




# Stochastic Spectral Approaches to Bayesian Inference

Prof. Nathan L. Gibson

Department of Mathematics



Applied Mathematics and Computation Seminar  
March 4, 2011

-  Y.M. Marzouk, H.N. Najm, and L.A. Rahn.  
Stochastic spectral methods for efficient Bayesian solution of inverse problems.  
*Journal of Computational Physics*, 224(2):560–586, 2007.
-  Y. Marzouk and D. Xiu.  
A Stochastic Collocation Approach to Bayesian Inference in Inverse Problems.  
*Communications in Computational Physics*, 6:826–847, 2009.
-  Y.M. Marzouk and H.N. Najm.  
Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems.  
*Journal of Computational Physics*, 228(6):1862–1902, 2009.

# Outline

## 1 Inference Background

# Outline

## 1 Inference Background

## 2 Posterior Surrogate

- Stochastic Collocation
- PC Stochastic Galerkin

# Outline

- 1 **Inference Background**
- 2 **Posterior Surrogate**
  - Stochastic Collocation
  - PC Stochastic Galerkin
- 3 **Polynomial Chaos**
  - PC Galerkin Example

# Outline

- 1 **Inference Background**
- 2 **Posterior Surrogate**
  - Stochastic Collocation
  - PC Stochastic Galerkin
- 3 **Polynomial Chaos**
  - PC Galerkin Example
- 4 **PC Stochastic Galerkin**
  - Generalizations
  - Convergence
  - Example
  - Distributed Parameters

# Inverse Problems

- Inverse problems arise from indirect observations of a quantity of interest.
- A physical system may be described by a (forward) model, which predicts some measurable features of the system given a set of parameters.
- The corresponding inverse problem consists of inferring these parameters from a set of observations of the features.

- Observations may be limited in number relative to the dimension or complexity of the parameter space.



- Observations may be limited in number relative to the dimension or complexity of the parameter space.
- Further, the action of the forward operator may include filtering or smoothing effects.

- Observations may be limited in number relative to the dimension or complexity of the parameter space.
- Further, the action of the forward operator may include filtering or smoothing effects.
- These features typically render inverse problems ill-posed in the sense that
  - no solution may exist
  - multiple solutions may exist
  - solutions may not depend continuously on the data

- Observations may be limited in number relative to the dimension or complexity of the parameter space.
- Further, the action of the forward operator may include filtering or smoothing effects.
- These features typically render inverse problems ill-posed in the sense that
  - no solution may exist
  - multiple solutions may exist
  - solutions may not depend continuously on the data
- In practical settings, where observations are inevitably corrupted by noise, this presents numerous challenges.

- Observations may be limited in number relative to the dimension or complexity of the parameter space.
- Further, the action of the forward operator may include filtering or smoothing effects.
- These features typically render inverse problems ill-posed in the sense that
  - no solution may exist
  - multiple solutions may exist
  - solutions may not depend continuously on the data
- In practical settings, where observations are inevitably corrupted by noise, this presents numerous challenges.
- Classical approaches to inverse problems have used regularization methods to impose well-posedness, and solved the resulting problems by optimization.

- Observations may be limited in number relative to the dimension or complexity of the parameter space.
- Further, the action of the forward operator may include filtering or smoothing effects.
- These features typically render inverse problems ill-posed in the sense that
  - no solution may exist
  - multiple solutions may exist
  - solutions may not depend continuously on the data
- In practical settings, where observations are inevitably corrupted by noise, this presents numerous challenges.
- Classical approaches to inverse problems have used regularization methods to impose well-posedness, and solved the resulting problems by optimization.
- However, important insights and methodologies emerge by casting inverse problems in the framework of statistical inference.

# Bayesian Inference

- In particular, the Bayesian setting for inverse problems offers
  - a rigorous foundation for inference from noisy data and uncertain forward models
  - a natural mechanism for incorporating prior information
  - a quantitative assessment of uncertainty in the inferred results

# Bayesian Inference

- In particular, the Bayesian setting for inverse problems offers
  - a rigorous foundation for inference from noisy data and uncertain forward models
  - a natural mechanism for incorporating prior information
  - a quantitative assessment of uncertainty in the inferred results
- Indeed, the output of Bayesian inference is not a single value for the model parameters, but a probability (**posterior**) distribution that summarizes all available information about the parameters.

# Bayesian Inference

- In particular, the Bayesian setting for inverse problems offers
  - a rigorous foundation for inference from noisy data and uncertain forward models
  - a natural mechanism for incorporating prior information
  - a quantitative assessment of uncertainty in the inferred results
- Indeed, the output of Bayesian inference is not a single value for the model parameters, but a probability (**posterior**) distribution that summarizes all available information about the parameters.
- From this posterior distribution, one may estimate means, modes, and higher-order moments, compute marginal distributions, or make additional predictions by averaging over the posterior.



Consider a forward problem

$$d^t = G(Z)$$

where  $d^t \in \mathbb{R}^{n_d}$  represents some true, observable data and  $Z = (Z_1, \dots, Z_{n_z})$  is a set (vector) of model parameters.

Consider a forward problem

$$d^t = G(Z)$$

where  $d^t \in \mathbb{R}^{n_d}$  represents some true, observable data and  $Z = (Z_1, \dots, Z_{n_z})$  is a set (vector) of model parameters.

In the Bayesian setting:

- $Z$  is a random variable.

Assume each component of  $Z$  has a (prior) probability distribution

$$F_i(z_i) = P(Z_i < z_i) \in [0, 1]$$

and a (prior) probability density  $\pi_i(z) = dF_i/dz_i$ . Assuming each  $Z_i$  mutually independent then  $\pi_Z(z) = \prod_{i=1}^{n_z} \pi_i(z_i)$  is the joint density function for  $Z$  (we will omit subscripts when convenient).

