STOCHASTIC SIMULATION, OPTIMIZATION, INFERENC

A diving bird's eye view

Petros Koumoutsakos

Computational Science
ETH Zurich
THE IMPERFECT PATHS TO KNOWLEDGE

THE UNIVERSE of PHYSICAL REALITIES

OBSERVATIONS

EXPERIMENTS

THEORY / MATHEMATICAL MODELS

COMPUTATIONAL MODELS

VERIFICATION

VALIDATION

DECISION

KNOWLEDGE

Discretization Errors

Modeling Errors

Observation Errors

Data Errors

ADAPTED FROM: R. Moser, T. Oden, O. Ghattas, UT Austin
VERIFICATION
VALIDATION
THE IMPERFECT PATHS TO KNOWLEDGE
OBSERVATIONS
Discretization
Errors
DECISION
Modeling
Errors
Data
Errors
#&'&(41/4/QUGT61FGP1)JCVVCU76#WUVKP
THEORY /
MATHEMATICAL
MODELS
COMPUTATIONAL
MODELS
VERIFICATION
THE UNIVERSE
of PHYSICAL
REALITIES
Observational
Errors
THE QUEST FOR KNOWLEDGE

- All men by nature desire knowledge (Aristotle)
- Man has an intense desire for assured knowledge (Einstein)
- Knowledge is Power (Bacon)

KNOWLEDGE:
- true, justified belief (Plato).
- understanding, as opposed to opinion.
- quantifiable relationships between facts/observations and ideas.
  "He believes it, but it isn't so," vs. "He knows it, but it isn't so." (Wittgenstein)
• Knowledge driven design of a series of experiments:
  ◦ Account for all sources of uncertainty in the experimental campaign
  ◦ Have a full description of the system state and its evolution.

What we observe is not nature itself, but nature exposed to our method of questioning.

Emmanuel Kant: *Our intellect does not draw its laws from nature, but tries – with varying degrees of success – to impose upon nature laws which it freely invents.*
The Classic view:

• Not sufficient just to detect patterns in events

• Describe events with **few fundamental, deterministic, causal principles.**

• Describes well “simple” deterministic systems.
  • “simple” refers to conceptual simplicity, rather than ‘technical’ system complexity.
  • e.g. a mechanical system with many moving objects interacting with each other; each object is described by simple equations that involve just a few variables.
Classical Approach to Human Knowledge

**Classic Knowledge**: Describe many facts with few fundamental principles.

- The number of facts (data, observations) used to derive such knowledge is usually **small**.
- The cost of collecting or generating these observations is **high**.
- The principles may **not be “useful”** (i.e. non-predictive).
Empirical knowledge: useful dependencies (i.e. predictive) estimated from data or derived from experience.

In contrast to first-principle knowledge, empirical knowledge:
• Describes properties of “complex” systems that lack credible first-principle models (‘complexity’ refers to a large number of observed parameters/variables);
• is statistical, i.e., allows to make non-deterministic predictions, at best;
• has a quantifiable practical utility for a given application (i.e. the “Oracle”)
• Highly facilitated by the digital processing and acquisition of knowledge
\[ \hat{\mathbf{u}}'(\mathbf{x}, t) = \sum_{i=1}^{n} a_i(t) \cdot \phi_i(\mathbf{x}) \]

\[ \int \langle u_j'(\mathbf{x}, t)u_h'(\mathbf{x}', t) \rangle \phi_i(\mathbf{x}') \, dx' = \lambda_i \phi_i(\mathbf{x}), \]

WHAT KIND OF KNOWLEDGE MAY WE NEED?
Designing a 'smart wing' for the Mars airplane

By David Kenwright

One of the most familiar and most inspiring stories in the annals of aviation is the tale of the Wright brothers' flight in Kitty Hawk in 1903. The story is often told as an example of the ability of the Wright brothers to see the future and create something that no one else could have imagined. However, the truth is that the Wright brothers were not the first to think about the idea of a flying machine. Many others had tried and failed, and it was only through a combination of persistence, hard work, and a willingness to experiment that the Wright brothers were able to achieve success.

The key to their success was the way in which they approached the problem of designing a flying machine. They started with a series of experiments in which they tested different materials and configurations for their wings. They then used those results to design and build a series of test vehicles, each of which was more advanced than the last. The final version of the Wright Flyer was a small, lightweight aircraft with austria and a single propeller, and it was this model that made the first successful powered flight in history.

In many ways, the Wright brothers' approach to designing a flying machine was similar to the approach that is taken today in the development of new technologies. In both cases, the process involves a combination of experimentation, analysis, and iteration. And just as the Wright brothers were able to achieve success by taking a systematic and methodical approach to their work, so too can we expect to see continued progress in the development of new technologies in the future.
EMPIRICAL/DIGITAL KNOWLEDGE: Self-Optimizing Machines

- the NOx emission for solutions in box 5 is about 35% smaller than for the given standard setting.
- the pulsations for solutions in box 1 is about 30dB smaller than for the given standard setting.

Are humans the only beings capable of "knowledge"?
MOVE

\[ D \xrightarrow{d_\mu} D_{\text{moves}} \]
\[ K \xrightarrow{k_\mu} K_{\text{moves}} \]

UN-BIND

\[ \rho \xrightarrow{d_\alpha} D \]
\[ \rho \xrightarrow{k_\alpha} K \]

\[ D \xrightarrow{d_\delta} \rho \]
\[ K \xrightarrow{k_\delta} \rho \]

\[ \rho \xrightarrow{d_\alpha} D \xrightarrow{d_\delta} \rho \]
\[ \rho \xrightarrow{k_\alpha} K \xrightarrow{k_\delta} \rho \]

Stochastic Simulations: R-leaping and Flavor SSA
Stochastic Optimisation: CMA-ES
Bayesian Uncertainty: The Hierarchical Approach
STOCHASTICS

1. Continuum - Fokker Planck
2. Discrete - Markov Processes
\[
\begin{align*}
\frac{d}{dt}x_i &= v_i \\
\frac{m}{dt}v_i &= F_i
\end{align*}
\]
CONTINUUM - Fokker Planck
2. DISCRETE - Example: Predator Prey - The Lotka-Volterra Mode

**Deterministic ODE’s**

\[
\frac{d[X]}{dt} = k_1[A][X] - k_2[X][Y] \\
\frac{d[Y]}{dt} = k_2[X][Y] - k_3[Y]
\]

**Stochastic Interpretation**

\[
P(x \rightarrow x + 1; y \rightarrow y) = k_1 a x \, dt \\
P(x \rightarrow x - 1; y \rightarrow y + 1) = k_2 x y \, dt \\
P(x \rightarrow x; y \rightarrow y - 1) = k_3 y \, dt \\
P(x \rightarrow x; y \rightarrow y) = 1 - (k_1 a x + k_2 x y + k_3 y) \, dt
\]

X: Sheeps, Y: Wolves
A: Grass

\[X + A \overset{k_1}{\rightarrow} 2X\]

\[X + Y \overset{k_2}{\rightarrow} 2Y\]

\[Y \overset{k_3}{\rightarrow} B\]
Spatially-Dependent Stochastic processes

- **Diffusion in 1-D**

  - A species $U$, whose elements are labeled by an index $i$

  - Diffusion as a set of events:
    
    $U_i \xrightarrow{k_{i,i+1}} U_{i+1}$
    
    $U_i \xrightarrow{k_{i,i-1}} U_{i-1}$
    
    $U_{i+1} \xrightarrow{k_{i+1,i}} U_i$
    
    $U_{i-1} \xrightarrow{k_{i-1,i}} U_i$

- Uniform Cells:
  
