$. Iglesias et.al [11] shows that, at each iteration, the ensemble U stays in the set $\mathcal{A}$, and thus the posterior distribution is also in $\mathcal{A}$. This makes the definition of U_0, and thus $\mathcal{A}$, a crucial design parameter.

Once we have the initial ensemble $U_0 = \{u^{(j)}\}_{j=1}^J$, define $z_0^{(j)} = \left(u^{(j)}, \mathcal{G}\left(u^{(j)}\right)\right)^T$. Now we can present the EKI algorithm, which is presented by Iglesias et al. in [11] and repeated below.

1. Impose artificial dynamics on the ensemble $\hat{z}_{n+1} = \Xi(z_n)$ and calculate the sample mean $\bar{z}_{n+1}$ and covariance C_{n+1}.

2. Update each ensemble member

$$z_{n+1}^{(j)} = I\hat{z}_{n+1}^{(j)} + K_{n+1}\left(y_{n+1}^{(j)} - H\hat{z}_{n+1}^{(j)}\right) = \left(I - K_{n+1}H\right)\hat{z}_{n+1}^{(j)} + K_{n+1}y_{n+1}^{(j)}$$

   where $K_n = C_n H^\perp (H C_n H^\perp + \Gamma)^{-1}$ is the Kalman gain.

3.2 Langevin Dynamics

The foundation for EKS is the evolution of an interacting set of particles $U = \{u^{(j)}\}_{j=1}^{J}$ according to overdamped (without inertia) Langevin dynamics:

$$\dot{u} = -\nabla\Phi_R(u) + \sqrt{2}\dot{W} \tag{15}$$

where W is a standard Brownian motion in $\mathbb{R}^d$. $-\nabla\Phi_R(u)$ is the gradient of the regularized loss function, and it will be estimated by differences. This is what makes EKS a derivative-free algorithm. The Langevin equation is a reversible diffusion process that is invariant to the posterior $\pi(u)$ [14]. This means that under conditions on $\Phi_R(u)$ that ensure ergodicity, this system transforms an arbitrary initial distribution into the desired posterior distribution over an infinite time horizon. [3].

Introducing a symmetric matrix $C \in \mathbb{R}^{d\times d}$ into the gradient descent scheme is a common approach to improve convergence time

$$\dot{u} = -C\nabla\Phi_r(u) + \sqrt{2C}\dot{W}. \tag{16}$$

In this case, C is chosen to be the empirical covariance between particles:

$$C(U) = \frac{1}{J}\sum_{k=1}^{J}\left(u^{(k)} - \bar{u}\right)\bigotimes\left(u^{(k)} - \bar{u}\right) \in \mathbb{R}^{d\times d}, \tag{17}$$

where $\bar{u} = \sum_{j=1}^{J} u^{(j)}$ is the sample mean. Thus U is evolved according to the system of stochastic differential equations (SDE):

$$\dot{u}^{(j)} = -C(U)\nabla\Phi_r\left(u^{(j)}\right) + \sqrt{2C(U)}\dot{W}^{(j)} \tag{18}$$

where $\{W^{(j)}\}$ is a collection of i.i.d Brownian motions in $\mathbb{R}^d$. This system can be rewritten as:

$$\dot{u}^{(j)} = -\frac{1}{J}\sum_{k=1}^{J}\langle D\mathcal{G}\left(u^{(j)}\right)\left(u^{(k)} - \bar{u}\right), \mathcal{G}\left(u^{(j)}\right) - y\rangle_\Gamma - C(U)\Gamma_0^{-1}u^{(j)} + \sqrt{2C(U)}\dot{W}^{(j)} \tag{19}$$

where $\langle A, B\rangle_\Gamma = \langle A, \Gamma^{-1}B\rangle = \langle \Gamma^{-\frac{1}{2}}A, \Gamma^{-\frac{1}{2}}B\rangle$ is the Γ-weighted inner product.

We now relate EKI and Langevin dynamics by presenting an alternative formulation of the former. Schillings and Stuart [15] define a continuous time version of EKI in a form

similar to equation 19 as the following system of SDEs:

$$\dot{u}^{(j)} = -\frac{1}{J}\sum_{k=1}^{J}\langle \mathcal{G}\left(u^{(k)}\right) - \bar{\mathcal{G}}, \mathcal{G}\left(u^{(j)}\right) - y\rangle_{\Gamma} u^{(k)} + C^{up}(U)\Gamma^{-1}\sqrt{\Sigma}\dot{W}^{(j)}. \tag{20}$$

When $\Sigma = 0$, thus eliminating the noise term, the algorithm works as a sequential optimizer to minimize $\Phi_R(u)$. For linear problems, when $\Sigma = \Gamma$, the ensemble $U = \{u^{(j)}\}_{j=1}^{J}$ is transformed from the prior to the posterior in one time unit.

$C^{up}(U)$ is defined as the empirical cross covariance of U:

$$C^{up}(U) := \frac{1}{J}\sum_{k=1}^{J}\left(u^{(k)} - \bar{u}\right)\bigotimes\left(\mathcal{G}(u^{(k)}) - \bar{\mathcal{G}}\right) \in \mathbb{R}^{d\times k} \tag{21}$$

$$\bar{\mathcal{G}} := \frac{1}{J}\sum_{k=1}^{J}\mathcal{G}(u^{(j)})$$

The long time dynamics has the undesirable property of ensemble collapse, meaning that all ensemble members will converge to the single point that minimizes $\Phi(u)$ [16]. EKI is a good optimization algorithm, but it is unsuitable for posterior sampling.