Consider a forward problem

$$d^t = G(Z)$$

where  $d^t \in \mathbb{R}^{n_d}$  represents some true, observable data and  $Z = (Z_1, \dots, Z_{n_z})$  is a set (vector) of model parameters.

In the Bayesian setting:

- $Z$  is a random variable.

Assume each component of  $Z$  has a (prior) probability distribution

$$F_i(z_i) = P(Z_i < z_i) \in [0, 1]$$

and a (prior) probability density  $\pi_i(z) = dF_i/dz_i$ . Assuming each  $Z_i$  mutually independent then  $\pi_Z(z) = \prod_{i=1}^{n_z} \pi_i(z_i)$  is the joint density function for  $Z$  (we will omit subscripts when convenient).

- Probability is used to express knowledge about true values of parameters. In other words, the prior and posterior probabilities represent degrees of belief about values before and after observing the data.

## Prior Probability Distribution

Information about the parameters (quantities of interest) may enter through the prior density  $\pi(Z)$ .

Examples include

- range of feasible values
- correlations
- shape or smoothness

- In practice there may exist measurement error so that the observed data does not match the predicted data.

- In practice there may exist measurement error so that the observed data does not match the predicted data.
- Assuming additive observational errors, we have

$$d = d^t + e = G(Z) + e,$$

where components of  $e \in \mathbb{R}^{n_d}$  are i.i.d. random variables with probability density function  $\pi_e$ .

- In practice there may exist measurement error so that the observed data does not match the predicted data.
- Assuming additive observational errors, we have

$$d = d^t + e = G(Z) + e,$$

where components of  $e \in \mathbb{R}^{n_d}$  are i.i.d. random variables with probability density function  $\pi_e$ .

- We make the usual assumption that  $e$  is also independent of  $Z$ .

- In practice there may exist measurement error so that the observed data does not match the predicted data.
- Assuming additive observational errors, we have

$$d = d^t + e = G(Z) + e,$$

where components of  $e \in \mathbb{R}^{n_d}$  are i.i.d. random variables with probability density function  $\pi_e$ .

- We make the usual assumption that  $e$  is also independent of  $Z$ .
- In this simple model,  $e$  may encompass both measurement error (e.g. sensor noise) and model error—the extent to which forward model predictions may differ from “true” values because of some unmodeled physics of the system.



- In practice there may exist measurement error so that the observed data does not match the predicted data.
- Assuming additive observational errors, we have

$$d = d^t + e = G(Z) + e,$$

where components of  $e \in \mathbb{R}^{n_d}$  are i.i.d. random variables with probability density function  $\pi_e$ .

- We make the usual assumption that  $e$  is also independent of  $Z$ .
- In this simple model,  $e$  may encompass both measurement error (e.g. sensor noise) and model error—the extent to which forward model predictions may differ from “true” values because of some unmodeled physics of the system.
- A typical choice is  $e_i \sim \mathcal{N}(0, \sigma^2)$ .

- In practice there may exist measurement error so that the observed data does not match the predicted data.
- Assuming additive observational errors, we have

$$d = d^t + e = G(Z) + e,$$

where components of  $e \in \mathbb{R}^{n_d}$  are i.i.d. random variables with probability density function  $\pi_e$ .

- We make the usual assumption that  $e$  is also independent of  $Z$ .
- In this simple model,  $e$  may encompass both measurement error (e.g. sensor noise) and model error—the extent to which forward model predictions may differ from “true” values because of some unmodeled physics of the system.
- A typical choice is  $e_i \sim \mathcal{N}(0, \sigma^2)$ .
- Note that if parameters in the noise model, or prior, are unknown, they may themselves be endowed with a prior and included as **hyperparameters** to be estimated from data. For example, some choices of  $\sigma$  above give “better” results than others.

## Likelihood Function

- Data enters the formulation through the **likelihood** which is defined to be  $\pi(d|z)$  and may be viewed as a function of  $z$ :

$$L(z) = \pi(d|z).$$

This is the **conditional probability** of  $d$  given  $z$ .

## Likelihood Function

- Data enters the formulation through the **likelihood** which is defined to be  $\pi(d|z)$  and may be viewed as a function of  $z$ :

$$L(z) = \pi(d|z).$$

This is the **conditional probability** of  $d$  given  $z$ .

- One estimate of a quantity of interest given data is that which maximizes the likelihood function (or minimizes  $-\log(L(z))$ ), called the MLE.

## MLE vs. MAP

- Information is generally available in the form  $P(\text{effect}|\text{cause})$  rather than  $P(\text{cause}|\text{effect})$  which is what we would like to determine.

## MLE vs. MAP

- Information is generally available in the form  $P(\text{effect}|\text{cause})$  rather than  $P(\text{cause}|\text{effect})$  which is what we would like to determine.
- For example,  $P(\text{symptom}|\text{disease})$  is easy to measure by observations, however  $P(\text{disease}|\text{symptom})$  is more important yet harder to determine.

## MLE vs. MAP

- Information is generally available in the form  $P(\text{effect}|\text{cause})$  rather than  $P(\text{cause}|\text{effect})$  which is what we would like to determine.
- For example,  $P(\text{symptom}|\text{disease})$  is easy to measure by observations, however  $P(\text{disease}|\text{symptom})$  is more important yet harder to determine.
- Rather than finding the maximizer to the former (MLE) we may instead wish to maximize the latter (MAP).

