Surrogate and Reduced Modeling for High-Dimensional Inverse Problems

Jinglai Li, Chad Lieberman, Youssef Marzouk, Karen Willcox

DOE Applied Mathematics Annual Program Meeting
October 17, 2011

DOE Awards DE-FG02-08ER25858 and DESC00025217 "Large-Scale Optimization for Bayesian Inference in Complex Systems"

Motivation: Large-scale statistical inverse problems

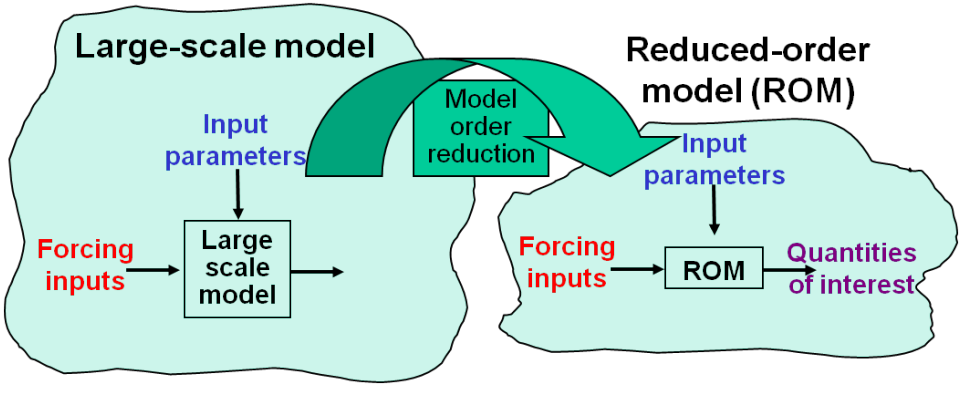
- Statistical inverse problem: formalized process of determining unobservable system properties and associated uncertainties through fusion of experimental data and computational models.
  - Especially challenging for large-scale PDE-based simulation models with high-dimensional parameters

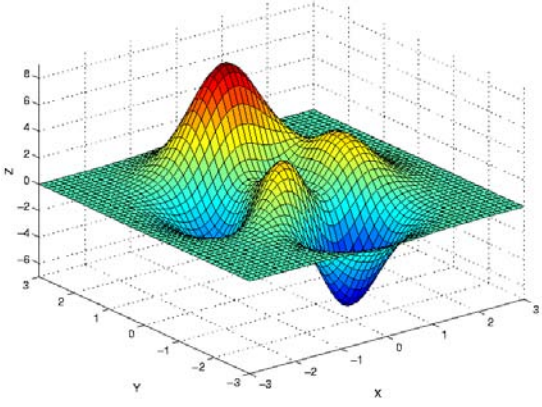
- Central to performing predictive simulations and ultimately decision under uncertainty for many DOE applications.
  - e.g., global climate change, nuclear waste repositories, groundwater contamination, carbon sequestration, clean combustion, coal gasification, nuclear fuel cycle, …

Research objectives

- Develop scalable numerical algorithms for large-scale Bayesian inversion in complex systems
  - Exploit the structure of the underlying mathematical model
  - Capitalize on advances in large-scale simulation-based optimization and inversion methods

- Develop new approaches using concepts from projection-based reduced-order modeling and stochastic spectral approximations
  - Combined with a goal-oriented view to overcome the challenges of high-dimensional parameters





Parameterized dynamical systems

Arising, for example, from spatial discretization of partial differential equations describing the system of interest.

$$\boxed{\begin{aligned} \dot{\mathbf{x}} &= \mathbf{A}(\mathbf{p})\mathbf{x} + \mathbf{B}(\mathbf{p})\mathbf{u} \\ \mathbf{y} &= \mathbf{C}(\mathbf{p})\mathbf{x} \end{aligned}} \qquad \boxed{\begin{aligned} \dot{\mathbf{x}} &= f(\mathbf{x}, \mathbf{p}, \mathbf{u}) \\ \mathbf{y} &= g(\mathbf{x}, \mathbf{p}, \mathbf{u}) \end{aligned}}$$

$\mathbf{x} \in \mathbf{R}^N$: state vector (e.g., flow unknowns)

$\mathbf{u} \in \mathbf{R}^{N_i}$: input vector (e.g., boundary forcing)

$\mathbf{p} \in \mathbf{R}^{N_p}$: parameter vector (e.g., properties)

$\mathbf{y} \in \mathbf{R}^{N_o}$: output vector (e.g., forces, moments)

Projection-based reduced-order models

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{A}(\mathbf{p})\mathbf{x} + \mathbf{B}(\mathbf{p})\mathbf{u} \\ \mathbf{y} &= \mathbf{C}(\mathbf{p})\mathbf{x} \end{aligned}$$

$\mathbf{x} \approx \mathbf{V}\mathbf{x}_r$

$$\begin{aligned} \mathbf{r} &= \mathbf{V}\dot{\mathbf{x}}_r - \mathbf{A}\mathbf{V}\mathbf{x}_r - \mathbf{B}\mathbf{u} \\ \mathbf{y}_r &= \mathbf{C}\mathbf{V}\mathbf{x}_r \end{aligned}$$

$\mathbf{W}^T\mathbf{r} = 0$

$$\begin{aligned} \mathbf{A}_r(\mathbf{p}) &= \mathbf{W}^T\mathbf{A}(\mathbf{p})\mathbf{V} \\ \mathbf{B}_r(\mathbf{p}) &= \mathbf{W}^T\mathbf{B}(\mathbf{p}) \\ \mathbf{C}_r(\mathbf{p}) &= \mathbf{C}(\mathbf{p})\mathbf{V} \end{aligned}$$

$$\begin{aligned} \dot{\mathbf{x}}_r &= \mathbf{A}_r(\mathbf{p})\mathbf{x}_r + \mathbf{B}_r(\mathbf{p})\mathbf{u} \\ \mathbf{y}_r &= \mathbf{C}_r(\mathbf{p})\mathbf{x}_r \end{aligned}$$