3.3 Ensemble Kalman Sampler

The starting point for EKS is the Langevin Dynamics system of equations 19, and in the linear case they coincide exactly. Garbuno-Inigo et al. [3] postulate that this relationship approximately holds in the nonlinear case as well, justifying the proposal of EKS as a derivative-free algorithm for sampling from the posterior distribution $\pi(u)$. If we approximate the derivative $D\mathcal{G}(u^{(j)})(u^{(k)} - \bar{u})$ with differences

$$D\mathcal{G}(u^{(j)})\left(u^{(k)} - \bar{u}\right) \approx \left(\mathcal{G}(u^{(k)}) - \bar{\mathcal{G}}\right), \tag{22}$$

then the Ensemble Kalman Sampler algorithm is defined by the system of SDEs

$$\dot{u}^{(j)} = -\frac{1}{J}\sum_{k=1}^{J}\langle \mathcal{G}\left(u^{(k)}\right) - \bar{\mathcal{G}}, \mathcal{G}\left(u^{(j)}\right) - y\rangle_{\Gamma} - C(U)\Gamma_0^{-1}u^{(j)} + \sqrt{2C(U)}\dot{W}^{(j)}. \tag{23}$$

Conceptually EKS and EKI are very similar, except for two key differences:

1. The noisy EKI introduces additional noise by generating noisy perturbations y_j of the data y, while EKS introduces this noise directly to the particles $u^{(j)}$.

2. Noisy EKI initializes the ensemble U at the prior distribution, while EKS explicitly accounts for the prior on the right hand side of the SDE.

The first point explains why EKS does not experience the same ensemble collapse as EKI. The second point allows the ensemble to be initialized arbitrarily, however it forces a dynamic that transforms the arbitrary initial distribution into samples of the posterior over an *infinite* time horizon.

4 Numerical Results

In this section we test the hypothesis that the Ensemble Kalman Sampler approximately samples from the posterior distribution $\pi(u)$. First, we replicate the numerical experiment conducted by Garbuno-Inigo et al. [3]: solve the 1D elliptic boundary value problem and compare the results to Random Walk Metropolis Hastings. We also solve the 1D deconvolution problem with EKS and compare it to the results generated by solving the Normal equation. This problem allows us to gauge the performance of EKS in higher (d=64) dimension.

4.1 Elliptic Boundary Value Problem

The elliptic boundary value problem is widely used [3,16,17] to test algorithms for solving inverse problems. For $x \in [0, 1]$, consider the elliptic boundary value problem defined by the differential equation

$$-\frac{d}{dx}\left(\exp(u_1)\frac{d}{dx}p(x)\right) = 1, \tag{24}$$

with boundary conditions $p(0) = 0$ and $p(1) = u_2$. It is popular because it has an explicit solution expressed by

$$p(x) = u_2 x + \exp(-u_1)\left(-\frac{x^2}{2} + \frac{x}{2}\right). \tag{25}$$

We are interested in estimating $U = (u_1, u_2)$, given noisy observations $y = (27.5, 79.7)$ at $x = \left(\frac{1}{4}, \frac{3}{4}\right)$. The observational noise $\eta \sim N(0, \Gamma)$ is assumed Gaussian, where $\Gamma = 0.01(I_2)$ and I_2 is the 2×2 identity matrix. The prior distribution is Gaussian $N(0, \Gamma_0)$, where $\Gamma_0 = 100(I_2)$. If we write $\mathcal{G}(u) = (p(x_1), p(x_2))$, then we have an inverse problem in the form of equation 5. Though EKS works with an arbitrary initial ensemble, we sample $U_0 \sim N(0, 1) \bigotimes U(90, 110)$ from the prior distribution used by Ernst et al. [16]. This means $u_1 \sim N(0, 1)$, and $u_2 \sim U(90, 110)$.

4.1.1 Markov Chain Monte Carlo Methods

The main references for this section are [18–21].

For simplicity (and because all computer simulations are discrete), we only present finite Markov chains here. All of the results hold similarly for the continuous case.

A Markov chain is a discrete stochastic process $\{X_n\}_{n=0}^{N}$ that follows the rule

$$\mathbb{P}\left(x_{i+1}|\{x_n\}_{n=0}^{i}\right) = \mathbb{P}\left(x_{i+1}|X_i\right).$$

This means that the new state x_{i+1} depends only on the current state x_i; the chain effectively forgets where it has been in the past. This Markovian property allows for the initial state x_0 to be picked arbitrarily. A Markov chain is defined by its initial state x_0; a state space $\mathcal{S} = \{0, 1, 2, ...\}$ of all possible states the chain can visit; and a transition kernel $P(i, j) := \mathbb{P}\{x_{n+1} = j|x_n = i\}$ that describes, for each pair of states (i, j), the probability of jumping from state i to state j. Because of the Markovian property, P remains the same at each iteration. If $\mathcal{S}$ has r elements, then P can be represented by an $r \times r$ matrix and the transition operation can be written as $x_{n+1} = x_n P$. Also, $x_n = x_0 P^n$.

For simple problems, we can graphically represent the transition kernel P by a state transition diagram. For example, the transition kernel

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}$$

is equivalent to the transition diagram pictured in figure 1 below.