  $k_{i,j} = \frac{D}{h^2}$

  $\mathcal{C}_i = \{c_i, c_{i+1}\}$

  $k_{i,j} = \frac{D}{h^2}$

  $c_{i-1} \leftrightarrow c_i \leftrightarrow c_{i+1}$

The Spatial Gray-Scott Model

\[
\frac{\partial u}{\partial t} = d_u \Delta u - uv^2 + F(1 - u), \\
\frac{\partial v}{\partial t} = d_v \Delta v + uv^2 - (F + \kappa)v.
\]

\[U + 2V \rightarrow 3V, \quad V \rightarrow P\]
Molecules per grid cell for a 300 x 300 grid

500  1000  5000  10000

$F = 0.04, \kappa = 0.06, t = 1000$
Stochastic Simulations of Reaction-Diffusion


\[
\frac{dp_\sigma(t)}{dt} = \sum_{\sigma'} \left\{ W(\sigma' \rightarrow \sigma)p_{\sigma'}(t) - W(\sigma \rightarrow \sigma')p_\sigma(t) \right\}
\]
Delta - Notch Signalling (Collier et. al, 1996)

Diagram showing the steady-state pattern of primary and secondary fates in an (a)7x7 and (b) 8x8 array of hexagonal cells [Black cells will adopt the primary fate and white cells the secondary fate]
Delta - Notch Signalling (Collier et al., 1996)

Deterministic

Stochastic
Juxtacrine Signaling

**Definition**
Intercellular signaling induced by physical cell contact. Mainly involved in
- developmental processes,
- wound healing,
- angiogenesis

**Examples**
- Receptor: Notch  Ligands: Delta, Jagged, Serrate
- Receptor: EGFR  Ligand: TGFα

**Pattern formation**
through positive and negative feedback loops
(autoinduced up-regulation and down-regulation of ligand and receptor expression)
Computational Models (EGFR/TGFα)

3 Species: $R$ Receptor, $L$ Ligand, $B$ Bound receptor (=R-L complex)

**Deterministic**

\[
\frac{\partial R_i}{\partial t} = -k_a R_i \langle L \rangle_i + k_d B_i - d_R R_i + \mathcal{P}_r(B_i)
\]

\[
\frac{\partial L_i}{\partial t} = -k_a \langle R \rangle_i L_i + k_d \langle B \rangle_i - d_L L_i + \mathcal{P}_l(B_i)
\]

\[
\frac{\partial B_i}{\partial t} = k_a R_i \langle L \rangle_i - k_d B_i - k_{\text{int}} B_i
\]

2x positive feedback:

\[
\mathcal{P}_l(B) = \frac{C_1^m B^m}{C_2^m + B^m}
\]

\[
\mathcal{P}_r(B) = \frac{C_4^m B^n}{C_5^m + B^n}
\]

Spatial operator $\langle \cdot \rangle_i =$ CONCENTRATION ON CELLS $j$

TOUCHING CELL $i$

**Stochastic** e.g. 1D

\[
R_i + L_{i+1} \xrightarrow{1/2 k_a} B_i
\]

\[
R_i + L_{i-1} \xrightarrow{1/2 k_a} B_i
\]

\[
B_i \xrightarrow{1/2 k_d} R_i + L_{i+1}
\]

\[
B_i \xrightarrow{1/2 k_d} R_i + L_{i-1}
\]

\[
L_i \xrightarrow{d_L} R_i \xrightarrow{d_R} B_i \xrightarrow{k_i} \text{DECAY 

& INTERNALIZATION}
\]

\[
\mathcal{P}_l(B_i) \xrightarrow{} L_i
\]

\[
\mathcal{P}_r(B_i) \xrightarrow{} R_i
\]

Feedback

*Owen & Sherratt, Mathematical Biosciences, 1998*
Stochastic vs. Det. simulation

**Spatiality:**
Species of different cells $i$ encoded as different species, i.e.

$$B_1, R_1, L_1, B_2, R_2 \ldots, L_N$$

For 70 cells: 770 reactions, prohibitively expensive with SSA → R-leap

**Patterning after 100h:**

<table>
<thead>
<tr>
<th>Cell ID</th>
<th>B</th>
<th>R</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>3000</td>
<td>462</td>
<td>3000</td>
</tr>
<tr>
<td>3000</td>
<td>3000</td>
<td>462</td>
<td>3000</td>
</tr>
<tr>
<td>4000</td>
<td>3000</td>
<td>462</td>
<td>3000</td>
</tr>
<tr>
<td>5000</td>
<td>3000</td>
<td>462</td>
<td>3000</td>
</tr>
<tr>
<td>6000</td>
<td>3000</td>
<td>462</td>
<td>3000</td>
</tr>
</tbody>
</table>

Initial condition: $B_j = R_j = 3000, L_j = 462 + 10\%$ Uniform noise.

Deterministic vs. Stochastic
Dynamics: Stochastic vs. Deterministic

Energy of 1-spike mode

\[ E = \sqrt{\sum_{j \in \text{cells}} |B_j - \frac{1}{2}(B_{j+1} + B_{j-1})|^2} \]

Deterministic simulation is **unable** to assess dynamics of noisy systems
ACCELERATED

STOCHASTIC SIMULATION ALGORITHMS
Chemical kinetics: Set-up

- Well stirred reaction volume $V$
- $N$ different species $S_1, S_2, \ldots, S_N$ in numbers $X_1, X_2, \ldots, X_N$
- random collisions and reactions through $M$ channels $R_1, R_2, \ldots, R_M$
- Experiment length $T$
Stochastic Simulation Algorithm (SSA)

- Gillespie 1977

- Example (Well-stirred volume $V$):
  - a single 2$^{\text{nd}}$ order reaction $A+B \rightarrow C$
  - Probability of $A$-$B$ collision within $dt \sim X_A X_B dt$
  - Probability of reaction within $dt$ is $k_{AB} X_A X_B dt$
  - Time until next $A$-$B$ reaction

$$\tau \sim \mathcal{E}(1/a)$$

$$a = k_{AB} X_A X_B$$
Chemical Kinetics

- **Macroscopic/Continuous approach**
  - Species concentrations $x_i = \frac{X_i}{V}$
  - Reaction rates $\sim$ reactant concentrations product
  - ODEs for $x_i$...
  - Result of a limiting process

- **Discrete approach**
  - Species concentrations are random variables
  - macroscopic approach $\sim$ expected value
  - with variance $\sim V^2$
  - For small volumes, and small $X_i$, variance blows up

\[ R_1 : X_1 + X_2 \rightarrow 2X_1 \]

Deterministic

\[
\begin{align*}
\frac{dx_1}{dt} &= k x_1 x_2 \\
\frac{dx_2}{dt} &= -\frac{dx_1}{dt}
\end{align*}
\]

SSA or BKL

\[
P(j = l) = \frac{a_l}{a_0}, \quad l = 1, \ldots, M
\]

\[
\delta t \sim \mathcal{E}(1/a_0)
\]

\[
X(t + \delta t) = X(t) + \nu_j
\]
Stochastic simulation: SSA

- For M reactions, time until any reaction
  \[ \tau \sim \mathcal{E}(1/a_0) \quad a_0 = \sum_{j=1}^{M} a_j \]

- Reaction index: point-wise distribution
  \[ p(j = l) = \frac{a_l}{a_0} \]

- The SSA simulates every reaction event!