$\mathbf{x} \in \mathbf{R}^N$: state vector

$\mathbf{p} \in \mathbf{R}^{N_p}$: parameter vector

$\mathbf{u} \in \mathbf{R}^{N_i}$: input vector

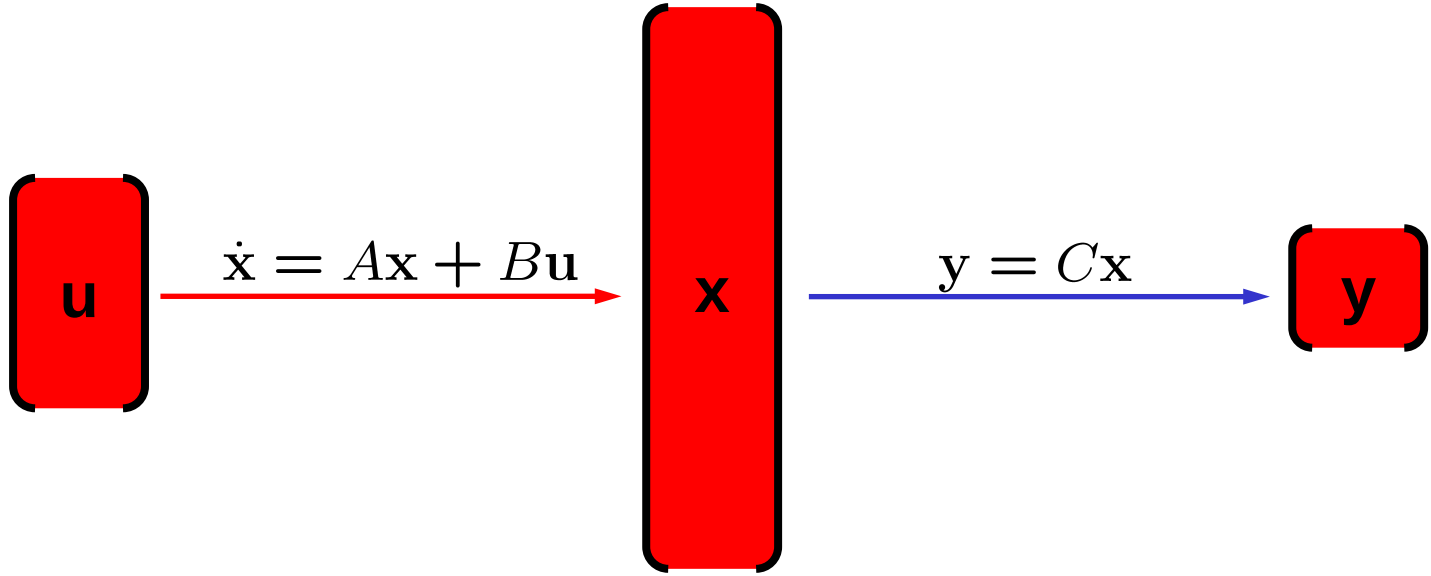
$\mathbf{y} \in \mathbf{R}^{N_o}$: output vector

$\mathbf{x}_r \in \mathbf{R}^n$: reduced state vector

$\mathbf{V} \in \mathbf{R}^{N \times n}$: reduced basis

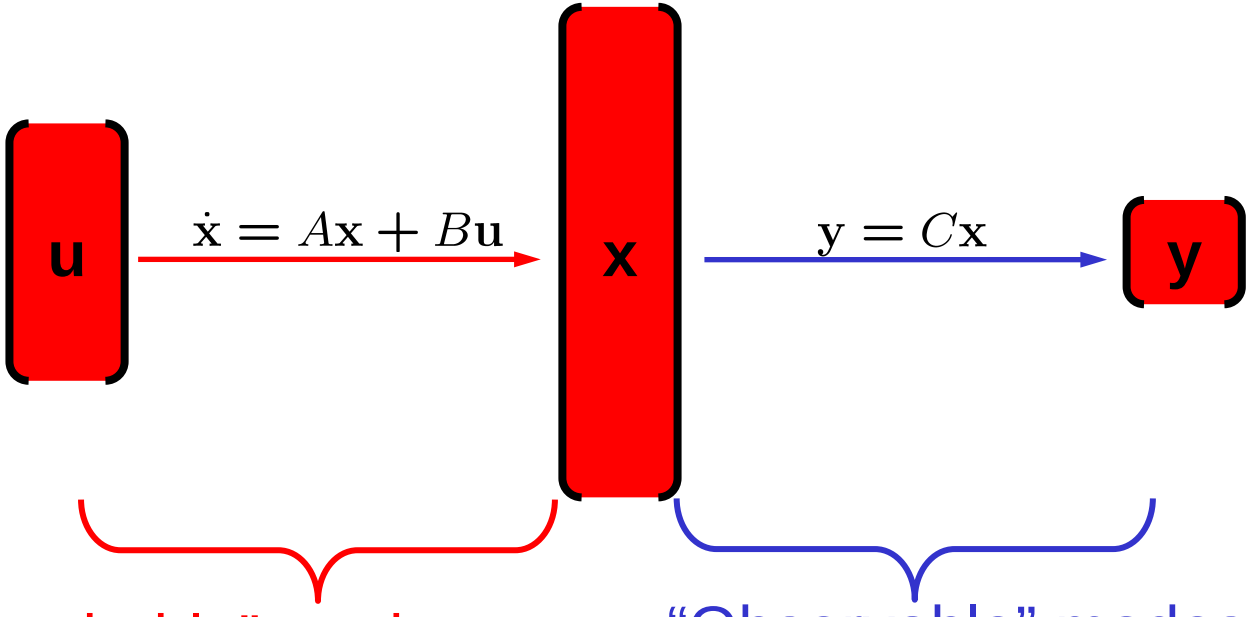
Why does model reduction work?

- Input $\rightarrow$ Output map is often much simpler than the full simulation model suggests



Why does model reduction work?

- In the linear case, the complexity of the input—output map can be quantified in rigorous terms



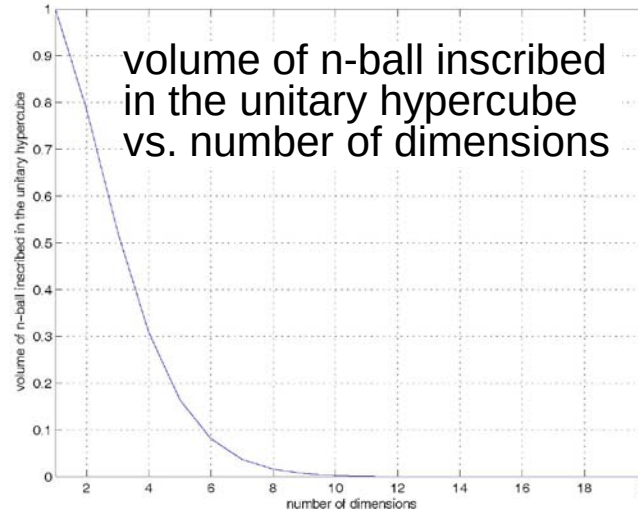
"Reachable" modes
- easy to reach
- dominant eigenmodes of a controllability Gramian