If there exists a valid path from state i to state j (if there exists a nonnegative integer n such that $P^n(i, j) > 0$), then we say j is *reachable* from i, or $i \to j$. Here, 4 is reachable from 1, but 5 is not reachable from 3. If two states are reachable from each other, such as 1 and 2, then we say they *communicate*, or $i \leftrightarrow j$. Because communication is an equivalence relation [20], we can partition a state transition diagram into *communicating classes*, where all elements communicate with each other within their own class. In our example, there are three communicating classes: $\{1, 2\}$, $\{3, 4\}$, and $\{5\}$. Note that because state 5 communicates with no other states, $\{5\}$ would be a valid communicating

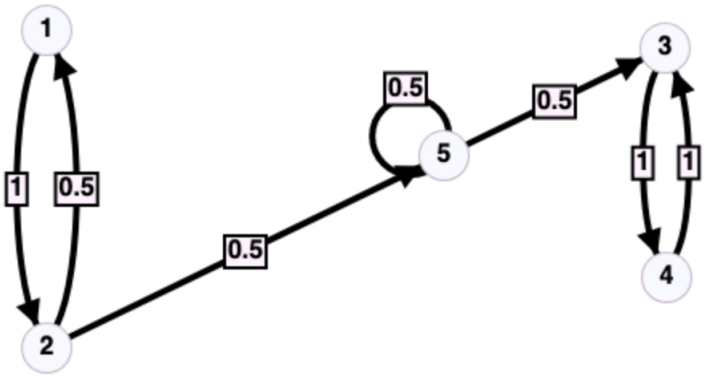


Figure 1. State transition diagram of kernel P introduced above. Edge weights represent transition probabilities.

class even without the self-loop. A Markov chain with only one communicating class is called *irreducible*.

There are two types of states, *recurrent* and *transient* ones. A state is recurrent if, upon leaving the state, we will eventually return to it with probability 1. All other states are transient. In our example, states 3 and 4 are recurrent. All states in the same class are of the same type, so we can define each of our communicating classes as either transient or recurrent.

Notice in our example that once we reach state 3 (or 4), that we will certainly return to it every two iterations. These states then have a period of two. The periodicity of a state i is defined as $d_i = gcd\{n : P^n(i,i) > 0\}$. A state i is *aperiodic* if $d_i = 1$. If two states i, j are in the same communicating class, then $d_i = d_j$.

For some Markov chains, there exists a *stationary distribution* π_*, where $\pi_* P = \pi_*$. That is to say that once the chain arrives at π_*, it will never leave. That Markov chain has become *stationary*. This parallels the recurrent class introduced above. In the simple example, $\{3, 4\}$ is the stationary distribution. It is clear that as $n \to \infty$, the probability that the chain settles in $\{3, 4\}$ approaches 1. We formalize this notion in the coming theorem, which states that under certain conditions, a Markov chain is guaranteed to converge to π_* (assuming it exists).

Theorem 5 (Basic Limit Theorem)**.** *Let X be an aperiodic, irreducible Markov chain, that is arbitrarily initiated at $x_0 \sim \pi_0$ and has a stationary distribution π_*. Then*

$$\lim_{n\to\infty} \pi_n(i) = \pi_*(i)$$

for all states $i \in \mathcal{S}$.

A *time reversible* Markov chain is one that satisfies the local balance equations

$$\pi(i)P(i,j) = \pi(j)P(j,i),$$

which basically tell us that the probability of moving from state i to j is the same as moving from j to i. A time reversible Markov chain is stationary [22].

The central idea behind Random Walk Metropolis Hastings (RWMH) is that we want to create a Markov chain that has the posterior distribution of the inverse problem as its stationary distribution π_*. If that chain is time reversible, then we can be certain that it will converge as we desire. Consider a proposal distribution $q(i,j)$ that generates a proposed new state j given that the current state is i. If $\pi(i)q(i,j) = \pi(j)q(j,i)$ for all $i,j \in \mathcal{S}$, then the transition kernel $P(i,j) = q(i,j)$ is time reversible. However, this is not generally the case.

Suppose there exists some $i,j \in \mathcal{S}$ such that $\pi(i)q(i,j) > \pi(j)q(j,i)$. Then the move $i \to j$ is likely to happen more often than the reverse move. We can balance this by adding an acceptance probability term $P(i,j) = q(i,j)\alpha(i,j)$. We accept new moves with probability $\alpha(i,j)$, else we stay at the current state. We do not want to restrict the less frequent direction, so set $\alpha(j,i) = 1$. The local balance equations are then

$$\pi(i)q(i,j)\alpha(i,j) = \pi(j)q(j,i)$$

which implies that

$$\alpha(i,j) = \frac{\pi(j)q(j,i)}{\pi(i)q(i,j)}.$$

Because α is a probability,

$$\alpha(i,j) = \min\{\frac{\pi(j)q(j,i)}{\pi(i)q(i,j)}, 1\}$$

The RWMH algorithm is a "random walk" because it uses a Gaussian proposal density q. That is, $x_{n+1} = x_n + \varepsilon$, where $\varepsilon \sim N(0,C)$. Tuning the covariance C is an effective way to tune the chain's convergence rate. Because the proposal density is symmetric, we can simplify α to

$$\alpha(i,j) = \min\{\frac{\pi(j)}{\pi(i)}, 1\}.$$

Then the RWMH transition kernel can be represented as the three step algorithm

1. Propose a new candidate $x_{n+1} = x_n + \varepsilon$
2. Calculate acceptance probability $\alpha(x_n, x_{n+1}) = \min\{\frac{\pi(x_{n+1})}{\pi(x_n)}, 1\}$
3. Let $u \sim U(0,1)$. If $u \leq \alpha(x_n, x_{n+1})$ then accept the move. Else set $x_{n+1} = x_n$.