- SSA: exact but slow

  - One timestep:
    - Sample \( \tau \)
    - Sample the index \( j \)
    - Update the \( X_i \), \( t = t + \tau \)
Stochastic simulation: Acceleration

- **τ leaping**: several reaction events over one time step,

- **Assumption**: reaction propensities $a_i$ remain essentially constant over $\tau$, in spite of several firings

- Over this given $\tau$, the number of reaction firings $K_{pj}^P$ is governed by a Poisson distribution

\[
K_{pj}^P \sim \mathcal{P}(a_j \tau)
\]

\[
X(t + \tau) = X(t) + \sum_{j=1}^{M} K_{pj}^P \nu_j.
\]

Cost $\sim M$ Poisson samplings
Stochastic Simulation Example

- **SSA**: exact but slow

- **τ leaping**: several reaction events over one time step,
  - **Assumption**: reaction propensities \( a_i \) remain essentially constant over \( \tau \), in spite of several firings
  - Over this given \( \tau \), the number of reaction firings \( K_j \) is governed by a Poisson distribution
  - but an approximation was made...

- **τ leaping** (ideal case)
\( \tau \)-leaping Consequences

- **\( \tau \)-leaping**: Can generate negative populations  
  
  - Binomial \( \tau \)-leaping: Approximate the unbounded Poisson distributions with Binomial ones

- **Modified \( \tau \)-leaping**:  
  - Critical reactions, i.e. those likely to drive some populations negative, handled by SSA
  - Other reactions advanced by \( \tau \)-leaping


Chatterjee et al., J. Chem. Phys. 2005

Approximation control: τ leaping

- τ “small enough” or bound the propensity change \( \forall j, \Delta a_j < \epsilon a_j \)

- want to impose, \( \forall j, \mathbb{E}(\Delta a_j) < \epsilon a_j \) to get τ estimate \( \forall j, \text{var}(\Delta a_j) < \epsilon a_j \)

- Only possible in a probabilistic sense

- Taylor expansion and truncation at first order gives bounds on τ

τ leaping: Bounds on propensity changes

- Use the absolute change \( \forall j, \Delta a_j < \epsilon a_0 \)
  - but if \( a_0 \) is large, slow reactions can see their propensities go through huge changes!

- Use the relative changes
  - Solves accuracy but...
    \( \forall j, \Delta a_j < \epsilon a_j \)
  - the computation of \( \mathbb{E}(\Delta a_j) \) and \( \text{var}(\Delta a_j) \) involves the determination of influences of reaction firings over all the propensities

- \( M^2 \) entries! Sparse but still heavy...
  \[ f_{jm}(x) = \sum_{l=1}^{N} \frac{\partial a_j(x)}{\partial x_l} \nu_{lm} \]
  = how much one firing of \( R_i \) changes \( a_j \)
**R-leaping : Accelerate SSA by reaction leaps**

- Leaps are in **prescribed number of reaction firings** \( L \) across all reaction channels

- Time increment \( \tau_L \) is Gamma-distributed \( \tau_L \sim \Gamma(L, 1/a_0(x)) \)

- In this interval we will have \( K_m \) firings of channel \( R_m \)

- with: \[ \sum_{m=1}^{M} K_m = L \]

- In R-leaping, as in SSA, the index \( j \) of every firing obeys a point-wise distribution \[ P(j = l) = \frac{a_l(x)}{a_0(x)} \text{ for } l = 1, \ldots, M. \]
R-leaping : One step

- Define $L$
  \[ \tau_L \sim \Gamma(L, 1/a_0(x)) \]

- Sample the index $j$
  \[ P(j = l) = \frac{a_l(x)}{a_0(x)} \text{ for } l = 1, \ldots, L. \]

- Number of reactions for channel $m$
  \[ K_m = \sum_{l=1}^{L} \delta_{l,m} \]

- Update species and time:
  \[ \mathbf{X}(t + \tau_L) = \mathbf{X}(t) + \sum_{j=1}^{M} K_j \nu_j \]
R-leaping: Accelerate SSA by reaction leaps

- L firings distributed across M reaction channels
  - In τ leaping: $K^p_j$ are independent Poisson variables.
  - In R-leaping, $K_j$ are not independent.

- L as a control parameter
  - System can be brought to a desired state X
  - Time is not a-priori specified
  - New approaches to controlling negative species

- In R-leaping, as in SSA, the index j of every firing obeys a point-wise distribution
R-leaping: How to Sample the M K_j

\[ R_0 \text{ Algorithm} \]

- Pointwise Sampling of L independent reaction indices
  \[ p(j = l) = \frac{a_l}{a_0} \]
  - Simple BUT scales with L - close to the work load of SSA!

The Ro-sampling scales with L and, in particular when compared with \( \tau \)-leaping that scales with M, the method is inefficient for large leap sizes, \( L \gg M \).
The distribution of $K_1$ is a binomial distribution: $\mathcal{B}(L, a_1(x)/a_0(x))$

and for every $m \in \{2, \ldots, M\}$ the conditional distribution of $K_m$

given the event $\{(K_1, \ldots, K_{m-1}) = (k_1, \ldots, k_{m-1})\}$ is

$$K_m \sim \mathcal{B}\left(L - \sum_{i=1}^{m-1} k_i, \frac{a_m(x)}{a_0(x) - \sum_{i=1}^{m-1} a_i(x)}\right).$$

This result is invariant under any permutation of the indices.
R-leaping: How to Sample the M K_j

\( R_0 \) Algorithm

- Pointwise Sampling of L independent reaction indices \( p(j = l) = \frac{a_l}{a_0} \)
- Simple BUT scales with L - close to the work load of SSA!

The \( R_0 \)-sampling scales with L and, in particular when compared with \( \tau \)-leaping that scales with M, the method is inefficient for large leap sizes, \( L \gg M \).

\( R_1 \) Algorithm

- Sampling M correlated binomial variables

\[ B(L, a_j/a_0) \]
- Create correlations with conditional distributions

If \( K_i = k_i, \forall i < m \),

\[ K_m \sim B \left( L - \sum_{i=1}^{m-1} k_i, \frac{a_m}{a_0 - \sum_{i=1}^{m-1} a_i} \right) \]
R-leaping: Sorting for efficiency

• Sampling the $M K_j$ efficiently

If $K_i = k_i$, $\forall i < m$, $K_m \sim B \left( L - \sum_{i=1}^{m-1} k_i, \frac{a_m}{a_0 - \sum_{i=1}^{m-1} a_i} \right)$

• When $\sum_{i=1}^{m-1} k_i = L$, sampling is done!

• Minimize the average $m$ by a permutation of the indices, such that $a_j$ is decreasing

$$a_M > a_3 > a_1 >> \cdots$$

• E.g.

[Diagram showing original loop and permutation of indices, with firing and binomial sampling highlighted]
R-leaping: Efficient Sampling

• Sampling the $\mathbf{M} \, K_j$ efficiently
  • $\mathbf{M}$ can be large (~$10^2$) for bio-chemical systems!
  • Efficient sampling effectively loops over a fraction of $\mathbf{M}$.
  • The larger the system, the bigger the payoff.
  • The more disparate the reaction rates are, the smaller the fraction.
• Price to pay: carry out re-ordering often enough (cheap!)