"Observable" modes
- generate large output energy
- dominant eigenmodes of an observability Gramian

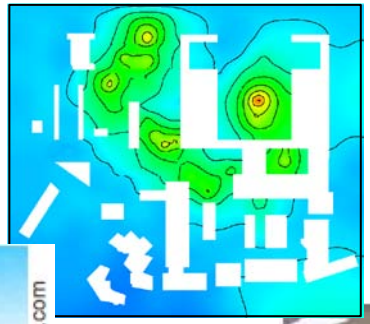
"Hankel singular values are to model order what singular values are to matrix rank." (Matlab hsvd documentation)

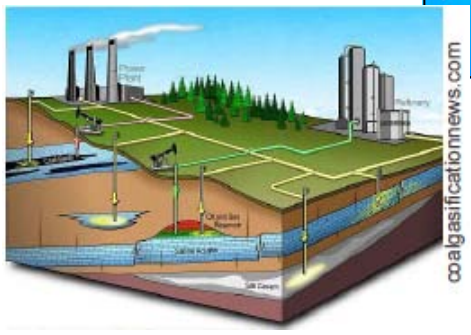
High-dimensional parameters: Why a challenge?

- Most model reduction methods sample the parameter space to build the basis



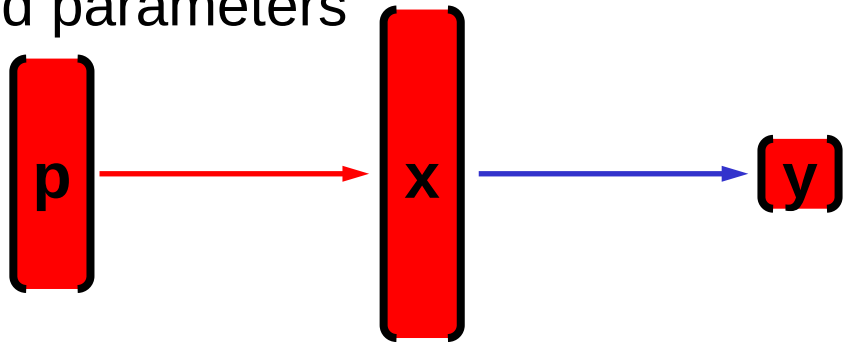
http://en.wikipedia.org/wiki/Sensitivity_analysis





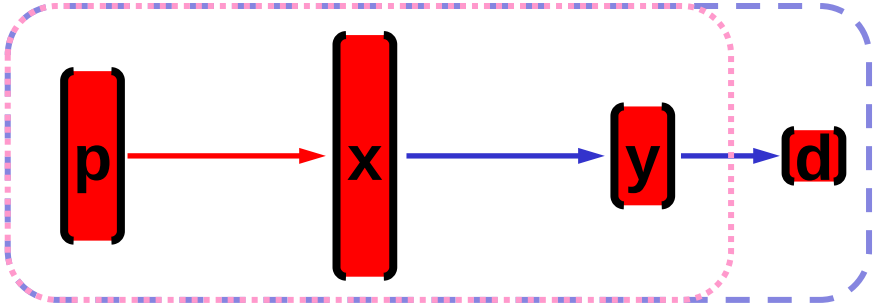
coalgasificationnews.com

- If we have many/distributed parameters (hundreds, thousands) can we really expect the input—output map to be low-dimensional?



High-dimensional parameters: There is hope

- Even if the parameter space is of high dimension, the outputs of interest are often of very low dimension
  - Engineering decisions are usually of low dimension (~1)
- If you have many outputs, is your system really encompassing the ultimate prediction/decision?

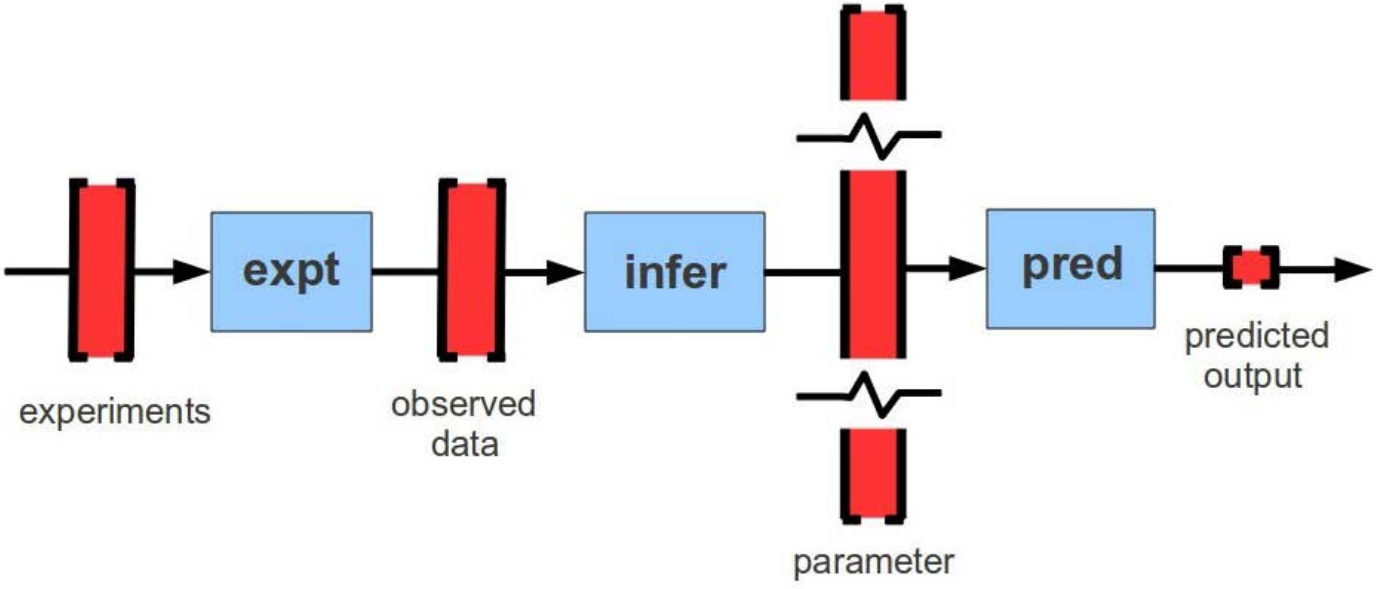


- Our approach:
  - Formulate the problem to account for the ultimate decision/prediction quantity of interest
  - Re-parameterize the problem in a goal-oriented manner (reduce parameter dimension) *(Lieberman, W., Ghattas; SISC 2010)*
  - Use optimization to search a high-dimensional space efficiently in a goal-oriented manner *(Bui-Thanh, W., Ghattas; SISC 2008)*

Inverse problem reformulation: Exploiting the data-to-prediction map

Lieberman, W.; in prep.