Roberts et al. [23] has shown that an optimal convergence rate can be found when the acceptance rate of proposed moves is 0.234. We used a constant covariance $C = 0.211(I_2)$, which gave an acceptance rate of 0.237 .

4.1.2 Results

For the primary results, we run EKS with $J = 10^4$ particles and ? iterations. With RWMH, we use $N = 10^5$ iterations and only using the last 10^4 samples for analysis, at which point the influences from the arbitrary initial state is negligible.

In figure 3 we see the results for RWMH on top and EKS on the bottom. The mean value of EKS samples is $(-2.72, 104.31)$, whereas the mean value of RWMH samples is $(-2.67, 104.44)$ and are marked by a red $\times$. Also drawn are the 1, 2, and 3 standard deviation ellipses (that is, 60, 95, and 99.7% confidence intervals). This shows us that the EKS samples have a slightly larger covariance. This discrepancy is not huge however; this can be seen in figure 2 where both ensembles are plotted together. EKS samples are plotted in red, with a green $\times$ marking the mean value. RWMH samples are plotted in black, with a white $\times$ marking the mean value.

We also used this problem to test performance of EKS with varying ensemble size $J \in [10, 10^4]$ and number of iterations $N \in [10, 10^3]$, compared to the RWMH results. We found that convergence was much more dependent on having sufficiently large N than for J. As long as $N \geq 160$, all values of J resulted in convergence to the RWMH solution. Empirical covariances were largely consistent among all experiments where $N \geq 160$.

Though the size of N is more influential to convergence, it is not to runtime. Through graphical analysis we see a linear relationship between N and runtime (see figure 4) and a roughly parabolic relationship between J and runtime (see figure 5).

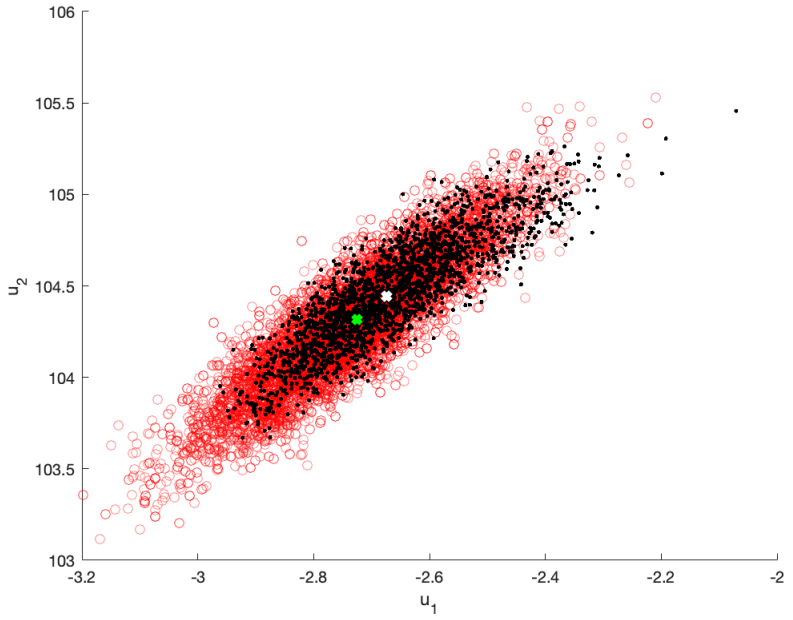


Figure 2. A scatter plot of samples generated by RWMH (black) and EKS (red) plotted together. Mean values are marked in green and white, respectively.

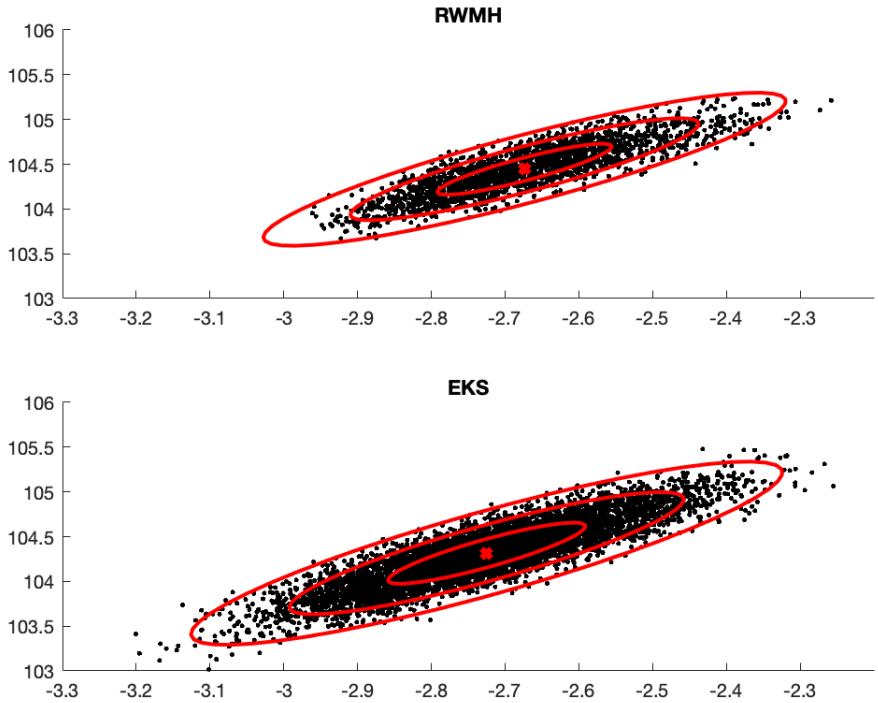


Figure 3. A scatter plot of samples generated by RWMH (top) and EKS (bottom) with confidence intervals (red). The x-axis is variable u_1 and the y-axis is u_2.