![Graph showing number of binomial samples per time step for LacY/LacZ activities in E. Coli, $\mathbf{M}=22$.]
Stochastic simulation: R-leaping

- Controlling the leap approximation
- All three methods of $\tau$ leaping are transposable to R-leaping
  - Absolute change of $a_j$
  - Relative change of $a_j$
  - Relative change of $a_j$ but efficiently through the relative changes in populations
Approximation control: τ leaping

• The L>1 fired reactions are changing the propensities
• τ "small enough" or bound the propensity change
  • want to impose, \( \forall j, \Delta a_j < \epsilon a_j \) to get τ estimate
  • Only possible in a probabilistic sense
    \[ \forall j, \mathbb{E}(\Delta a_j) < \epsilon a_j \]
• Taylor expansion and truncation at first order gives bounds on τ
  \[ \forall j, \text{var}(\Delta a_j) < \epsilon a_j \]
Stochastic simulation: R-leaping

- Controlling negative species with $L$ as a leap parameter
  - Bound $L$ by maximum number of firings $L_j$ allowed by reactants
    - Strict control of negative species
  - Bound the expected number of firings for reaction $j$ by $L_j$
    - Introduces negative species

- Introduce trade-off
  - With parameter $\theta$

\[
L < L_j
\]

Too strict!

\[
L \frac{a_j}{a_0} < L_j
\]

Too lax!

\[
L \leq (1 - \theta)L_j + \theta \frac{a_0}{a_j} L_j
\]

Tunable compromise between efficiency and accuracy
Results

• Accuracy and complexity
• Decaying-dimerizing system
• no negative species

\[ \epsilon = 0.05 \]
\[ \epsilon = 0.03 \]
\[ \epsilon = 0.01 \]

Histrogram of \( X_2 \) at \( t = 10 \)

Histogram error vs. CPU time

\( S_1 \overset{c_1}{\rightarrow} 0, \ S_1 + S_1 \overset{c_2}{\rightarrow} S_2, \ S_2 \overset{c_3}{\rightarrow} S_1 + S_1, \ S_2 \overset{c_4}{\rightarrow} S_3 \)

\[ c_1 = 1, \ c_2 = 0.002, \ c_3 = 0.5, \ c_4 = 0.04 \]

\[ X_1(0) = 4150, \ X_2(0) = 39565, \ X_3(0) = 3445 \]
Results

- LacZ/LacY genes expression and enzymatic/transport activities of LacZ/LacY proteins in E. Coli

- Histogram errors vs CPU time

- Efficient sampling offers factor 2 in speed w.r.t. modified $\tau$-leaping!

\[
\begin{align*}
\epsilon &= 0.005 \\
\epsilon &= 0.1 \\
\epsilon &= 0.2
\end{align*}
\]
Summary

• **R-leaping**, an accelerated stochastic algorithm that is complementary to existing τ-leaping algorithms

• Efficient binomial sampling offers computational savings for large systems with disparate rates
  • Efficient sampling exploits size and stiffness of system.
  • Can be transposed to τ-leaping algorithms (!)...  

• **Treatment of negative species** with a tunable compromise efficiency-accuracy
  • An alternative to modified τ-leaping, which essentially recurs to SSA when in trouble
Accelerated SSC - Part II

- Flow Averaging Integrators
  - Recapitulation of the Method
  - An Example: van der Pol's Oscillator

- Flow Averaging for Stochastic Simulations
  - Proposed Method
  - Numerical Examples & Observations

- Cutoff Phenomena
  - A Cutoff Phenomenon in Card Shuffling (work of Diaconis)
  - A Cutoff Phenomenon in FLAVOR-SSA
Consider the following system of stiff ODEs

\[ \dot{u} = G(u) + \frac{1}{\epsilon} F(u) \]

\[ \epsilon \ll 1 \]

with the *legacy* integration scheme:

\[ \bar{u}_{t+\tau} = \Phi_{\tau}^{1/\epsilon}(\bar{u}_t) \]

Explicit Integrators, 1st order \[ \tau \ll \epsilon \]
FLow AVeraging integratOR

FLAVOR: $\bar{u}_{t+h} = \left( \Phi^{0}_{\frac{h}{M} - \tau} \circ \Phi^{1/\epsilon}_{\tau} \right)^{M} (\bar{u}_{t})$

$h \sim O(1) \text{ step}$

$M : \text{number of substeps used to “average” the flow of the dynamical system, where:}$

$$\delta \triangleq \frac{h}{M}$$

$$0 < \tau \ll \epsilon \ll \delta \ll 1$$

Tao, Owhadi, & Marsden, Non-intrusive and structure preserving multiscale integration of stiff ODEs, SDEs and Hamiltonian systems with hidden slow dynamics via flow averaging, Multiscale Model. Simul. 2010
Average the flow of the ODE system by turning on/off the fast dynamics.
van der Pol’s Oscillator:

\[
\begin{align*}
\dot{x} &= \frac{1}{\epsilon} \left( x + y - \frac{1}{3} x^3 \right) \\
\dot{y} &= -\epsilon x
\end{align*}
\]

\[\epsilon = 10^{-3}, \quad \tau = 5 \cdot 10^{-5}, \quad \delta = 10^{-2}\]

\[x(0) = 1, \quad y(0) = 1, \quad T = 5/\epsilon = 5000.\]
Recall SSA:

\[ a_k(X) := c_k \Omega \prod_{j=1}^{N} \left\{ \frac{X_j(X_j - 1) \cdots (X_j - r_j + 1)}{\Omega r_j} \right\} \]

The Algorithm:

\[ \epsilon = \min_k \left\{ \frac{a_k - 1}{a_k} \right\} \quad \text{stiffness parameter (rate disparity)} \]

Here distinguish two scales (can be hierarchical)
**FLow AVeraging integratOR + SSA**

SSA

\[(X_n, t_n) = \left( \Phi^{(1/\varepsilon)} \right) (X_{n-1}, t_{n-1})\]

**FLAVOR-SSA:**

\[(X_{n+1}; t_{n+1}) = \left( \Phi^{(\xi/\varepsilon)} \circ \Phi^{(1/\varepsilon)} \right) (X_{n-1}; t_{n-1})\]

where \(\xi \in [0, 1]\)

\[
\delta = \delta(\xi) \sim \mathcal{E} \left( \frac{1}{\hat{a}_0} \right)
\]

\[
\hat{a}_0(t, \xi) := \frac{\xi}{\varepsilon} \sum_i \tilde{a}_i^{(fast)}(t) + \sum_j a_j^{(slow)}(t)
\]

\[\xi = 0 \implies \text{Largest Speedup}\]

\[\xi = 1 \implies \text{SSA}\]

Is there an optimal \(\xi\)?


**FLow AVeraging integratOR + SSA**

Convergence in Distribution

\[
S_1 \xleftrightarrow{\mathcal{O}(1/\epsilon)} S_2 \xleftrightarrow{\mathcal{O}(1)} S_3
\]

\[
S \xrightarrow{\mathcal{O}(1)} S
\]

Cumulative Distribution

large transition w.r.t the error
FLow AVeraging integrat/OR + SSA

\[ S_1 \overset{O(1/\epsilon)}{\leftrightarrow} S_2 \overset{O(1)}{\leftrightarrow} S_3 \]

\[ \epsilon = 10^{-4} \]

Error in the Cumulative Distribution

Error in the Density Function

\[ \xi = \epsilon \]
How to pick $\xi$?  

(Set $\xi = A^* \epsilon$, $A > 1$)

$S_1 \leftrightarrow S_2$,  \hspace{0.5cm} $\epsilon$ not explicit in this example

$S_1 \leftrightarrow S_3$,  \hspace{0.5cm} Reactions become stiff during the simulation

$2S_2 + S_3 \leftrightarrow 3S_4$.