- Experimental data: low-dimensional $O(10^2)$
- Parameter: high-dimensional $O(10^5)$
- Prediction output of interest: low-dimensional $O(1)$



- Identify reduced parameter subspace in which to perform the inference, *with the goal of accurate predictions* → "inference for prediction"

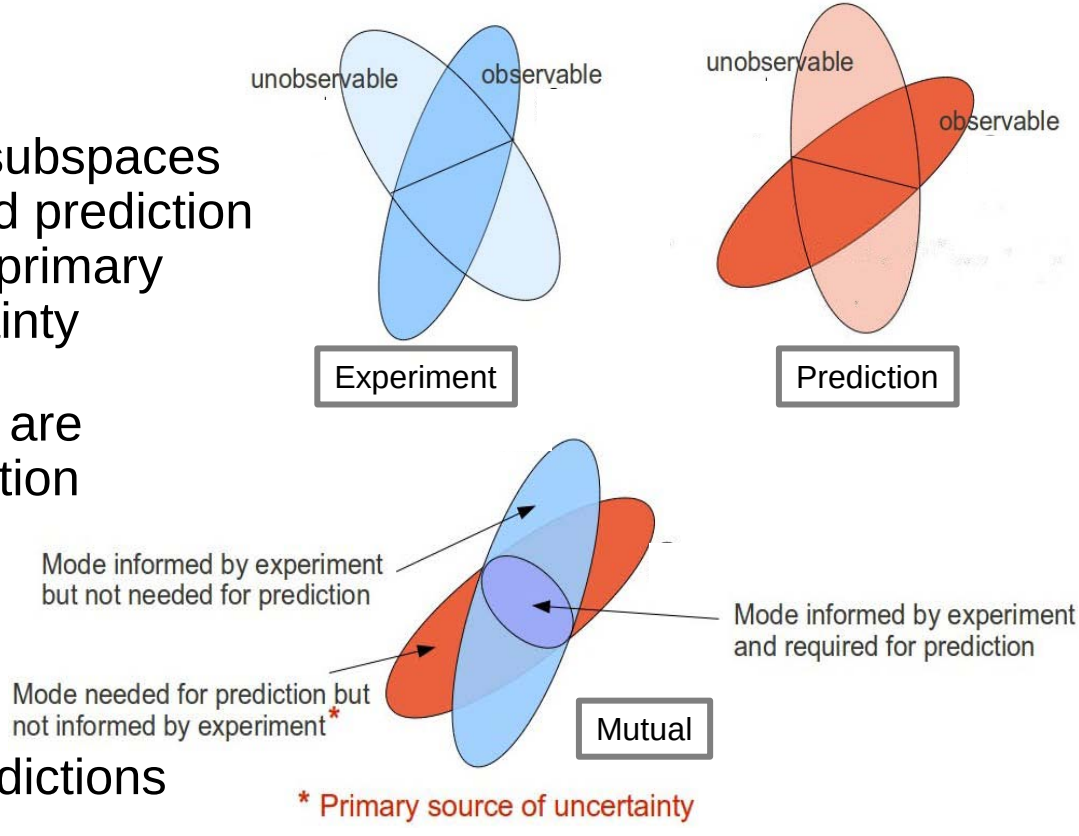
- An example where making the system boundary larger may make the problem "simpler"

A control-theoretic approach to "inference for prediction"

Lieberman, W.; in prep.

- Parameter (high-dimensional): $\mathbf{p}$
- Measurements (sparse): $\mathbf{y}_e = \mathbf{O}_e \mathbf{p}$
- Prediction outputs of interest: $\mathbf{y}_p = \mathbf{O}_p \mathbf{p}$

- Separate observable subspaces for experiment ($\mathbf{V}_e$) and prediction ($\mathbf{V}_p$) processes reveal primary contributors of uncertainty
- Determine modes that are experiment and prediction observable ($\mathbf{W}$)
- Sacrifice inversion accuracy but maintain accuracy in output predictions



Inference for prediction: Linear case

Lieberman, W.; in prep.

Truncated SVD:

- Inverse problem solution is the vector in the range of O_e that minimizes the data mismatch

$$\begin{aligned}\hat{\mathbf{p}}^{TSVD} &= \mathbf{V}_e \left\{ \arg\min_{\mathbf{a}\in R^r} \frac{1}{2}\|\mathbf{O}_e\mathbf{V}_e\mathbf{a} - \mathbf{y}_e\|_2^2 \right\} \\ \mathbf{y}_p^{TSVD} &= \mathbf{O}_p\hat{\mathbf{p}}^{TSVD}\end{aligned}$$

Inference for prediction:

- Find a low-dimensional basis $\mathbf{W}$ such that only the modes of $\mathbf{p}$ informed by experiment <u>and</u> required for prediction are inverted

$$\begin{aligned}\hat{\mathbf{p}}^{IFP} &= \mathbf{W} \left\{ \arg\min_{\mathbf{b}\in R^s} \frac{1}{2}\|\mathbf{O}_e\mathbf{W}\mathbf{b} - \mathbf{y}_e\|_2^2 \right\} \\ \mathbf{y}_p^{IFP} &= \mathbf{O}_p\hat{\mathbf{p}}^{IFP}\end{aligned}$$

Inference for prediction: Linear theory

Algorithm 1: IFP Basis Generation (TSVD)

Step 1. Define $\mathbf{G} = \mathbf{V}_e \mathbf{L}_e^{-1} \mathbf{V}_e \mathbf{O}_e^T$, where $\mathbf{V}_e \mathbf{L}_e \mathbf{V}_e^T$ is the eigendecomposition of $\mathbf{O}_e^T \mathbf{O}_e$.

Step 2. Compute the eigendecomposition $\mathbf{\Psi} \mathbf{\Sigma}^2 \mathbf{\Psi}^T$ of the "Gramian product" $\mathbf{G}^T \mathbf{O}_p^T \mathbf{O}_p \mathbf{G}$.