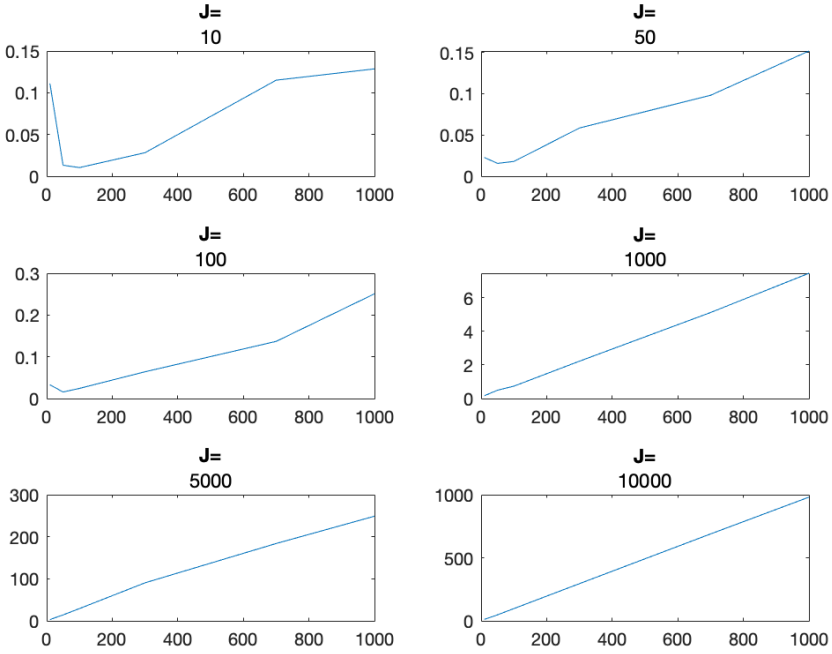


Figure 4. For several ensemble sizes J, the number of iterations N (x-axis) plotted vs the runtime in seconds (y-axis).

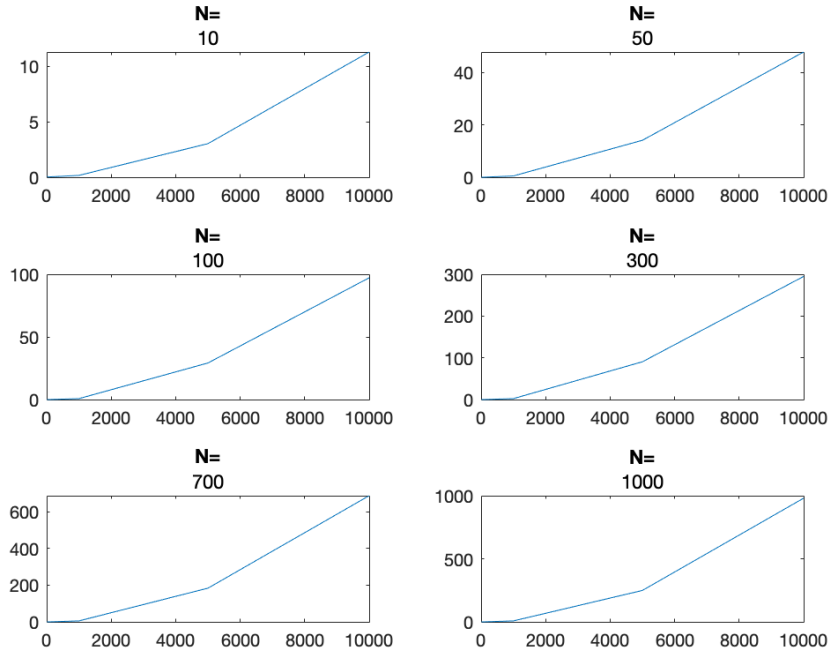


Figure 5. For several iteration counts N, the ensemble size J (x-axis) plotted vs the runtime in seconds (y-axis).

4.2 One Dimensional Deconvolution

The 1D deconvolution problem is similar to the backwards heat equation presented in the introduction. It is the inverse of 1D convolution, a function on two vectors – the *signal* $f \in \mathbb{R}^d$ and the *convolution kernel* $g \in \mathbb{R}^k$ – with output $h \in \mathbb{R}^d$ defined via

$$h(i) = (f * g)(i) = \sum_{j=1}^{k} g(j) \cdot f\left(i - j + \frac{m}{2}\right),$$

where $*$ is the convolution operator. The intuition behind convolution is to think of the kernel g as a filter sliding over the signal f. $h(i)$ is then found by centering g at $f(i)$ and take the dot product of all overlapping terms. For example, take $f = [1, 2, 3, 4, 5, 6]$ and $g = [10, 20, 30]$. Then

$$\begin{aligned}
h(1) &= 1 * 20 + 2 * 30 = 80 \\
h(2) &= 1 * 10 + 2 * 20 + 3 * 30 = 140 \\
h(3) &= 2 * 10 + 3 * 20 + 4 * 30 = 200 \\
h(4) &= 3 * 10 + 4 * 20 + 5 * 30 = 260 \\
h(5) &= 4 * 10 + 5 * 20 + 6 * 30 = 320 \\
h(6) &= 5 * 10 + 6 * 20 = 170
\end{aligned}$$

and thus $h = [80, 140, 200, 260, 320, 170]$. The kernel $g = [10, 20, 30]$ does not have any particular mathematical meaning, but this operation can have many interpretations depending on the definition of g. For example $g = \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right]$ produces a windowed average of f, effectively blurring it. The identity kernel is $g = [0, 1, 0]$ and results in $h = f$. The kernel $g = [0, 0, 1]$ will shift f to the left by 1. With the above example, $h = [2, 3, 4, 5, 6, 0]$.