Error vs. Speedup over SSA

HERE: A-priori selected $\xi$

$\xi \approx \epsilon$
Consider the Riffle Shuffle

1 shuffle results in 2 ascending sequences, 2 shuffles, 4 sequences, etc.

E.g. how many shuffles does it take to randomize a deck of cards?
Cutoff Phenomenon in Card Shuffling

Table 1. Distance to stationarity for repeated shuffles of 52 cards

<table>
<thead>
<tr>
<th>k</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>P^k - \pi</td>
<td></td>
<td></td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.924</td>
<td>0.624</td>
</tr>
</tbody>
</table>

Stationary distribution: \( \pi = \lim_{k \to \infty} P^k_x = \frac{1}{n!} \)

Number of riffle shuffles: \( k \)
can be shown that

\[ \| P^\xi(t) - \pi \|_2 < \exp \left\{ -\lambda t \frac{\xi}{\epsilon} \right\} \]

\( P^\xi(t) \) solution of the FLAVORized Master Equation

\[ \pi = \lim_{t \to \infty} P(t) \]

\( \lambda \) so-called spectral gap, positive, & independent of \( \epsilon \)

Levin, Peres, & Wilmer, Markov Chains & Mixing Times, Amer. Math. Soc. 2009

optimal value for the parameter \( \xi \) can now be chosen, i.e.

\[ \xi > \epsilon \]
Spatially-Dependent SSA

- Diffusion in 1-D
- A species $U$, whose elements are labeled by an index $i$
- Diffusion as a set of reactions of the form:

\[
\begin{align*}
U_i & \xrightarrow{k_{i,i+1}} U_{i+1} \\
U_i & \xrightarrow{k_{i,i-1}} U_{i-1}
\end{align*}
\]

Uniform Cells: $k_{i,j} = \frac{D}{h^2}$
**Space: Finite Volume Method for AMR**

Requirements for Numerical Method:

1) Conservation
2) Convergence to Continuum Equations
3) Convergence in Distribution

### i) uniform (standard FD) scheme

\[
\frac{dU_i^{(s)}}{dt} = \frac{D}{h^2} \left( U_i^{(s)} - U_i^{(s)} \right) + O(h^2).
\]

### ii) multiresolution scheme

\[
\frac{dU_i^{(s)}}{dt} = \frac{4D}{3h^2} \left( 2(U_j^{(s)} + U_k^{(s)}) - U_i^{(s)} \right) + O(h)
\]

\[
\frac{dU_j^{(s)}}{dt} = \frac{dU_k^{(s)}}{dt} = \frac{2D}{3h^2} \left( U_i^{(s)} - 2(U_j^{(s)} + U_k^{(s)}) \right) + O(h)
\]

\[
\frac{dU_i^{(s)}}{dt} = - \left( \frac{dU_j^{(s)}}{dt} + \frac{dU_k^{(s)}}{dt} \right)
\]
Diffusion with Gradient-Based Refinement

$u_0 = \sin(\pi x) \sin(\pi y) \quad x, y \in [0, 1]$
Fisher’s Reaction-Diffusion System in 2-D

\[ X_1 + X_2 \rightarrow 2X_1 \]
• Gray-Scott Reaction-Diffusion System in 2-D
Spatially Dependent Stochastic Simulation Algorithms

**R and tau-Leaping: Coupled Reaction and Diffusion**

\[ u(t + \tau) = u(t) + \sum_{c \in C} \left( \sum_{r \in \mathbb{R}} \nu_r^c P(a_r(u^c), \tau) + \sum_{i \in I} \sum_{d \in \mathbb{D}} \nu_d^{i,c} P\left(\frac{d_i u^c_i}{6dt^2}, \tau\right) \right) \]


D. Rossinelli, B. Bayati, P. Koumoutsakos, Accelerated Algorithms for Stochastic Spatial Simulations, (submitted)

**Hybrid**

Deterministic Diffusion + Stochastic Reactions with Tau-Leaping

\[ u(x, t + \tau) = u(x, t) + M_1(\Delta_d M_2(u(x, t))) + f_s(u(x, t)) \]

◊ **Conversion Method for Continuum to Discrete Models**

\[ M_1 : \mathbb{R}^N_+ \rightarrow \mathbb{N}^N \]

\[ \hat{\Gamma}(x_{i,j,k}) = \hat{\Gamma}_N(x_{i,j,k}) + k_{i,j,k}, \quad \text{for } i = 1 \ldots N_i, j = 1 \ldots N_j, \text{ and } k = 1 \ldots N_k \]

\[ k_{a,b,c} = B \left( L - \sum_{\alpha=1}^{a} \sum_{\beta=1}^{b} \sum_{\gamma=1}^{c-1} k_{\alpha,\beta,\gamma}, \frac{p(x_{a,b,c})}{1 - \sum_{\alpha=1}^{a} \sum_{\beta=1}^{b} \sum_{\gamma=1}^{c-1} p(x_{\alpha,\beta,\gamma})} \right) \]
Stochastic Diffusion on Non-uniform Discretizations: Optimization

STOCHASTIC OPTIMISATION
Task: **minimize an objective function** *(fitness function, loss function)* in continuous domain

\[ f : \mathcal{X} \subseteq \mathbb{R}^n \rightarrow \mathbb{R}, \quad x \mapsto f(x) \]

**Black Box scenario** *(direct search)*

- gradients are not available or not useful
- problem domain specific knowledge is used only within the black box, e.g. within an appropriate encoding
- .......

**Search costs:** number of function evaluations
Problem Statement - Continuous Domain Search/Optimization

- **Goal**
  - fast convergence to the global optimum
  - solution \( x \) with small function value \( f(x) \) with least search cost

- **Typical Examples**
  - shape optimization (e.g. using CFD)
  - model inference
  - parameter calibration

- **Problems**
  - exhaustive search is infeasible
  - naive random search takes too long
  - deterministic search is not successful / takes too long

**Approach:** Stochastic search, Evolutionary Algorithms
Objective Function Properties

We assume $f : \mathcal{X} \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$ to be non-linear, non-separable and to have at least moderate dimensionality, say $n \ll 10$.

Additionally, $f$ can be:

- non-convex
- multimodal
- non-smooth, derivatives not easy to compute
- discontinuous, plateaus
- ill-conditioned
- noisy
- ...

**Goal**: cope with any of these function properties they are related to real-world problems.
Probabilistic Methods

$f : \mathbb{R}^{n} \rightarrow \mathbb{R}$

Parents
\( \{x_k\}_{k=1}^{\mu} \)

Adapt \( \theta \)

Sample \( P(x \mid \theta) \)

Select

Evaluate

Offspring
\( \{x_k\}_{k=1}^{\lambda} \)
Models in Stochastic Optimisation

- **Model to render search more efficient than random sampling**
- **Meta-models to replace expensive evaluations in the selection process**

**Diagram Details**:
- Parents: \( \{ x_k \}_{k=1}^{\mu} \)
- Offspring: \( \{ x_k \}_{k=1}^{\lambda} \)
- **Adapt**: \( \theta \)
- Sample: \( P(\theta) \)
- Evaluate: Select and Evaluate process

**Equation**:
\[
P(\theta) \Rightarrow \theta \Rightarrow \text{Adapt} \Rightarrow P(\theta) \Rightarrow \text{Sample} \Rightarrow \{ x_k \}_{k=1}^{\mu} \Rightarrow \text{Parents} \Rightarrow \text{Select} \Rightarrow \text{Evaluate} \Rightarrow \lambda \Rightarrow \text{Offspring} \]
PROBABILITY SAMPLING