Step 3. Define the IFP basis:

$$\mathbf{W} = \mathbf{G} \mathbf{\Psi} \mathbf{\Sigma}^{-1/2}.$$

Theoretical Result: The prediction y_p^{IFP} obtained by using Algorithm 1 to generate the basis $\mathbf{W}$ is equal to the prediction obtained by the TSVD approach:

$$y_p^{IFP} = y_p^{TSVD}$$

Similar algorithms for analogous treatment of Tikhonov-regularized inverse problems and linear-Gaussian statistical inverse problems.

Posterior predictive covariance

- Statistical setting: parameter estimated by a probability distribution; prediction inherits stochasticity from parameter estimate.
- Linear Gaussian case: mean of the prediction $\mathrm{E}[\mathbf{y}_p]$ is given by IFP solution of an associated Tikhonov-regularized problem
- Prediction covariance $\mathrm{cov}(\mathbf{y}_p)$ is essential for treating the uncertainty (e.g., in design or control)

Theoretical Result: The IFP methodology leads to prediction covariance essentially for free, since

$$\mathbf{O}_p \mathbf{\Gamma}_\pi \mathbf{O}_p^T = \mathbf{O}_p \mathbf{W} \mathbf{\Sigma} \mathbf{W}^T \mathbf{O}_p^T$$

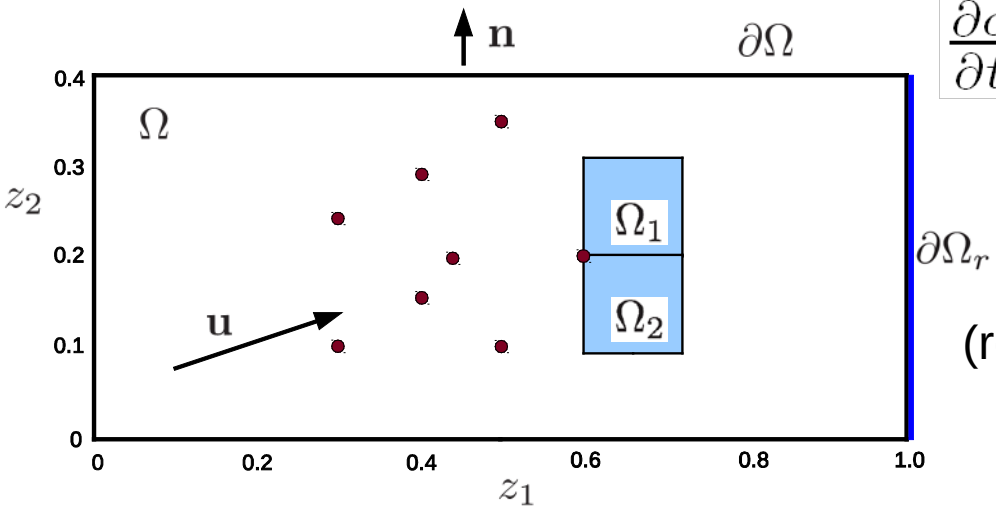
where

$$\mathbf{\Gamma}_\pi = (\Gamma_0^{-1} + \sigma^{-2} \mathbf{O}_e^T \mathbf{O}_e)^{-1}$$

is the covariance of the parameter estimate.

- Once the posterior predictive mean is computed by IFP, the covariance can be obtained for the cost of matrix multiplications.

Advection-diffusion example

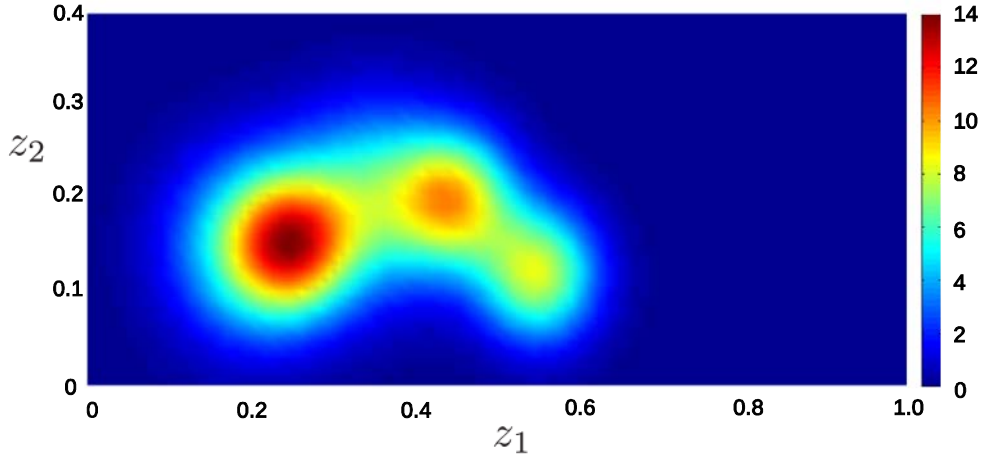


$$\frac{\partial c}{\partial t} = -\kappa \nabla^2 c + \mathbf{u} \cdot \nabla c$$

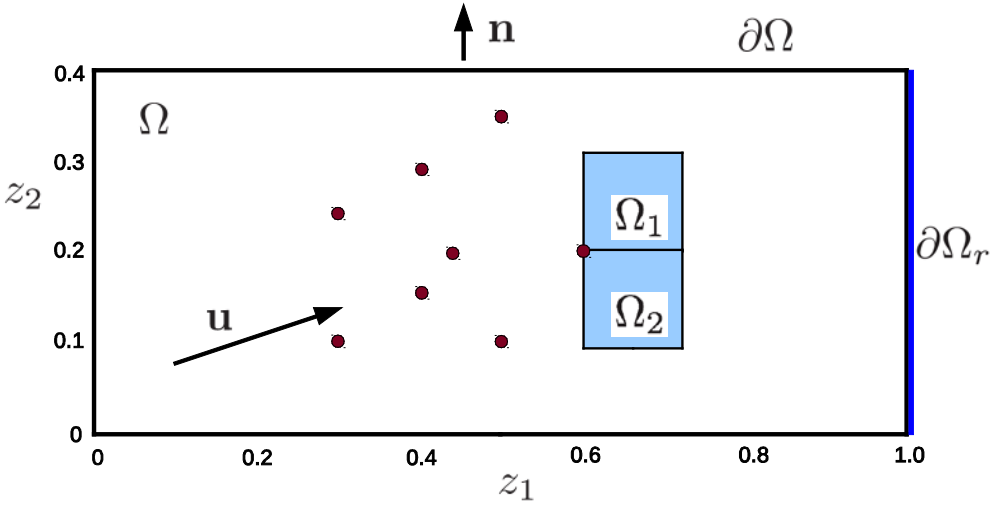
Computational domain

(red dots are sensor locations)