For this reason we employ Bayes' Rule

$$\pi(z|d) = \frac{\pi(d|z) \pi(z)}{\int \pi(d|z) \pi(z) dz}$$

where again  $\pi(z)$  is the prior density of  $Z$  and  $\pi(z|d)$ , the density of  $Z$  conditioned on  $d$ , is the posterior density of  $Z$ . To emphasize the dependence on  $z$  one may write

$$\pi^d(z) = \frac{L(z) \pi(z)}{\int L(z) \pi(z) dz}.$$



For this reason we employ Bayes' Rule

$$\pi(z|d) = \frac{\pi(d|z) \pi(z)}{\int \pi(d|z) \pi(z) dz}$$

where again  $\pi(z)$  is the prior density of  $Z$  and  $\pi(z|d)$ , the density of  $Z$  conditioned on  $d$ , is the posterior density of  $Z$ . To emphasize the dependence on  $z$  one may write

$$\pi^d(z) = \frac{L(z) \pi(z)}{\int L(z) \pi(z) dz}.$$

In the presence of additive noise as defined above, the likelihood becomes

$$L(z) = \pi(d|z) = \prod_{i=1}^{n_d} \pi_e(d_i - G_i(z))$$

For this reason we employ Bayes' Rule

$$\pi(z|d) = \frac{\pi(d|z) \pi(z)}{\int \pi(d|z) \pi(z) dz}$$

where again  $\pi(z)$  is the prior density of  $Z$  and  $\pi(z|d)$ , the density of  $Z$  conditioned on  $d$ , is the posterior density of  $Z$ . To emphasize the dependence on  $z$  one may write

$$\pi^d(z) = \frac{L(z) \pi(z)}{\int L(z) \pi(z) dz}.$$

In the presence of additive noise as defined above, the likelihood becomes

$$L(z) = \pi(d|z) = \prod_{i=1}^{n_d} \pi_e(d_i - G_i(z))$$

We now see that the posterior density may be evaluated at any  $z$  for a cost of an evaluation of  $G(z)$ .

- Consider a general, time-dependent PDE with dependence on a set of random variables  $Z \in \mathbb{R}^{n_z}$

$$u_t(x, t, z) = \mathcal{L}(u), \quad D \times (0, T] \times \mathbb{R}^{n_z}$$

where  $D$  is some spatial domain and  $T > 0$  is some fixed time, along with appropriate boundary and initial conditions, and where  $\mathcal{L}$  is a (nonlinear) differential operator.

- Consider a general, time-dependent PDE with dependence on a set of random variables  $Z \in \mathbb{R}^{n_z}$

$$u_t(x, t, z) = \mathcal{L}(u), \quad D \times (0, T] \times \mathbb{R}^{n_z}$$

where  $D$  is some spatial domain and  $T > 0$  is some fixed time, along with appropriate boundary and initial conditions, and where  $\mathcal{L}$  is a (nonlinear) differential operator.

- We define the observation map  $g : \mathbb{R}^{n_u} \rightarrow \mathbb{R}^{n_d}$  relating the solution  $u$  of the system to the true observable data

$$d^t = g(u)$$

then the forward model becomes

$$d^t = G(Z) = g \circ u(Z)$$

and computation of  $G$  is presumed to be expensive.

- Therefore, while (to the extent that  $\pi(z)$  and  $\pi(e)$  are known)  $\pi^d(z)$  can be evaluated at any  $z$ , this is far from being useful. The posterior probability distribution is typically not of analytical form and, especially in high dimensions, cannot be easily interrogated.

- Therefore, while (to the extent that  $\pi(z)$  and  $\pi(e)$  are known)  $\pi^d(z)$  can be evaluated at any  $z$ , this is far from being useful. The posterior probability distribution is typically not of analytical form and, especially in high dimensions, cannot be easily interrogated.
- A typical use of Bayesian estimation is in the computation of integrals over the posterior

$$I[f] = \int f(z)\pi^d(z)dz = \int f(z)L(z)\pi(z)dz$$

for example, the **posterior expectation** of  $f$  is  $\mathbb{E}_\pi f = I[f]/I[1]$ .

- Therefore, while (to the extent that  $\pi(z)$  and  $\pi(e)$  are known)  $\pi^d(z)$  can be evaluated at any  $z$ , this is far from being useful. The posterior probability distribution is typically not of analytical form and, especially in high dimensions, cannot be easily interrogated.
- A typical use of Bayesian estimation is in the computation of integrals over the posterior

$$I[f] = \int f(z)\pi^d(z)dz = \int f(z)L(z)\pi(z)dz$$

for example, the **posterior expectation** of  $f$  is  $\mathbb{E}_\pi f = I[f]/I[1]$ .

- Note that approximations could be constructed utilizing sampling  $z^{(j)}$  from the prior (MC), or a deterministic quadrature rule could be used at discrete nodes  $z^{(k)}$  on the support of the prior, but each requires numerous evaluations of  $G$ .

## Posterior Surrogate

- An alternate approach that may accelerate evaluation of Bayesian integrals and other characterizations of the posterior is to employ spectral representations of uncertain parameters and field quantities, specifically polynomial chaos (PC) expansions for random variables and stochastic processes, in order to create a **posterior surrogate** for efficient Bayesian inference in inverse problems.



## Posterior Surrogate

- An alternate approach that may accelerate evaluation of Bayesian integrals and other characterizations of the posterior is to employ spectral representations of uncertain parameters and field quantities, specifically polynomial chaos (PC) expansions for random variables and stochastic processes, in order to create a **posterior surrogate** for efficient Bayesian inference in inverse problems.
- In particular, we propagate prior uncertainty through the forward model, thus yielding a polynomial approximation  $\tilde{G}$  of the forward solution over the support of the prior.