- **Goal**
  - Sampling from the whole distribution
  - Computation of probability weighted moments for QoI

- **Typical Examples**
  - model selection
  - robust predictions
  - model uncertainty quantification

- **Problems**
  - Probabilities are not invertible
  - High dimensional problems
  - Naive sampling takes too long/impossible
  - Laplace approximation is not sufficient

**Approach:** Markov Chain Monte Carlo
BIONSPIRED OPTIMIZERS: BACTERIA CHEMOTAXIS

Model: (Dahlquist, 1976)

1. \( U \): constant

2. \( T \sim P_e(T_0) \)
   \( T_0 \sim \) food gradient and accumulated food

3. \( \Theta \): \( P_g(\Theta_0) \)
   \( \Theta_0 \sim \) duration of previous run and the particular environment

Parameters optimized for various cost functions

Optimal Bacteria

Fig. 11. Paths of two bacteria on $F_1(x, y)$ with and without strategy parameter refinement. Start point $(0, 0)$, $r_{init} = 10^{-10}$. Without strategy parameters refinement ($\rightarrow$), $r_{init} = 10^{-10}$, $n_{pop} = 0$, number of function evaluations: 10806. With strategy parameters refinement ($\rightarrow$), $r_{init} = 10^{-7}$, $n_{pop} = 10$, number of function evaluations: 1172.

Fig. 12. Convergence plot for the two bacteria paths shown in Fig. 11. Without parameter refinement ($\rightarrow$). With parameter refinement ($\cdots$).

Fig. 13. Paths of two bacteria on $F_2 = (x-1)^n + (y-1)^n$. Start point $(0, 0)$, $n_{pateax} = 3$, $A = 10^6$, $B = 4.0$, $r_{init} = 10^{-7}$. Without plateau speedup ($\rightarrow$), number of steps: 2156. With plateau speedup ($\cdots$), number of steps: 365.
Bacteria Optimization for Inverse Airfoil Design
DESIGN ADAPTIVITY: Self-Optimizing Machines

Designing a 'smart wing' for the Mars airplane

BY DAVID KENWRIGHT

Symmetrical Airfoil – maximum thickness

Foam core Flexible skin

Undercambered Airfoil – maximum deflection

Micro-servos

Symmetrical Airfoil – minimum thickness

Load bearing Spars

The wing, which the team will build and test next year, will mimic flying insects in three ways: first, it will continuously sense its environment, process this information electrically, and adjust its inner shape or control surfaces in an appropriate flight mode. Second, these adjustments will be made by

Figure 1: Photograph of a model "smart" wing in a wind tunnel being tested using micro-servo actuators. The wings apply forces to the three main wings and can simulate the flight in 1/500th of a second. The prototype wing can simulate hundreds of different kinds of aerodynamic results.

CREDIT: NASA Tech Reports 1999
SPATIALLY ADAPTIVE CONTROLLERS

Actuation on Each Panel:
Steady Tangential Wall Motion

\[ \mathbf{X}^g = (x_1^g, x_2^g, ..., x_{16}^g) \text{ with } x_i^g \in [-1, +1], \quad i = 1, ..., 16 \]
Drag reduction $\sim 60\%$
Histograms of population values (over each panel)
Critical actuators
3D OPTIMIZATION FROM 2D OPTIMIZED FUNCTIONS

Optimize Rotating Cylinders - 3D to 2D

\[ \Omega(t) = \Omega_1(t) + \Omega_2(t) \sin(2\pi ft + \phi) \]

- Optimization via self-organising maps

- Happiness and Stagnation

- Clustering Genetic Algorithms

- Optimal Differential Evolution
  R. Gämperle, S.D. Müller, P. Koumoutsakos A parameter study for differential evolution, Advances in intelligent systems, fuzzy systems, evolutionary computation, 10 (2002):

- .....
Uncomfortable Statements

We conclude for the sphere function: The amplification factor $F$ should not be smaller than a certain value to avoid that the population converges before arriving at the minimum. On the other hand, the amplification factor $F$ should not be chosen too large because the number of function evaluations increases as $F$ increases. The crossover constant $CR$ should not be too large to avoid that the perturbations get too high and the convergence speed decreases. However, a small $CR$ decreases diversity and might cause the strategy to get stuck. For the same size of the population DE/best/2/bin and DE/rand/1/bin perform similarly. However, for a small population it is more likely to find the minimum using DE/best/2/bin instead of DE/rand/1/bin due to the improved diversity of the trial vectors. The best results were achieved with DE/best/2/bin.
Optimization Challenges

- **Efficiency**: (compressed representation of a varying environment)
- **Adaptivity**: (small changes, emergence)
- **Programmability**: (same inputs & state rules, same outputs)
Optimization Challenges

Efficiency
(compressed representation of a varying environment)

Adaptivity
(small changes, emergence)

Machine Learning
= Adaptively Create Efficient Models of the Environment
Optimization as Sampling

\[ \tilde{x} \sim P(x, \hat{\theta}) \]

\[ \hat{\theta} = \arg \max_{\theta} \int J(x) P(x, \theta) \, dx \]
Evolution Strategies

(1 + 1) - ES  maximize \( f(x) \)

From parent \( m \) sample one offspring  \( x = m + Q(x; m) \)

Compute the ratio  \( a = \frac{f(x)}{f(m)} \)

If  \( a > 1 \) select  \( m \leftarrow x \)
(1 + 1)-ES  

maximize $f(x)$

From parent $m$ sample one offspring $x = m + Q(x; m)$

Compute the ratio $a = \frac{f(x)}{f(m)}$

If $a > 1$ select $m \leftarrow x$

Probability of success

- $1/2$
- $1/5$

“too small”
Why Step-Size Control?

\[ f(x) = \sum_{i=1}^{n} x_i^2 \]

in \([−2.2, 0.8]^n\)

for \(n = 10\)
New search points are sampled normally distributed

\[ x_i \sim m + \sigma N_i(0, C) \quad \text{for } i = 1, \ldots, \lambda \]

as perturbations of \( m \), where \( x_i, m \in \mathbb{R}^n, \sigma \in \mathbb{R}_+, C \in \mathbb{R}^{n \times n} \)

where

- the mean vector \( m \in \mathbb{R}^n \) represents the favorite solution
- the so-called step-size \( \sigma \in \mathbb{R}_+ \) controls the step length
- the covariance matrix \( C \in \mathbb{R}^{n \times n} \) determines the shape of the distribution ellipsoid

here, all new points are sampled with the same parameters

KEY ISSUE: how to update \( m, C, \) and \( \sigma \).
Recombination:: Update $m$::

\[ x_i = m + \sigma \mathcal{N}(0, C) = m + \sigma z_i \]

\[ f(x_1: \lambda) \leq \cdots \leq f(x_\lambda: \lambda) \]

\[ m \leftarrow \sum_{i=1}^{\mu} w_i x_{i: \lambda} = m + \sigma \sum_{i=1}^{\mu} w_i z_{i: \lambda} = m + \sigma \langle z \rangle_w \]

\[ \sum_{i=1}^{\mu} w_i = 1 \quad w_1 \geq \cdots \geq w_\mu \geq 1 \]
Mutate $C$ to increase the probability of successful steps to appear again

$$C \leftarrow (1 - LR_{cov})C + LR_{cov}\langle z \rangle_w \langle z \rangle_w^T$$

The rank-one update has been found independently in several domains\textsuperscript{6 7 8 9}

\begin{itemize}
    \item \textsuperscript{6} Kjellstrom&Taxen 1981. Stochastic Optimization in System Design, IEEE TCS
    \item \textsuperscript{7} Hansen&Ostermeier 1996. Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation, ICEC
    \item \textsuperscript{8} Ljung 1999. System Identification: Theory for the User
    \item \textsuperscript{9} Haario et al 2001. An adaptive Metropolis algorithm, JSTOR
\end{itemize}
We combine live cell microscopy, single particle tracking and trajectory segmentation with system identification:

\[
\langle z \rangle_w \langle z \rangle_w^T = -\langle z \rangle_w (-\langle z \rangle_w)^T
\]

\[
p_c \leftarrow (1 - LR_c)p_c + LR_c \langle z \rangle_w
\]
"Cumulation" is a widely used technique and also known as

- exponential smoothing in time series, forecasting
- exponentially weighted *moving average*
- *iterate averaging* in stochastic approximation
- *momentum* in the back-propagation algorithm for ANNs
- ...