Prescribed initial condition



Inference for prediction: Truncated SVD

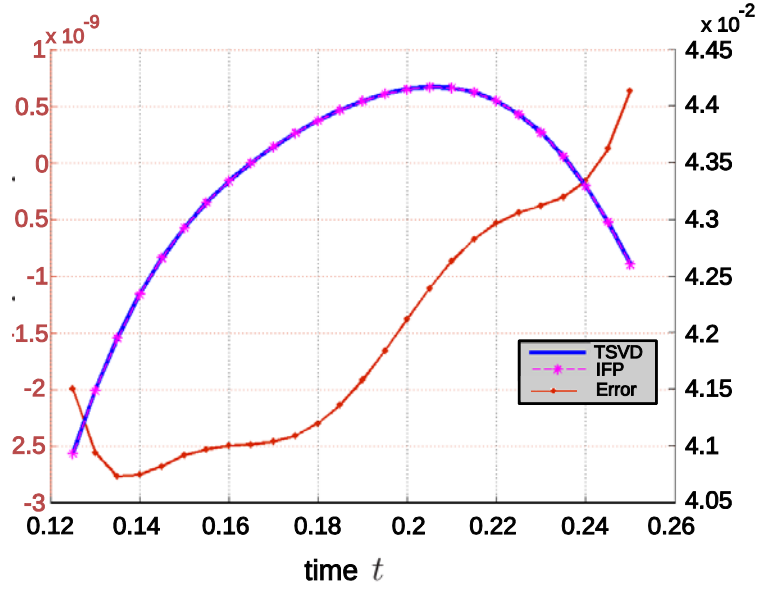


Prediction output of interest

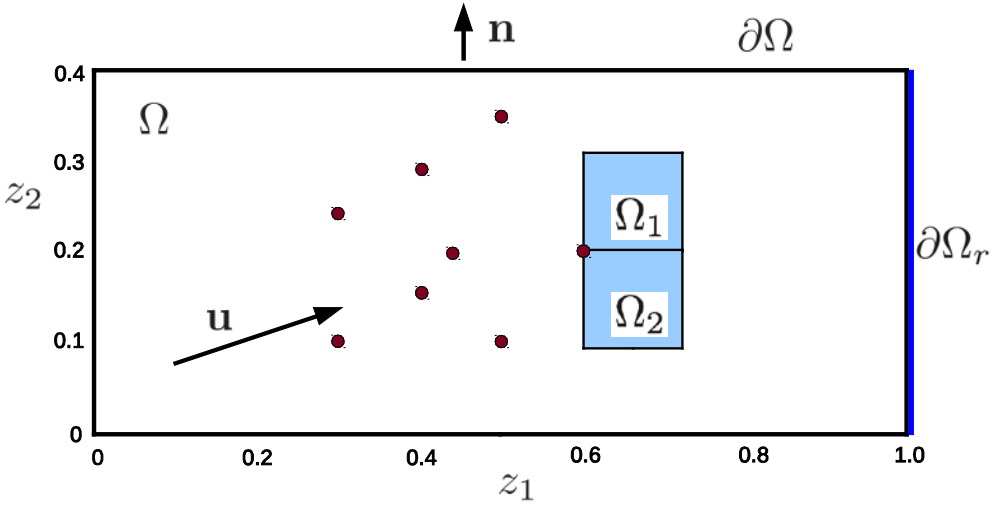
$$y_p(t) = \int_{\Omega_1} c dz_1 dz_2$$

Comparing predictions from IFP and TSVD approaches

r=54 ($\mathbf{V}_e$)
s=15 ($\mathbf{W}$)



Inference for prediction: Tikhonov regularization

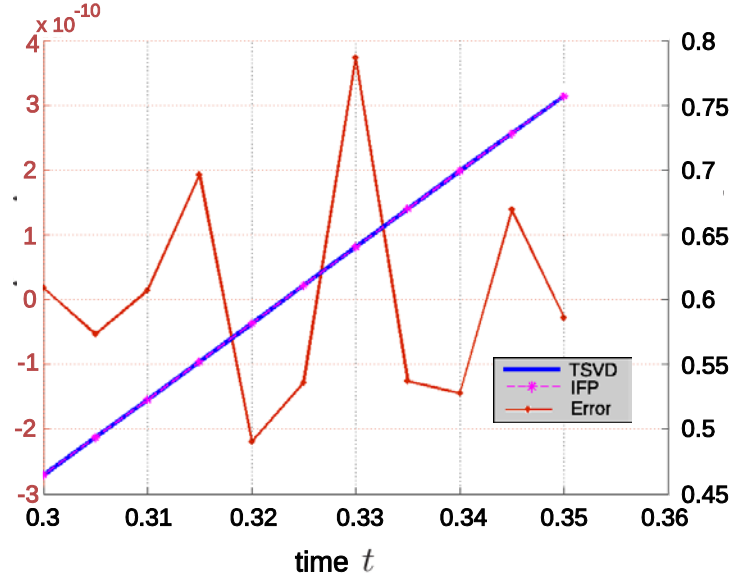


Prediction output of interest

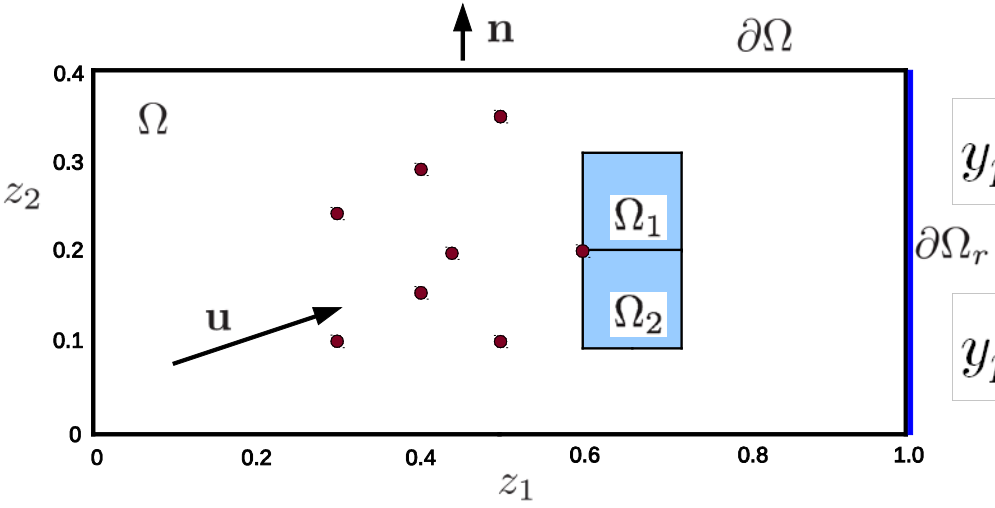
$$y_p(t) = \int_{\partial\Omega_r} c\mathbf{u} \cdot \mathbf{n} dz_2$$

