Uncertainty Quantification in MD Simulations: Forward Propagation and Parameter Inference

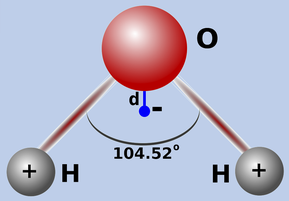
JOHNS HOPKINS UNIVERSITY WHITING SCHOOL OF ENGINEERING

O.M. Knio[1], F. Rizzi[1], H.N. Najm[2], B.J. Debusschere[2], K. Sargsyan[2], M. Salloum[2], H. Adalsteinsson[2]

[1]Johns Hopkins University, [2]Sandia National Laboratories

Motivation

- Water has been extensively and successfully investigated experimentally, but large uncertainty still characterizes water models widely employed in simulations.
- Common approaches are based on empirical force-field descriptions: e.g. RWK, SPC/E, TIP4P, TIP4P-Ew and TIP5P.
- Capable of reproducing *some* physical properties of water with good accuracy.



TIP4P water molecule

Goals

- Develop a framework to quantify uncertainty in Molecular Dynamics (MD) simulations of water, employing Polynomial Chaos (PC) representation of uncertainties.
- Analyze and control the effect of parametric uncertainty and intrinsic (thermal) noise.
- Determine or refine a set of force-field parameters for TIP4P water with a reformulated Bayesian inference approach using PC surrogates (Marzouk *et al.* 2007).

MD Simulations

- Cubic domain (xyz): 37.2 Å^3.
- Periodic boundary conditions.
- TIP4P water: 1728 molecules.
- NPT simulations: P=1 atm, T=298 K.
- LAMMPS (lammps.sandia.gov).
- Focus on time-averaged density (ρ), self-diffusivity (D) and enthalpy (H) collected during the *steady* state.
- Uncertainty in the force-field parameters $\Rightarrow$ MD predictions (ρ, D, H) become random variables.

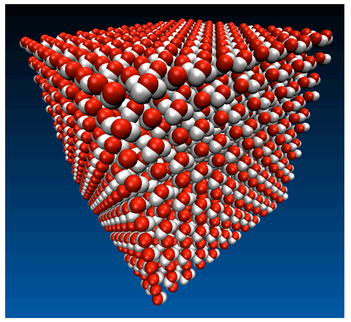


Figure: Atomistic system.

UQ Methods

- PC expansion of an $\mathbb{R}$-valued random variable (RV) X:

$$X = \sum_{i=0}^{\infty} c_i \Psi_i(\xi)$$

with orthogonal polynomials $\Psi_i(\xi)$ of normalized RV ξ:

$$\xi \sim \mathcal{U}[-1,1] \Longrightarrow \Psi_i \quad \text{Legendre polynomials}$$

- Generalizes to higher-dimensional spaces: $X = \mathcal{F}(\xi_1, \ldots, \xi_n)$.

1) Non-Intrusive Spectral Projection (NISP):

$$c_i = \frac{1}{\langle \Psi_i, \Psi_i \rangle} \int_\Omega X \Psi_i(\xi) p_f(\xi) d\xi$$

where $p_f(\xi)$ is the probability density of ξ.

- Typically computed using quadrature rules.
- Feasible for low-dimensional problems due to large computational costs associated with high-dimensional cases.
- Regularity of X with respect to ξ is typically assumed.

2) Inference-Based Spectral Methods exploit a set of data $\{X_j\}_{j=1}^N$:

$$\pi\left(c_1, c_2, \ldots \middle| \{X_j\}_{j=1}^N\right) \propto \mathcal{L}\left(\{X_j\}_{j=1}^N \middle| c_1, c_2, \ldots\right) \textit{Prior}$$

where π and $\mathcal{L}$ are the posterior and likelihood, respectively.

- No restrictions on the sampling scheme.
- Joint posterior for the c_i's, i.e. *non-deterministic* expansion.

Forward Propagation: NISP approach

- Uncertain force-field parameters $\boldsymbol{\theta} = \{\varepsilon, \sigma, d\}$, with (ε, σ) defining the Lennard-Jones potential. Stochastic reformulation of the forward problem:

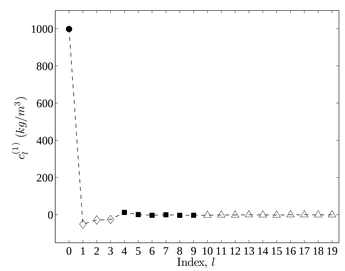
$$\boldsymbol{\theta} = \boldsymbol{f}(\boldsymbol{\xi}) \qquad (1)$$

- Let $\boldsymbol{G} = \{\rho, D, H\}$ be the set of uncertain observables. Assume:

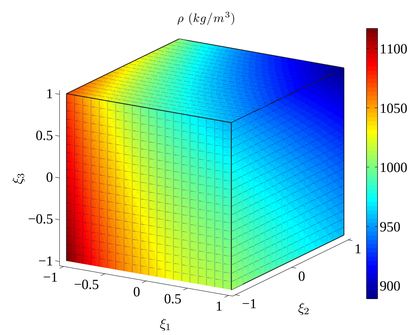
$$\overline{G}_k(\boldsymbol{\xi}) \approx M_k(\boldsymbol{\xi}) = \sum_{l=0}^{P} c_l^{(k)} \Psi_l(\boldsymbol{\xi}), \qquad k = 1,2,3, \qquad (2)$$

where the overline denotes averaging over m replicas.