## Posterior Surrogate

- An alternate approach that may accelerate evaluation of Bayesian integrals and other characterizations of the posterior is to employ spectral representations of uncertain parameters and field quantities, specifically polynomial chaos (PC) expansions for random variables and stochastic processes, in order to create a **posterior surrogate** for efficient Bayesian inference in inverse problems.
- In particular, we propagate prior uncertainty through the forward model, thus yielding a polynomial approximation  $\tilde{G}$  of the forward solution over the support of the prior.
- This approximation then enters the likelihood function, resulting in an approximate posterior density that is inexpensive to evaluate.

## Two main approaches

- Stochastic Galerkin methods:  
Exact computation of the coefficients of  $\tilde{G}$  via Galerkin projection results in an **intrusive** stochastic spectral methodology, in which polynomial chaos representations of the unknown parameters lead to a **reformulation** of the governing equations of the forward model.

## Two main approaches

- Stochastic Galerkin methods:  
Exact computation of the coefficients of  $\tilde{G}$  via Galerkin projection results in an **intrusive** stochastic spectral methodology, in which polynomial chaos representations of the unknown parameters lead to a **reformulation** of the governing equations of the forward model.
- Stochastic Collocation:  
Uses numerical quadrature rather than Galerkin projection to approximate the coefficients in the spectral expansion. A key advantage of stochastic collocation is that it requires only a finite number of uncoupled deterministic simulations, with no reformulation of the governing equations of the forward model.

## Comments on Stochastic Collocation

- preferable in very high dimensions: many methods already exist for addressing high input dimensionality via efficient low-degree integration formulae or sparse grids

## Comments on Stochastic Collocation

- preferable in very high dimensions: many methods already exist for addressing high input dimensionality via efficient low-degree integration formulae or sparse grids
- can deal with highly nonlinear problems that are challenging, if not impossible, to handle with stochastic Galerkin methods

## Comments on Stochastic Collocation

- preferable in very high dimensions: many methods already exist for addressing high input dimensionality via efficient low-degree integration formulae or sparse grids
- can deal with highly nonlinear problems that are challenging, if not impossible, to handle with stochastic Galerkin methods
- a spectral representation may also be applied to arbitrary functionals of the forward solution (e.g., location of a root of a solution)

## Comments on Stochastic Collocation

- preferable in very high dimensions: many methods already exist for addressing high input dimensionality via efficient low-degree integration formulae or sparse grids
- can deal with highly nonlinear problems that are challenging, if not impossible, to handle with stochastic Galerkin methods
- a spectral representation may also be applied to arbitrary functionals of the forward solution (e.g., location of a root of a solution)
- does not depend on Galerkin projection step and therefore can be more flexible in the choice of prior/polynomial pair.



- This process involves:
  - ① constructing PC expansions for each unknown parameter and field quantity, according to probability distributions that include the support of the prior

- This process involves:
  - 1 constructing PC expansions for each unknown parameter and field quantity, according to probability distributions that include the support of the prior
  - 2 substituting these expansions into the governing equations and using Galerkin projection to obtain a coupled system of equations for the PC mode strengths of  $\tilde{G}$

- This process involves:
  - 1 constructing PC expansions for each unknown parameter and field quantity, according to probability distributions that include the support of the prior
  - 2 substituting these expansions into the governing equations and using Galerkin projection to obtain a coupled system of equations for the PC mode strengths of  $\tilde{G}$
  - 3 solving this system

- This process involves:
  - 1 constructing PC expansions for each unknown parameter and field quantity, according to probability distributions that include the support of the prior
  - 2 substituting these expansions into the governing equations and using Galerkin projection to obtain a coupled system of equations for the PC mode strengths of  $\tilde{G}$
  - 3 solving this system
  - 4 forming an expression for the posterior density based on the resulting PC expansions of forward model predictions, then exploring this posterior density with an appropriate sampling strategy.

- This process involves:
  - 1 constructing PC expansions for each unknown parameter and field quantity, according to probability distributions that include the support of the prior
  - 2 substituting these expansions into the governing equations and using Galerkin projection to obtain a coupled system of equations for the PC mode strengths of  $\tilde{G}$
  - 3 solving this system
  - 4 forming an expression for the posterior density based on the resulting PC expansions of forward model predictions, then exploring this posterior density with an appropriate sampling strategy.
- In this scheme, sampling can have negligible cost; nearly all the computational time is spent solving the system in Step 3.

- This process involves:
  - 1 constructing PC expansions for each unknown parameter and field quantity, according to probability distributions that include the support of the prior
  - 2 substituting these expansions into the governing equations and using Galerkin projection to obtain a coupled system of equations for the PC mode strengths of  $\tilde{G}$
  - 3 solving this system
  - 4 forming an expression for the posterior density based on the resulting PC expansions of forward model predictions, then exploring this posterior density with an appropriate sampling strategy.
- In this scheme, sampling can have negligible cost; nearly all the computational time is spent solving the system in Step 3.
- Depending on model non-linearities and the necessary size of the PC basis, this computational effort may be orders of magnitude less costly than exploring the posterior via direct sampling.