Comparing predictions from IFP and Tikhonov-regularized approaches

r=4005 (=n)
s=11 ($\mathbf{W}$)



Inference for prediction: Statistical inverse problem



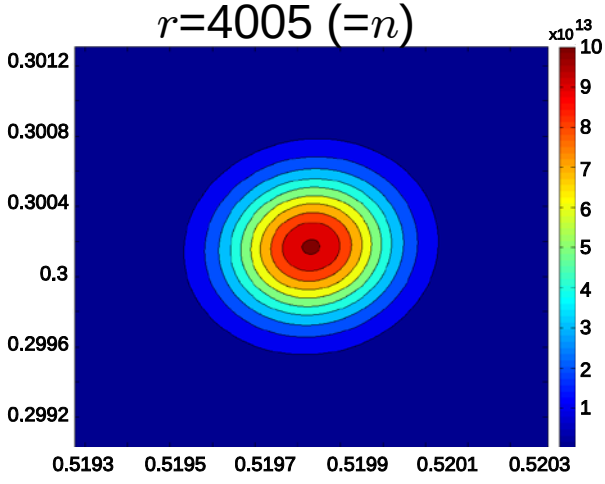
Prediction outputs of interest

$$y_{p_1} = \int_{25\Delta t}^{50\Delta t} \int_{\Omega_1} c dz_1 dz_2 dt$$

$$y_{p_2} = \int_{25\Delta t}^{50\Delta t} \int_{\Omega_2} c dz_1 dz_2 dt$$

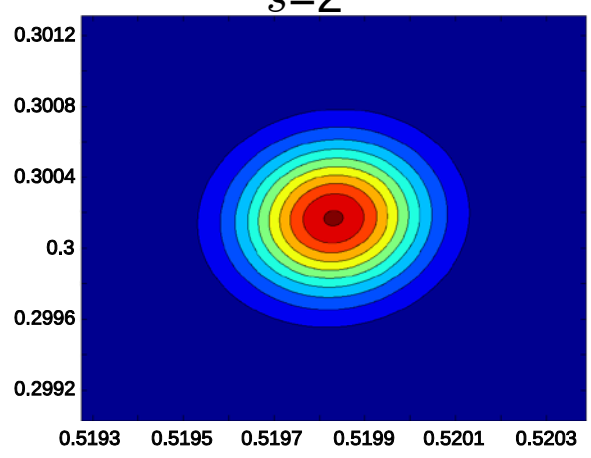
Traditional approach

r=4005 (=n)



IFP approach

s=2



Posterior predictive density function contours

Adaptive surrogates for nonlinear problems

- **Nonlinear forward models:** constructing an accurate surrogate over the entire parameter space may be prohibitive
  - Posterior concentrates on a small fraction of the prior support; particularly for high-dimensional problems
  - Localizing a surrogate mitigates the impact of nonlinearity
- **Inference problems:** can we construct a surrogate only over the support of the posterior? How to do this before characterizing the posterior?
- New **adaptive** approach, based on the cross-entropy method and importance sampling:
  - Construct a sequence of “cheap” surrogates and biasing distributions that converges to the posterior
  - Surrogates (e.g., polynomial chaos expansions) remain *local* and *low-order*

Adaptive surrogates for nonlinear problems

- Overall procedure:
  - Seek a biasing distribution that is close to the posterior $\pi^{\text{post}}(p) \propto L(p)\pi(p)$ (*p* is the parameter, *L* is the likelihood function, π is the prior)
  - Pick biasing distribution *q(p)* from a simple family of distributions, parameterized by *v*

$$\min_v D_{\text{KL}}\left(\pi^{post}(p) \,\|\, q(p;v)\right) \;\leftrightarrow\; \max_v \int L(p)\pi(p)\log q(p;v)\, dp$$

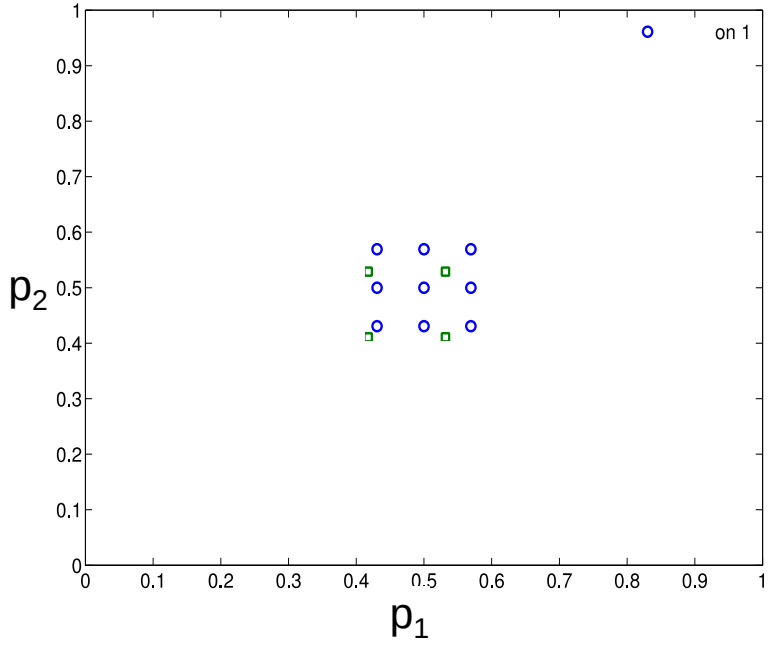
- Iterative approach:
  - Use sequential importance sampling to estimate the integral above, with a sequence of biasing distributions $q(p,v_m)$
  - At each iteration use a localized *surrogate,* based on *q(p)*, to evaluate the likelihood function

$$v_{m+1} = \arg\max_v \frac{1}{n}\sum_{i=1}^{n} L(p^{(i)})\log q(p^{(i)},v)\frac{\pi(p)}{q(p^{(i)},v_m)}$$