- NISP using Gauss quadrature with 7 points along each dimension.



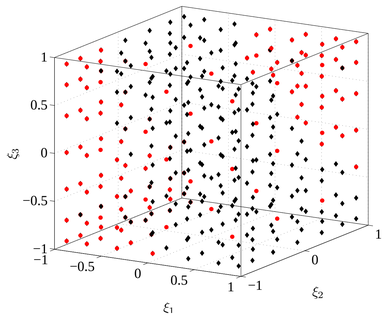
Spectrum of PC coefficients for ρ.



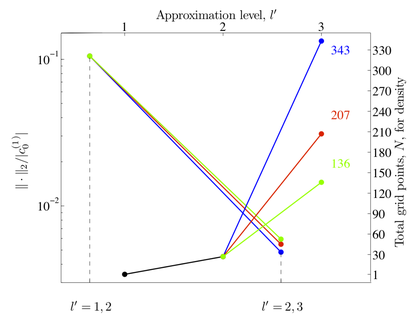
3D volume of $M_1(\boldsymbol{\xi})$.

Forward Propagation: Inference-Based approach

- Sampling technique combining nested Fejér grids and adaptive selection of new points at high resolution levels ($\ell \geq 3$).
- Choose $\xi^i : Z_k(\xi^i) \geq \lambda,\ i = 1, \ldots, N_\ell$, where Z_k represents the normalized "error" for M_k between two successive levels.



Full grid vs. reduced grid ($\lambda = 0.40$) for ρ.



Diagnostics as a function of the approximation level and tolerance: $\lambda = 0$, $\lambda = 0.25$, $\lambda = 0.40$.

Parameter Inference

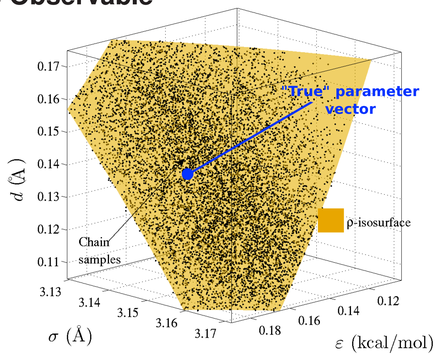
- Focus on a synthetic problem where presumed "true" values $\{\hat{\sigma}, \hat{\varepsilon}, \hat{d}\}$ are used to generate a collection of noisy data, $\boldsymbol{Q}_k$, for ρ ($k=1$), D ($k=2$) and H ($k=3$), then Bayesian inference is exploited to recover the "true" set of driving parameters.
- Exploit (1) and (2) to develop a PC surrogate-based Bayesian inference approach (Marzouk *et al.* 2007) with likelihood:

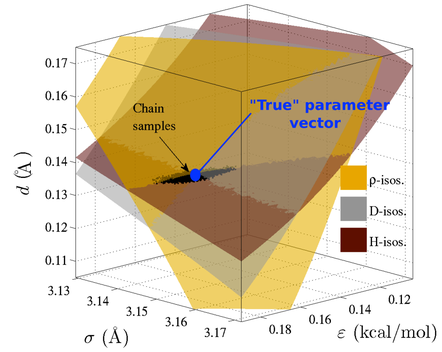
$$\mathcal{L}\left(\boldsymbol{Q}_1, \boldsymbol{Q}_2, \boldsymbol{Q}_3, \middle| \{\xi_1, \xi_2, \xi_3\}\right) = \prod_{k=1}^{3} \prod_{i=1}^{N} p_{\gamma_k}\left(Q_k^i - M_k(\xi_1, \xi_2, \xi_3)\right),$$

- Significant cost savings in evaluating the forward model for different parameter values during Markov chain Monte Carlo.
- We assume $\{\gamma_k\}_{k=1}^3 \sim \mathcal{N}(0, \tilde{\sigma}_k^2)$, with hyperparameters $\{\tilde{\sigma}_k^2\}_{k=1}^3$.
- $N = 10$ observations for each observable, i.e. a total of 30 data points, and cubic PC expansions as surrogate model.

Single Observable

- Chain moves over the isosurface extracted from the surrogate model $M_1(\xi)$ defined as $\xi : M_1(\xi) = \overline{\boldsymbol{Q}}_1$.
- Underdetermined problem.
- Large uncertainties in the inferred values.



Three Observables

- Chain localizes at the intersection of the isosurfaces extracted from the surrogate model and defined as $\xi : M_k(\xi) = \overline{\boldsymbol{Q}}_k$, $k = 1,2,3$.
- Surfaces of constant ρ, D and H intersect at a point, yielding a well-defined problem.
- True value is recovered with good accuracy and small uncertainty.