Let  $\xi$  be a random variable with density  $\pi_\xi(\xi) = w(\xi)$  and assume (say)  $\xi \sim U[-1, 1]$ . Then the normalized Legendre polynomials  $\{\psi_i(\xi)\}$  are orthogonal with respect to  $w$ , i.e.,

$$\langle \psi_i, \psi_j \rangle = \int \psi_i(\xi) \psi_j(\xi) w(\xi) d\xi = \delta_{ij}.$$

Let  $\xi$  be a random variable with density  $\pi_\xi(\xi) = w(\xi)$  and assume (say)  $\xi \sim U[-1, 1]$ . Then the normalized Legendre polynomials  $\{\psi_i(\xi)\}$  are orthogonal with respect to  $w$ , i.e.,

$$\langle \psi_i, \psi_j \rangle = \int \psi_i(\xi) \psi_j(\xi) w(\xi) d\xi = \delta_{ij}.$$

Suppose  $X$  is a random variable with a known **polynomial chaos (PC) expansion**

$$X = \sum_{i=0}^P x_i \psi_i(\xi).$$

Then the orthogonality property can be used to calculate the truncated PC expansion of a function  $f(X)$  by projecting onto the PC basis

$$\tilde{f} = \sum_{k=0}^P f_k \psi_k(\xi), \quad f_k = \langle f(X), \psi_k \rangle.$$



Let  $\xi$  be a random variable with density  $\pi_\xi(\xi) = w(\xi)$  and assume (say)  $\xi \sim U[-1, 1]$ . Then the normalized Legendre polynomials  $\{\psi_i(\xi)\}$  are orthogonal with respect to  $w$ , i.e.,

$$\langle \psi_i, \psi_j \rangle = \int \psi_i(\xi) \psi_j(\xi) w(\xi) d\xi = \delta_{ij}.$$

Suppose  $X$  is a random variable with a known **polynomial chaos (PC) expansion**

$$X = \sum_{i=0}^P x_i \psi_i(\xi).$$

Then the orthogonality property can be used to calculate the truncated PC expansion of a function  $f(X)$  by projecting onto the PC basis

$$\tilde{f} = \sum_{k=0}^P f_k \psi_k(\xi), \quad f_k = \langle f(X), \psi_k \rangle.$$

This orthogonal projection minimizes the error  $\|f - \tilde{f}\|$  on the space spanned by  $\{\psi_k\}_{k=0}^P$ .

Suppose that the behavior of  $f$  can be expressed as  $\mathcal{O}(f, X) = 0$ , where  $\mathcal{O}$  is some deterministic operator. Substituting PC expansions for  $f$  and  $X$  into this operator and requiring the residual to be orthogonal to  $\psi_j$  for  $j = 0 \dots P$  yields a set of coupled, deterministic equations for the PC coefficients  $f_k$ :

$$\left\langle \mathcal{O} \left( \sum_{k=0}^P f_k \psi_k, \sum_{i=0}^P x_i \psi_i \right), \psi_j \right\rangle = 0, \quad j = 0, \dots, P.$$

Suppose that the behavior of  $f$  can be expressed as  $\mathcal{O}(f, X) = 0$ , where  $\mathcal{O}$  is some deterministic operator. Substituting PC expansions for  $f$  and  $X$  into this operator and requiring the residual to be orthogonal to  $\psi_j$  for  $j = 0 \dots P$  yields a set of coupled, deterministic equations for the PC coefficients  $f_k$ :

$$\left\langle \mathcal{O} \left( \sum_{k=0}^P f_k \psi_k, \sum_{i=0}^P x_i \psi_i \right), \psi_j \right\rangle = 0, \quad j = 0, \dots, P.$$

This Galerkin approach is known as “intrusive” spectral projection due to the reformulation of the forward problem, in contrast to “non-intrusive” approaches in which the inner product is evaluated by sampling or quadrature, thus requiring repeated evaluations of  $f(X)$  corresponding to different realizations of  $\xi$ .

For simplicity assume  $n_z = n_d = n_u = 1$  and let  $g$  be the identity map. Consider the ODE model

$$\dot{u} = -Zu, \quad Z = Z(\xi) = \xi, \quad \xi \sim \mathcal{N}(0, 1).$$

For simplicity assume  $n_z = n_d = n_u = 1$  and let  $g$  be the identity map. Consider the ODE model

$$\dot{u} = -Zu, \quad Z = Z(\xi) = \xi, \quad \xi \sim \mathcal{N}(0, 1).$$

We apply a truncated Polynomial Chaos expansion in terms of orthogonal Hermite polynomials  $H_j$  to the solution  $u$ :

$$u(t, \xi) = \sum_{k=0}^P u_k(t) \psi_k(\xi).$$

For simplicity assume  $n_z = n_d = n_u = 1$  and let  $g$  be the identity map. Consider the ODE model

$$\dot{u} = -Zu, \quad Z = Z(\xi) = \xi, \quad \xi \sim \mathcal{N}(0, 1).$$

We apply a truncated Polynomial Chaos expansion in terms of orthogonal Hermite polynomials  $H_j$  to the solution  $u$ :

$$u(t, \xi) = \sum_{k=0}^P u_k(t) \psi_k(\xi).$$

Taking the Galerkin Projection onto  $\text{span}(\{\psi_i\}_{i=0}^P)$ , the ODE becomes

$$\dot{u}_j + (j + 1)u_{j+1} + u_{j-1} = 0, \quad j = 0, \dots, P,$$

Letting  $\vec{u}$  represent the vector containing  $u_0(t), \dots, u_P(t)$  the system of ODEs can be written

$$\dot{\vec{u}} + M\vec{u} = \vec{0}.$$

## Comments on PC Stochastic Galerkin

- While the coupled system is larger than the original model, it need only be solved once to determine the coefficients involved in the approximate model  $\tilde{G}$ .
- Note also that this approximate model is independent of the parameters of the prior density of  $Z$ , it merely depended on the choice of type distribution of  $\xi$ .

## Generalizations

- For arbitrary observation map  $g$ , recalling our notation  $G(Z) = g \circ u(Z)$ , we define instead the PC expansion  $\tilde{G} = \sum_{i=0}^P g_k \psi(\xi)$  and compute  $g_k$  using the procedure above.