"Cumulation" conducts a *low-pass* filtering, but there is more to it...
Recap::

\[ x_i = m + \sigma z_i \]

\[ m \leftarrow m + \sigma \langle z \rangle_w \]

\[ p_c \leftarrow (1 - c_c)p_c + \sqrt{1 - (1 - c_c)^2} \frac{1}{\sum_{i=1}^{\mu} w_i^2} \langle z \rangle_w \]

\[ C \leftarrow (1 - c_{cov})C + c_{cov} p_c p_c^T \]

\[ z_i \sim \mathcal{N}(0, C) \]

\[ \langle z \rangle_w = \sum_{i=1}^{\mu} w_i z_i ; \lambda \]
There is more information out there

\[ Z = \sum_{i=1}^{\mu} w_i z_i ; \lambda z_i^T ; \lambda \]

\[ C \leftarrow (1 - c_{cov}) C + c_{cov} Z \]
### Rank-$\mu$ update

\[ x_i = m + \sigma z_i \]

\[ m \leftarrow m + \sigma \langle z \rangle_w \]

\[ p_c \leftarrow (1 - c_c) p_c + \sqrt{1 - (1 - c_c)^2} \sqrt{\frac{1}{\sum_{i=1}^{\mu} w_i^2} \langle z \rangle_w} \]

\[ C \leftarrow (1 - c_{cov}) C + c_{cov} \frac{1}{\mu_{cov}} p_c p_c^T + c_{cov} \left(1 - \frac{1}{\mu_{cov}}\right) Z \]

\[ z_i \sim \mathcal{N}(0, C) \]

\[ \langle z \rangle_w = \sum_{i=1}^{\mu} w_i z_{i;\lambda} \]

\[ \mu_{cov} = \sqrt{\frac{1}{\sum_{i=1}^{\mu} w_i^2}} \]

\[ Z = \sum_{i=1}^{\mu} w_i z_{i;\lambda} z_{i;\lambda}^T \]
The rank-μ update

- increases the possible learning rate in large populations
  roughly from $2/n^2$ to $\mu_w/n^2$
- can reduce the number of necessary generations roughly from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$
  given $\mu_w \propto \lambda \propto n$

Therefore the rank-μ update is the primary mechanism whenever a large population size is used
  say $\lambda \geq 3n + 10$

The rank-one update

- uses the evolution path and reduces the number of necessary function evaluations
  to learn straight ridges from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$

Rank-one update and rank-μ update can be combined

---

The CMA-ES

**Input:** \( m \in \mathbb{R}^n, \sigma \in \mathbb{R}_+, \lambda \)

**Initialize:** \( C = I, \) and \( p_c = 0, p_\sigma = 0 \)

**Set:** \( c_c \approx 4/n, c_\sigma \approx 4/n, c_1 \approx 2/n^2, c_\mu \approx \mu_w/n^2, c_1 + c_\mu \leq 1, d_\sigma \approx 1 + \sqrt{\frac{\mu_w}{n}} \)

and \( w_{i=1}...\lambda \) such that \( \mu_w = \frac{1}{\sum_{i=1}^{\mu} w_i^2} \approx 0.3\lambda \)

**While not terminate**

\[ x_i = m + \sigma y_i, \quad y_i \sim \mathcal{N}_i(0, C), \quad \text{for } i = 1, \ldots, \lambda \]

Not covered on this slide: termination, restarts, useful output, boundaries and encoding
Invariance

- Empirical performance results
  - from benchmark functions
  - solved real world problems

  are only useful if they do generalize to other problems

- Invariance is a strong non-empirical statement about generalization

  generalizing (identical) performance from a single function to a whole class of functions

  invariance is important for the evaluation of search algorithms
• Invariance Under Monotonically Increasing Function (rank only update)

• Translational Invariance

• Rotational Invariance
Invariance Under Monotonically Increasing Functions

Rank-based algorithms

Update of all parameters uses only the ranks

\[ f(x_1^{\lambda}) \leq f(x_2^{\lambda}) \ldots \leq f(x_\lambda^{\lambda}) \]

\[ g(f(x_1^{\lambda})) \leq g(f(x_2^{\lambda})) \ldots \leq g(f(x_\lambda^{\lambda})) \quad \forall g \]

\( g \) is strictly monotonically increasing
\( g \) preserves ranks

Credit: Whitley 1989. The GENITOR algorithm and selection pressure: Why rank-based allocation of reproductive trials is best, ICGA
BAYESIAN

UNCERTAINTY QUANTIFICATION
MODEL : Molecular Dynamics

\[
\frac{dx_p}{dt} = u_p
\]

\[
m \frac{du_p}{dt} = \sum_{i=1}^{N} F_p \equiv -\frac{\partial U}{\partial x_p}
\]
Water Flow in Carbon Nanotubes

EXPERIMENTS
Sources of Uncertainty in Water-Graphite Systems

\[ \phi_{\text{LJ}}(r_{ij}) = 4\varepsilon_{\text{LJ}} \left( \frac{\sigma_{\text{LJ}}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{\text{LJ}}}{r_{ij}} \right)^{6} \]

K. Osborne III (2009)
Water Droplets on Graphene: Wetting Depends on MD Potentials

**Water conduction through the hydrophobic channel of a carbon nanotube**

G. Hummer*, J. G. Rasaiah† & J. P. Howorka†

**Molecular Dynamics Simulation of Contact Angles of Water Droplets in Carbon Nanotubes**

Thomas Weiser,*† Jan H. Walther,* Richard L. Jacks,† Torun Halsevings,² Flavio Rocca,² and Pietro Parrinello*

Contact angle (°)

-20 -15 -10 -5 0

Binding energy of a water monomer (kJ/mol)

-160 -140 -120 -100 -80 -60 -40 -20 0

**Scattering of water from graphite: simulations and experiments**

Nikola Marković,*†, Patrik U. Anderson, Mats B. Någård, Jan B.C. Petersson¹

**Hydrogen bond structure of liquid water confined in nanotubes**

M.C. Gorilillo, J. Martí*
“Theories have to be judged in terms of their probabilities in light of the evidence.”

“Life’s most important questions are, for the most part, nothing but probability problems.”