The task of deconvolution is to approximate the original signal f, given a noisy observation of the output h and the known convolution kernel g. In other words, recover f from

$$h = f * g + \eta.$$

Now we outline the specific deconvolution problem we will solve. Figure 6 shows the true signal f in blue, and the noisy observation h in red.

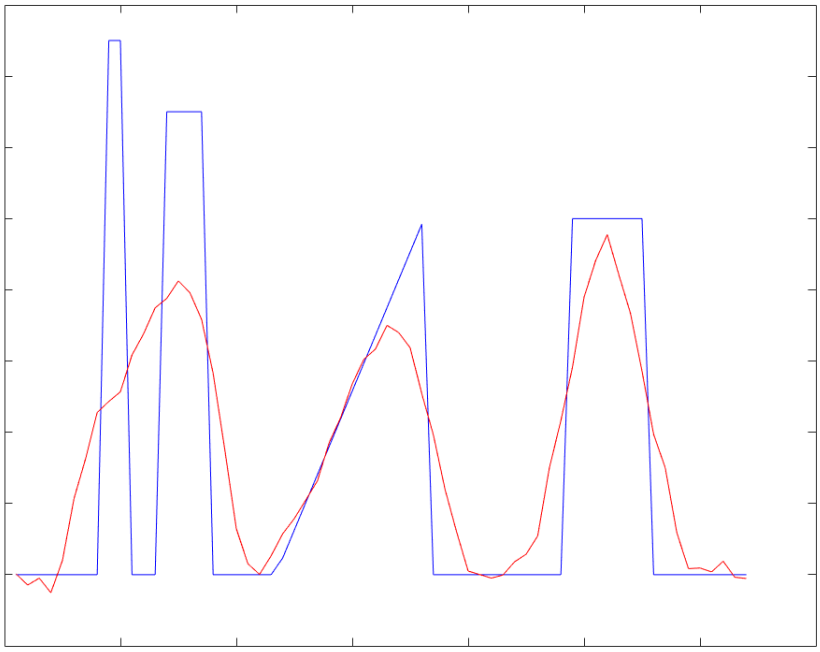


Figure 6. The true signal f (blue) and the noisy observation h (red). The x-axis is position and the y-axis is signal intensity.

The convolution kernel g is defined by

$$g(x) := \begin{cases} (x+a)^2(x-a)^2 & \text{if } x \in [-a, a] \\ 0 & \text{elsewhere,} \end{cases} \tag{26}$$

where a is a constant. If we define the matrix $A \in \mathbb{R}^{d \times d}$ such that the ith row of A corresponds to the ith application of the convolution kernel on f. For our simple example above, where $g = [10, 20, 30]$,

$$A = \begin{bmatrix} 20 & 30 & 0 & 0 & 0 & 0 \\ 10 & 20 & 30 & 0 & 0 & 0 \\ 0 & 10 & 20 & 30 & 0 & 0 \\ 0 & 0 & 10 & 20 & 30 & 0 \\ 0 & 0 & 0 & 10 & 20 & 30 \\ 0 & 0 & 0 & 0 & 10 & 20 \end{bmatrix}.$$

Then the convolution can be written as a linear system $h = Af$. We factor in the obser-

vation noise $\eta \sim \Gamma$

$$h = Af + \eta,$$

and obtain a linear inverse problem. This can be solved via the Normal equation

$$\hat{f}_{Normal} = \left(A^T\Gamma^{-1}A + \Gamma_0^{-1}\right)^{-1}\left(A^T\Gamma^{-1}y\right),$$

where Γ and Γ_0 are the covariance of the Gaussian observation noise and of the Gaussian prior, respectively. This result will be used as a baseline to gauge the effectiveness of our Ensemble Kalman Sampler to solve inverse problems.

4.2.1 Ensemble Kalman Sampler for 1D Deconvolution

Here we show how to solve inverse problems with EKS . Garbuno-Inigo et al. [3] describe a linearly implicit split-step discretization scheme to approximate equation 23 by

$$\begin{aligned} u_{n+1}^{(*,j)} &= u_n^{(j)} - \Delta t_n \frac{1}{J}\sum_{k=1}^{J}\langle \mathcal{G}\left(u_n^{(k)}\right) - \bar{\mathcal{G}}, \mathcal{G}\left(u_n^{(j)}\right) - y\rangle_\Gamma u_n^{(k)} - \Delta t_n C(U_n)\Gamma_0^{-1}u_{n+1}^{(*,j)} \\ u_{n+1}^{(j)} &= u_{n+1}^{(*,j)} + \sqrt{2\Delta t_n C(U_n)}\xi_n^{(j)} \end{aligned} \tag{27}$$

where $\xi_n^{(j)} \sim N(0, I)$. We use the same adaptive time-step Δt_n as Kovachki and Stuart [24].