## Generalizations

- For arbitrary observation map  $g$ , recalling our notation  $G(Z) = g \circ u(Z)$ , we define instead the PC expansion  $\tilde{G} = \sum_{i=0}^P g_k \psi(\xi)$  and compute  $g_k$  using the procedure above.
- If polynomials of degree less than or equal to  $N$  are used, then the number of terms in the expansion is  $P = \frac{(n_z + N)!}{n_z! N!} - 1$ .

## Generalizations

- For arbitrary observation map  $g$ , recalling our notation  $G(Z) = g \circ u(Z)$ , we define instead the PC expansion  $\tilde{G} = \sum_{i=0}^P g_k \psi(\xi)$  and compute  $g_k$  using the procedure above.
- If polynomials of degree less than or equal to  $N$  are used, then the number of terms in the expansion is  $P = \frac{(n_z+N)!}{n_z! N!} - 1$ .
- If the prior is such that it is difficult to write a PC expansion for  $Z$  then a change of variables  $Z = h(\xi)$  may be assumed where the support of the density of  $h(\xi)$  includes that of the prior. The change of variables formula says that

$$\pi_{\xi}(\xi) = \pi_z(h(\xi)) |\det(Dh(\xi))|$$

where  $Dh$  is the Jacobian matrix of  $h$ . This imposes that  $h$  be a diffeomorphism.

## Convergence

- To establish convergence of the PC-based Bayesian algorithm, we quantify the difference between the approximate posterior  $\tilde{\pi}_N^d$  computed with an  $N^{\text{th}}$  degree PC expansion and the exact posterior  $\pi^d$ , via Kullback-Leibler divergence.

# Convergence

- To establish convergence of the PC-based Bayesian algorithm, we quantify the difference between the approximate posterior  $\tilde{\pi}_N^d$  computed with an  $N^{\text{th}}$  degree PC expansion and the exact posterior  $\pi^d$ , via Kullback-Leibler divergence.
- The Kullback-Leibler divergence (KLD) measures the difference between probability distributions and is defined, for probability density functions  $\pi_1(z)$  and  $\pi_2(z)$ , as

$$D(\pi_1 \parallel \pi_2) := \int \pi_1(z) \log \frac{\pi_1(z)}{\pi_2(z)} dz.$$

- It is always non-negative, and  $D(\pi_1 \parallel \pi_2) = 0$  when  $\pi_1 = \pi_2$ .

## Theorem ([2] 4.1)

*Under certain assumptions, if*

$$\|G(Z) - G_N(Z)\|_{L^2_{\pi_Z}} \leq C N^{-\alpha}, \alpha > 0$$

*then*

$$D(\tilde{\pi}_N^d \| \pi^d) \leq C N^{-\alpha}, \alpha > 0.$$

## Theorem ([2] 4.1)

*Under certain assumptions, if*

$$\|G(Z) - G_N(Z)\|_{L^2_{\pi_Z}} \leq C N^{-\alpha}, \quad \alpha > 0$$

*then*

$$D(\tilde{\pi}_N^d \| \pi^d) \leq C N^{-\alpha}, \quad \alpha > 0.$$

Thus if the forward problem converges exponentially fast, the inverse problem converges at least as fast.

## Theorem ([2] 4.1)

*Under certain assumptions, if*

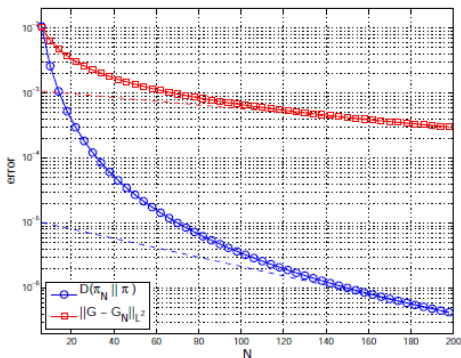
$$\|G(Z) - G_N(Z)\|_{L^2_{\pi_Z}} \leq C N^{-\alpha}, \quad \alpha > 0$$

*then*

$$D(\tilde{\pi}_N^d \| \pi^d) \leq C N^{-\alpha}, \quad \alpha > 0.$$

Thus if the forward problem converges exponentially fast, the inverse problem converges at least as fast.

- A similar result holds for  $G_N^Q(Z)$  where  $Q$  denotes the order of the numerical quadrature rule used in Stochastic Collocation.



[2] Fig. 2. Convergence of the forward model  $G$  in  $L^2$  norm, and posterior density  $\pi^d$  in Kullback-Leibler distance  $D(\tilde{\pi}_N^d \parallel \pi^d)$  from the direct posterior to the approximate posterior, versus  $N$ .



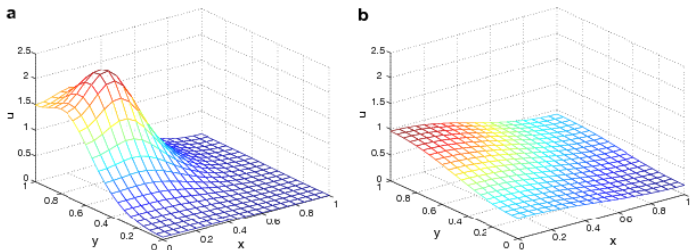
Consider the dimensionless diffusion model on a  $2D$  square domain

$$\frac{\partial u}{\partial t} = \nabla^2 u + \frac{s}{2\pi\sigma^2} \exp\left(\frac{-\|\vec{q} - \vec{x}\|^2}{2\sigma^2}\right) [1 - H(t - \tau)], \quad (0, T] \times S$$