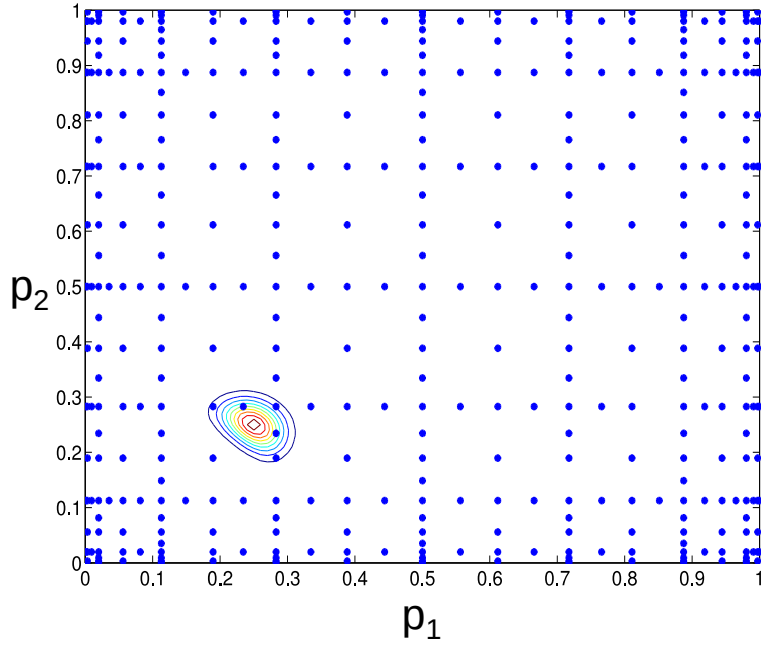
- Example: use Gaussian biasing distributions and Hermite polynomial chaos surrogates

Adaptive surrogates for nonlinear problems

- Example: 2-D source inversion problem
  (model evaluation points and *posterior density contours)*



adaptive surrogate



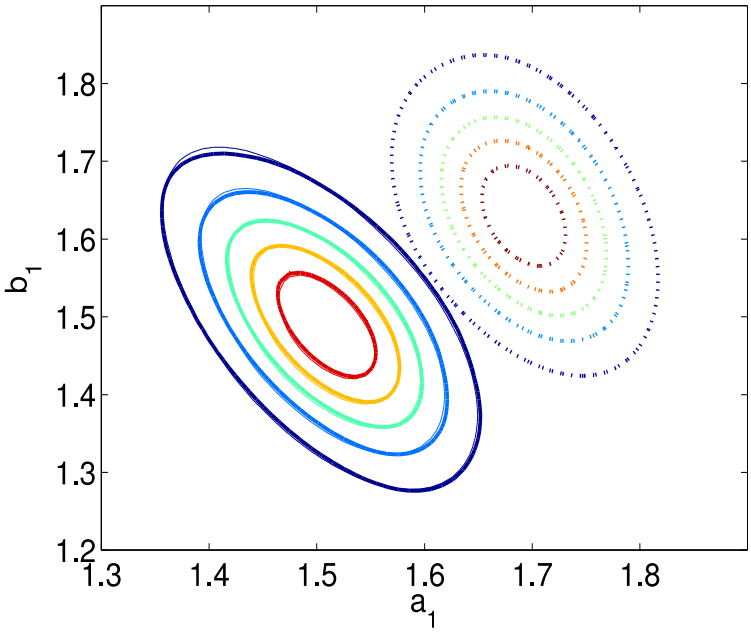
global surrogate

- Sparse grids used to construct polynomial chaos surrogates in both cases
- Number of model evaluations/polynomial order selected to ensure comparable accuracy!

Adaptive surrogates for nonlinear problems

- Example: nonlinear inverse heat conduction problem
  - Infer boundary heat flux from internal temperature measurements; note temperature-dependent conductivity *c(u)*
  - Heat flux parameterized with Fourier modes (11 dimensions)

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(c(u)\frac{\partial u}{\partial x}\right)$$

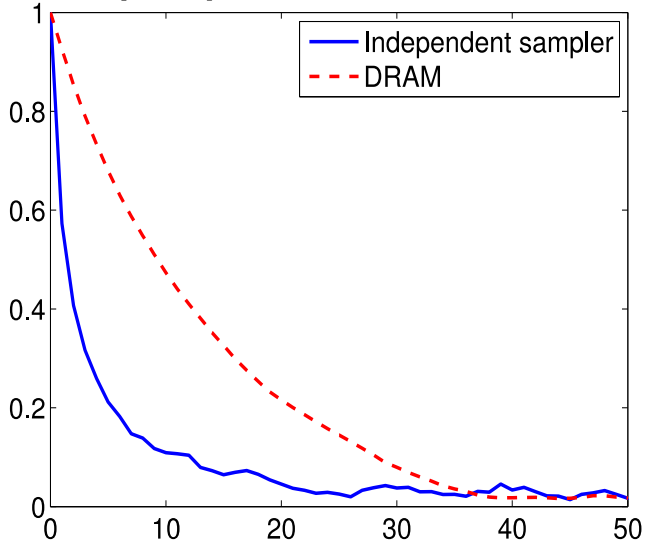


(thick solid line) full model
(thin solid line) **adaptive surrogate**
(dotted line) global surrogate

	# model evals	polynomial order	$D_{KL}\left(\pi \,\|\, \tilde{\pi}_{surr}\right)$
global surrogate	35929	5	8.37
adaptive surrogate	5763	2	0.0032

Adaptive surrogates for nonlinear problems

- Final biasing distribution also provides a good foundation for efficient MCMC sampling (e.g., use as proposal in an *independence sampler)*



- Method is not limited to polynomial chaos surrogates or Gaussian biasing distributions:
  - Projection-based reduced order models
  - *Mixtures* of *exponential family* distributions
- Dimensionality reduction is necessary and complementary

Conclusions

- Surrogate models provide dramatic speedups in solution of large-scale inverse problems
- For problems with high-dimensional parameter spaces, use goal-oriented approaches to overcome curse of dimensionality:
  - Formulate the problem to account for the ultimate decision/prediction quantity of interest and re-parameterize the problem in a goal-oriented manner
  - Adaptive approach using stochastic optimization methods to construct surrogates that are accurate over the support of the *posterior* distribution.
- In the linear case, our approach has rigorous theory and connections to balanced truncation
- Not all problems are amenable to model reduction
  - But many are, especially if you keep your goal in mind