In implementation it is helpful to rearrange the first step as

$$u_{n+1}^{(*,j)} = \left(I + \Delta t_n C(U_n)\Gamma_0^{-1}\right)^{-1}\left(u_n^{(j)} - \Delta t_n \frac{1}{J}\sum_{k=1}^{J}\langle \mathcal{G}\left(u_n^{(k)}\right) - \bar{\mathcal{G}}, \mathcal{G}\left(u_n^{(j)}\right) - y\rangle_\Gamma u_n^{(k)}\right).$$

4.2.2 Results

After running the algorithm, calculate the empirical conditional mean estimator (definition 2.7) of the final ensemble as

$$U_{CM} = \frac{1}{J}\sum_{j=1}^{J} u_n^{(j)}.$$

The conditional mean estimate of the EKS result is written as $\hat{f}_{EKS}$. Plotted with confidence intervals (CI) the results are shown in Figure 8 (red CI). This can be compared to

the result $\hat{f}_{Normal}$ obtained by the Normal equation, as shown in Figure 9 (blue CI).

We can see that both approaches do a good job at recreating the source signal f. EKS successfully found the double peaks in the beginning (these were lost in the observation h), and does good job at tracking the final two peaks. We can look at the convergence of the conditional mean estimate as a function of time in figure 7, where we see that the EKS error $\|\hat{f}_{EKS} - f\|_2$ approaches the error of the Normal solution $\|\hat{f}_{Normal} - f\|_2$ as the number of iterations N grows large. More formally,

$$\lim_{n\to\infty} \|\hat{f}_{EKS} - f\|_2 \approxeq \|\hat{f}_{Normal} - f\|_2.$$

From this we conclude that $\hat{f}_{EKS}$ approximately converges to $\hat{f}_{Normal}$, and that EKS approximately solves the 1D deconvolution problem.

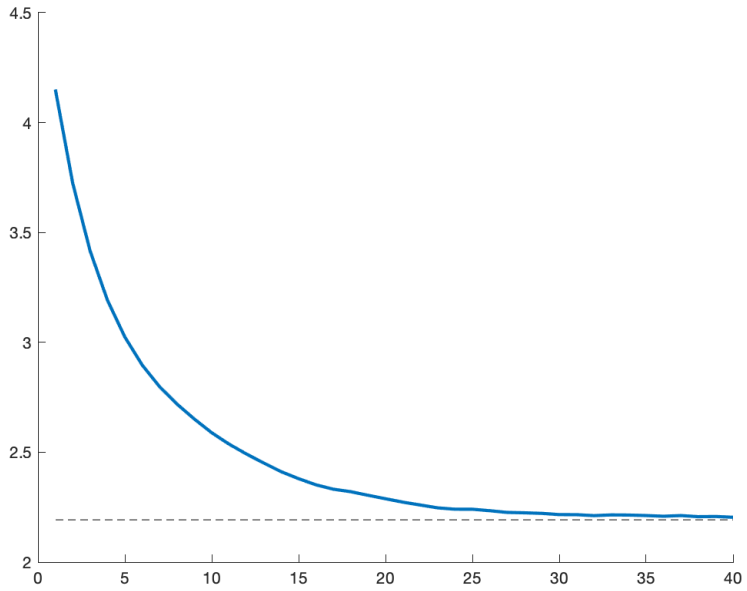


Figure 7. The evolution of the EKS error $\|\hat{f}_{EKS} - f\|_2$ (blue) compared to the Normal solution error $\|\hat{f}_{Normal} - f\|_2$ (dotted), with respect to number of iterations complete.

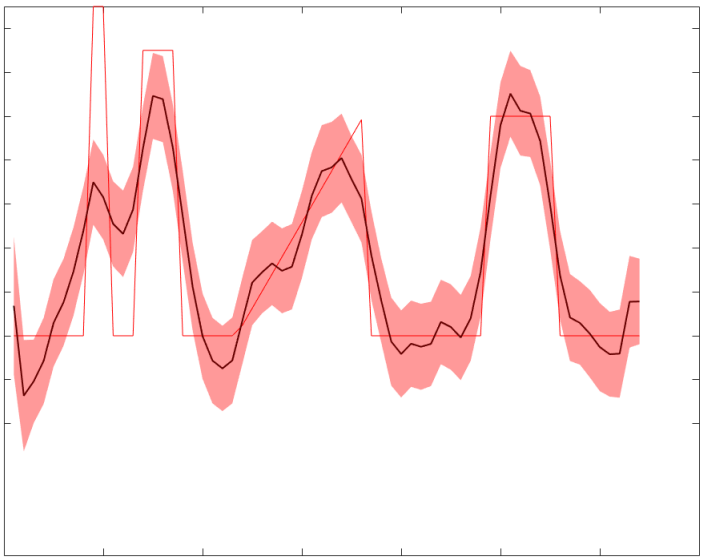


Figure 8. The Conditional Mean estimate $\hat{f}_{EKS}$ (black) of the EKS results with confidence intervals (red highlight) compared to the truth (red).