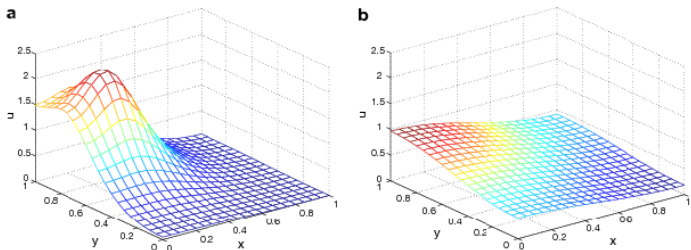
$$\nabla u \cdot \hat{n} = 0, \quad \partial S$$

$$u(\vec{x}, 0) = 0, \quad S$$

where  $S = [0, 1] \times [0, 1]$ ,  $t \in [0, T]$ . The gaussian source term is active on a finite interval of time  $[0, \tau]$ , centered at  $\vec{q}$  with size and strength given by  $\sigma$  and  $s$ , respectively.



[1] Fig. 1. Scalar field  $u$  in the deterministic forward problem for  $(q_1, q_2) = (0.25, 0.75)$  at (a)  $t = 0.05$  and (b)  $t = 0.15$ .



[1] Fig. 1. Scalar field  $u$  in the deterministic forward problem for  $(q_1, q_2) = (0.25, 0.75)$  at (a)  $t = 0.05$  and (b)  $t = 0.15$ .

- Note that for later times the inverse problem becomes more ill-posed since as the solution flattens it becomes more dominated by noise.

- Each of  $q$ ,  $\sigma$ ,  $s$ ,  $\tau$  could be parameters to be determined.
- In this example we seek to infer the location of the source  $q = (q_1, q_2)$  (with all other parameters known) based on (noisy) observations of  $u$  at several locations over a few discrete times,

$$\begin{aligned}d^t &= \{u(\vec{x}_i, t_j)\}_{i,j=1}^{M,N}, \quad MN = n_d \\ &= \{u(\vec{x}_\ell, t_\ell)\}_{\ell=1}^{n_d}\end{aligned}$$

- We assume in advance only that  $q_1$  and  $q_2 \in [0, 1]$ , thus the prior is  $U[0, 1]$  for each component.
- We further assume that errors between “real-world” measurements and model predictions (encapsulating measurement error, model error and simulation error) are i.i.d.  $\mathcal{N}(0, \zeta^2)$  with known  $\zeta$ .

- Given that the prior is a uniform distribution, we choose to employ Legendre Polynomials.
- We further define “standard” random variables  $\xi_i \sim U[-1, 1]$ ,  $i = 1, 2$  so that  $q = h(\xi)$  defines the (linear) map from  $[-1, 1]$  to  $[0, 1]$  for each component of  $q$

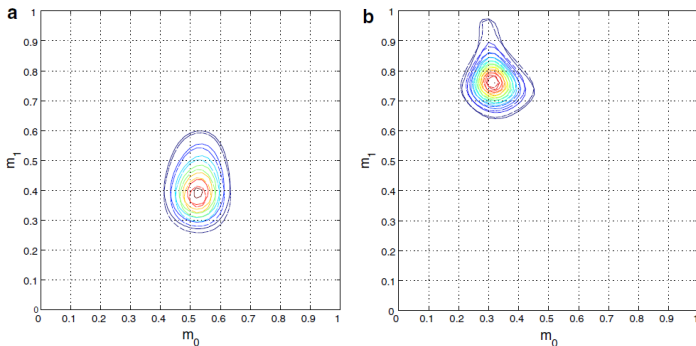
$$h_i = 0.5 + 0.5\xi_i.$$

- Then each component of  $q$  can be written in a PC expansion,

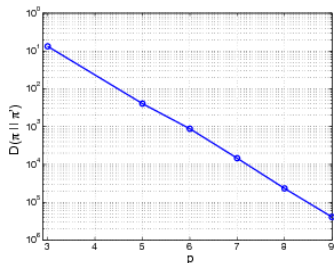
$$q_i = g_i(\xi) = \sum_{k=0}^P g_{ik} \psi_k(\xi)$$

(where only two are actually needed).

- Using stochastic Galerkin method we obtain a PC expansion for each prediction of the scalar field  $\tilde{G}_\ell(\xi)$  which replaces  $G_\ell$  in the likelihood function.



[1] Fig. 16. Contours of the posterior density. Solid lines are obtained with direct evaluations of the forward problem and dashed lines from Uniform-Legendre PC expansions with  $p = 9$ .



[1] Fig. 7. Kullback-Leibler distance  $D(\tilde{\pi} || \pi)$  from the direct posterior to the PC-reformulated posterior, versus  $p$ .

## Distributed Parameters

- Consider that the unknown parameters are actually functions of space or time, i.e., spatial or temporal fields (distributed parameter estimation).



## Distributed Parameters

- Consider that the unknown parameters are actually functions of space or time, i.e., spatial or temporal fields (distributed parameter estimation).
- Estimating fields rather than parameters typically increases the ill-posedness of the inverse problem, since one is recovering an infinite-dimensional object from finite amounts of data.

## Distributed Parameters

- Consider that the unknown parameters are actually functions of space or time, i.e., spatial or temporal fields (distributed parameter estimation).
- Estimating fields rather than parameters typically increases the ill-posedness of the inverse problem, since one is recovering an infinite-dimensional object from finite amounts of data.
- One approach is to discretize the field on a finite set of grid points, however this presents difficulties for stochastic spectral approaches as the size of a PC basis does not scale favorably with dimension.

## Karhunen-Loève (K-L) expansion

- Ideally, one should employ a representation that reflects how much information is truly required to capture variation among realizations of the unknown field.

## Karhunen-Loève (K-L) expansion

- Ideally, one should employ a representation that reflects how much information is truly required to capture variation among realizations of the unknown field.
- To this end, they introduce a **Karhunen-Loève (K-L) expansion** based on the prior random process, transforming the inverse problem to inference on a truncated sequence of weights of the K-L modes.

## Karhunen-Loève (K-L) expansion

- Ideally, one should employ a representation that reflects how much information is truly required to capture variation among realizations of the unknown field.
- To this end, they introduce a **Karhunen-Loève (K-L) expansion** based on the prior random process, transforming the inverse problem to inference on a truncated sequence of weights of the K-L modes.
- In particular, the K-L representation of a scaled Gaussian process prior defines the uncertainty that is propagated through the forward model with a stochastic Galerkin scheme.

## Karhunen-Loève (K-L) expansion

- Ideally, one should employ a representation that reflects how much information is truly required to capture variation among realizations of the unknown field.
- To this end, they introduce a **Karhunen-Loève (K-L) expansion** based on the prior random process, transforming the inverse problem to inference on a truncated sequence of weights of the K-L modes.
- In particular, the K-L representation of a scaled Gaussian process prior defines the uncertainty that is propagated through the forward model with a stochastic Galerkin scheme.
- The deterministic forward model, originally specified by (a system of) partial differential equations, is thus replaced by stochastic PDEs.

- Let  $M(x)$  be a real-valued field endowed with a Gaussian process prior with mean  $\mu(x)$  and covariance kernel  $C$ , we denote this as  $M \sim \mathcal{GP}(\mu, C)$ .

- Let  $M(x)$  be a real-valued field endowed with a Gaussian process prior with mean  $\mu(x)$  and covariance kernel  $C$ , we denote this as  $M \sim \mathcal{GP}(\mu, C)$ .
- Introduce the corresponding  $K$ -term K-L representation of  $M$

$$M_K(x, \omega) = \mu(x) + \sum_{k=1}^K \sqrt{\lambda_k} c_k(\omega) \phi_k(x).$$



- Let  $M(x)$  be a real-valued field endowed with a Gaussian process prior with mean  $\mu(x)$  and covariance kernel  $C$ , we denote this as  $M \sim \mathcal{GP}(\mu, C)$ .
- Introduce the corresponding  $K$ -term K-L representation of  $M$

$$M_K(x, \omega) = \mu(x) + \sum_{k=1}^K \sqrt{\lambda_k} c_k(\omega) \phi_k(x).$$

- The eigenvalues  $\lambda_k$  and eigenfunctions  $\phi_k$  satisfy

$$\int C(x_1, x_2) \phi_k(x_2) dx_2 = \lambda_k \phi_k(x_1).$$

- Let  $M(x)$  be a real-valued field endowed with a Gaussian process prior with mean  $\mu(x)$  and covariance kernel  $C$ , we denote this as  $M \sim \mathcal{GP}(\mu, C)$ .
- Introduce the corresponding  $K$ -term K-L representation of  $M$

$$M_K(x, \omega) = \mu(x) + \sum_{k=1}^K \sqrt{\lambda_k} c_k(\omega) \phi_k(x).$$

- The eigenvalues  $\lambda_k$  and eigenfunctions  $\phi_k$  satisfy

$$\int C(x_1, x_2) \phi_k(x_2) dx_2 = \lambda_k \phi_k(x_1).$$

- In this example the  $c_k$  are Gaussian and independent,  $c_k \sim \mathcal{N}(0, 1)$ .

- Consider the following model problem

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x} \left( v(x) \frac{\partial u}{\partial x} \right) + S(x, t)$$

where  $S$  is a known gaussian source term active on a finite interval of time, and  $v(x)$  is to be inferred from (noisy) observations of  $u$  at discrete points in space and time.

- Consider the following model problem

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x} \left( v(x) \frac{\partial u}{\partial x} \right) + S(x, t)$$

where  $S$  is a known gaussian source term active on a finite interval of time, and  $v(x)$  is to be inferred from (noisy) observations of  $u$  at discrete points in space and time.

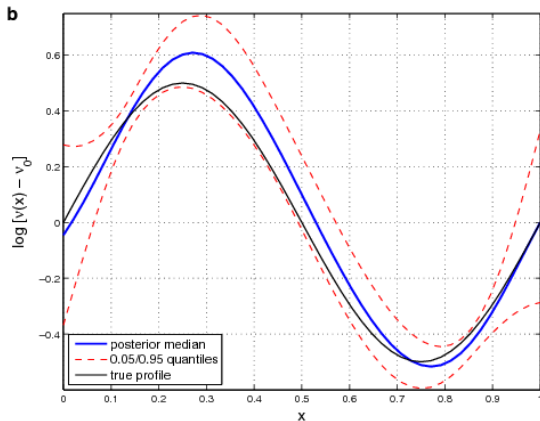
- Taking a truncated K-L expansion of  $v$  leaves coefficients  $c_k$  (each with Gaussian prior) to be determined conditioned on data.

- Consider the following model problem

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x} \left( v(x) \frac{\partial u}{\partial x} \right) + S(x, t)$$

where  $S$  is a known gaussian source term active on a finite interval of time, and  $v(x)$  is to be inferred from (noisy) observations of  $u$  at discrete points in space and time.

- Taking a truncated K-L expansion of  $v$  leaves coefficients  $c_k$  (each with Gaussian prior) to be determined conditioned on data.
- Treat the vector of  $c_k$  as  $\vec{c} = g(\xi)$  and proceed as before with PC and Galerkin projection to solve the stochastic forward problem.



[3] Fig 8b. K-L-based inversion of the sinusoidal log-diffusivity profile,  
 $K = 10$ .