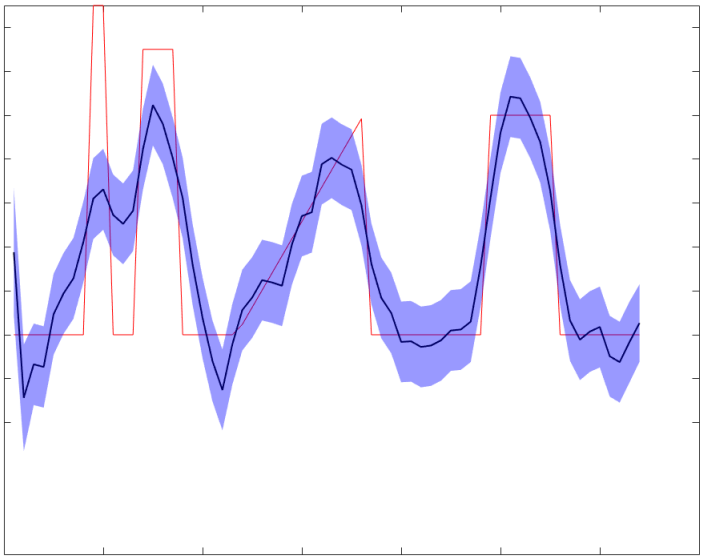


Figure 9. True Conditional Mean estimate $\hat{f}_{Normal}$ (black) from the Normal equation with confidence intervals (blue highlight) compared to the truth (red).

5 Conclusion

The main goal of this paper was to introduce the Ensemble Kalman Sampler, originally proposed by Garbuno-Inigo et al [3] as a derivative free algorithm for sampling from the posterior distribution $\pi(u)$ (equation 7) and thus solving Bayesian inverse problems. EKS is built upon the Langevin equation (19) and is a variant of Ensemble Kalman Inversion (equation 20) where noise is added directly to the particles.

We evaluate the performance of EKS against the 1D elliptic boundary value problem and the 1D deconvolution problem. For the former, we compare its sampling properties against RWMH and find that they have comparable performance. We also analyze the runtime and see that runtime scales linearly with number of iterations N and quadratically with ensemble size J. For the latter, we analyze the conditional mean estimate of EKS and see that it performs just as well as the Normal equation solution.

REFERENCES

[1] Matthew J. Hancock. The 1-d heat equation. https://ocw.mit.edu/courses/mathematics/18-303-linear-partial-differential-equations-fall-2006/lecture-notes/heateqni.pdf, 2006. [Online: accessed 16-November-2020].

[2] Jacques Hadamard. *Lectures on Cauchy's problem in linear partial differential equations*. Dover, New York, NY, 1952. This book has also been published by Dover in 2003.

[3] Alfredo Garbuno-Inigo, Franca Hoffmann, Wuchen Li, and Andrew M. Stuart. Interacting langevin diffusions: Gradient structure and ensemble kalman sampler. *Applied Dynamical Systems*, 19(1), 2020.

[4] Hanne Kekkonen. Bayesian inverse problems. http://www.damtp.cam.ac.uk/research/cia/files/teaching/Inverse_Problems_Lent_2018/19y_01m_16d_LectureNotes.pdf, 2019. [Online: accessed 04-December-20].

[5] Tapio Helin. Bayesian inversion: Theoretical perspective. https://courses.helsinki.fi/sites/default/files/course-material/4542044/Lecture_notes_part1.pdf, January 2018. [Online: accessed 04-December-20].

[6] George Casella and Roger Berger. *Statistical Inference*. Duxbury Resource Center, June 2001.

[7] Masoumeh Dashti and Andrew M. Stuart. The bayesian approach to inverse problems, 2013.

[8] Guillaume Bal. Introduction to inverse problems. https://www.stat.uchicago.edu/~guillaumebal/PAPERS/IntroductionInverseProblems.pdf, July 2019. [Online: accessed 04-December-20].

[9] Daniel Sanz-Alonso, Andrew M. Stuart, and Armeen Taeb. Inverse problems and data assimilation, 2018.

[10] Martin Hairer. An introduction to stochastic pdes, 2009.

[11] Marco A Iglesias, Kody J H Law, and Andrew M Stuart. Ensemble kalman methods for inverse problems. *Inverse Problems*, 29(4):045001, Mar 2013.

[12] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45.

[13] Jan Mandel. A brief tutorial on the ensemble kalman filter, 2009.

[14] Grigorios A. Pavliotis. *Stochastic processes and applications: diffusion processes, the Fokker- Planck and Langevin equations*, volume 60. Springer, New York, NY, 2014.

[15] Claudia Schillings and Andrew M. Stuart. Analysis of the ensemble kalman filter for inverse problems. *Numerical Analysis*, 55(3):1264–1290.

[16] Oliver G. Ernst, Björn Sprungk, and Hans-Jörg Starkloff. Analysis of the ensemble and polynomial chaos kalman filters in bayesian inverse problems. 3:823–851, 2015.

[17] Michael Herty and Giuseppe Visconti. Kinetic methods for inverse problems, 2018.

[18] Roger Peng. Advanced statistical computing. `https://bookdown.org/rdpeng/advstatcomp/`, May 2020. [Online: accessed 04-December-20].

[19] Joe Chang. Stochastic processes. `http://www.stat.yale.edu/~pollard/Courses/251.spring2013/Handouts/Chang-MarkovChains.pdf`, February 2007. [Online: accessed 04-December-20].

[20] Karl Sigman. Communication classes and irreducibility for markov chains. `http://www.columbia.edu/~ww2040/4701Sum07/4701-06-Notes-MCII.pdf`. [Online: accessed 04-December-20].

[21] Hossein Pishro-Nik. Introduction to probability. `https://www.probabilitycourse.com/chapter11/11_2_1_introduction.php`. [Online: accessed 04-December-20].

[22] Siddhartha Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.

[23] G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk metropolis algorithms. *Ann. Appl. Probab.*, 7(1):110–120, 02 1997.

[24] Nikola B Kovachki and Andrew M Stuart. Ensemble kalman inversion: a derivative-free technique for machine learning tasks. *Inverse Problems*, 35(9):095005, Aug 2019.