“What we know is not much. What we do not know is immense.”
Bayes: Models and Data

\[ P(A|B)P(B) = P(B|A)P(A) \]

\[
\begin{align*}
P(A|B) & = \frac{P(B|A)P(A)}{P(B)} \\
\text{posterior} & \quad \text{likelihood} \\
\text{prior} & \quad P(B) \\
\end{align*}
\]

\[ A \rightarrow \text{Hypothesis/Model} \]
\[ B \rightarrow \text{DATA} \]
Bayesian UQ: Calibration and Model Selection

Experimental Data: \( D \)

Use observations to select the model classes and estimate their parameter values such that the model predictions best fit the data.

**PARAMETER ESTIMATION**

\[
f(\theta_i|D, MD_i) = \frac{f(D|\theta_i, MD_i) \pi(\theta_i|MD_i)}{f(D|MD_i)}
\]

- Experiments
- Physical limitations
- Past studies
- Expert elicitation

**MODEL CLASS SELECTION**

\[
Pr(MD_i|D) = \frac{f(D|MD_i)Pr(MD_i)}{f(D)}
\]

**Evidence of Model Class**

\[
f(D|MD_i) = \int f(D|\theta_i, MD_i) \pi(\theta_i|MD_i) d\theta_i
\]

tradeoff between data-fitting and information gain from the data.
Evolution Strategies vs Markov Chain Monte Carlo

<table>
<thead>
<tr>
<th>(1 + 1)-ES</th>
<th>maximize ( f(x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>From parent ( m ) sample one offspring ( x = m + Q(x; m) )</td>
<td></td>
</tr>
<tr>
<td>Compute the ratio ( a = \frac{f(x)}{f(m)} )</td>
<td></td>
</tr>
<tr>
<td>If ( a &gt; 1 ) select ( m \leftarrow x )</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MH—MCMC</th>
<th>sample ( P(x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>From parent ( m ) sample one offspring ( x = m + Q(x; m) )</td>
<td></td>
</tr>
<tr>
<td>Compute the ratio ( a = \frac{P(x)}{P(m)} )</td>
<td></td>
</tr>
<tr>
<td>If ( a &gt; 1 ) select ( m \leftarrow x )</td>
<td></td>
</tr>
<tr>
<td>ELSE with ( \text{prob}(a) ) select ( m \leftarrow x )</td>
<td></td>
</tr>
</tbody>
</table>
Sampling vs Finding Optima of a Distribution

DRAM

CMA-ES
p = annealing parameter

$p = 1$?

YES

p = 0

STOP

NO

$p++$

Parents \( \{ \theta_k \}_{k=1}^{\lambda} \)

rank according to likelihood

Offsprings \( \{ \theta_k \}_{k=1}^{\mu} \)

update p

Rejuvenate

\( \text{MH} (\mathcal{N}(\theta_1, \Sigma_\theta), \text{length}(i)) \)

Resample \( \{ \theta_l \}_{\text{leader}}^{\text{length}} \)

Select: Poisson sampling with replacement
Rejuvenate

MH (N (θ₁, Σᵦ), length(i))

Original scheme: uneven chain length proportional to the number of times a sample is resampled

New scheme: use uniform chain length (add burn-in to each chain for better convergence)
X-TMCMC

Improve efficiency of rejuvenate using MH

Reduce number of function evaluation by
1. using Langevin-based proposals
2. construct surrogate model with Kriging (aka Gaussian processes)
Approximate Bayesian Computation (ABC)

stochastic model

\[ M(\theta) \sim \pi_M \]

parameters

\[ p(x|\theta, M) = \pi_M(\theta) \]

augmented parameter space

\( (\theta, x) \) model outcome (r.v.)

joint posterior

\[ p((\theta, x)|D, M) \propto p(\theta|M)p(x|\theta, M)p(D|(\theta, x), M) \]

prior model likelihood

approximate likelihood

\[ P(x \in \mathcal{N}_\epsilon(D)|x, M) \]

with statistics

\[ \mathcal{N}_\epsilon(D) = \{ x : \rho(\eta(x), \eta(D)) \leq \epsilon \} \]

discrepancy tolerance

approximate marginal posterior

\[ p(\theta|M, D) \propto p(\theta|M)P(x \in \mathcal{N}_\epsilon(D)|\theta, M) \]
Why Approximate Bayesian?

- Measurement Error \( e^d \sim N(\mu^d, \Sigma^d) \)
- Computational Error \( e^f \sim N(\mu^f, \Sigma^f) \)
- Model Error \( e^m \sim N(\mu^m, \Sigma^m) \)

Few molecules, rare events, out of equilibrium simulations

Valid?

Hierarchical Coarse Graining: from atomistic to mesoscale, full access to molecular ensemble data

**Approximate Bayesian Computing for MD**

MD as a Stochastic model

\[ M(\theta) \sim p(x|\theta) \]

(parameters)

Approximate likelihood

\[ p(\rho(\eta(D), \eta(x)) \leq \delta|x, MD) \]

(discrepancy)

\[ (\theta, x) \]

(model outcome (r.v.))

**Which Discrepancy to use?**

Kullback-Leibler divergence

\[ \rho(x, y) = D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx \]

(relative entropy)

K-L quantifies the overlap between two molecular ensembles*

Target Data is now the entire molecular ensemble

---

ABC-SubSim

\( \text{len} = \text{Markov chain length inside MH} \)

\( D = \text{experimental data} \)

\( x_i = M(\theta_i) \) (model evaluation)

\[ \pi(\theta) \]

\[ \rho(x_i, D) \]

Select: \((100 / \text{len})\)% best samples, each chain has a length \text{len}
ABC-SubSim for Helium

Parameters:

\[ V_{LJ}(r; \sigma, \epsilon) = 4\epsilon \left[ \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^{6} \right] \]

Data:

Boltzmann factor distribution:

\[ f_B = \left\langle \exp \left(-\frac{H}{T k_B} \right) \right\rangle \]

\[ \rho(\eta(x), \eta(y)) = \sqrt{\left( \frac{\mu_x - \mu_y}{\mu_x} \right)^2 + \left( \frac{\sigma_x - \sigma_y}{\sigma_x} \right)^2} \]

Gaussian

\[ \rho(\eta(x), \eta(y)) = \left( \sum_{k=1}^{4} \left( \frac{q_k(x) - q_k(y)}{q_k(x)} \right)^2 \right)^{1/2} \]

Quantiles

Kullback-Leibler

\[ \rho(\eta(x), \eta(y)) = D_{KL}(P||Q) \]
**EXAMPLE: MD SIMULATIONS FOR WATER**

- **Lennard-Jones Potential**
  \[
  \phi(r) = 4\epsilon_{LJ} \left( \frac{\sigma_{LJ}}{r} \right)^{12} - \left( \frac{\sigma_{LJ}}{r} \right)^{6}
  \]

- **MD parameter calibration - $\theta$**
  - 0-0 Lennard-Jones ($\epsilon_{LJ}, \sigma_{LJ}$)
  - 0-0, 0-H Coulomb charges ($q$)
Calibration and robust prediction

Bayesian Calibration of TIP5P-E water MD parameters

Robust Prediction of TIP5P-E water model radial distribution function
water contact angle

parameters: $\varepsilon$, $r_{\text{cut}}$

HPC challenges for UQ in nanoscale flows:
large, variable, computational cost per calibration or propagation sample

Resources: 1200 cores, 32GPUs

Friction coefficient and slip length of water inside CNTs

500 samples - 2 days

water transport in CNTs

200 samples - 7 days

Enhanced Flow in Carbon Nanotubes

Measurement of the Rate of Water Translocation through Carbon Nanotubes
Xingcai Qin, et al., NanoLetters, 11, 2173, 2011

Fast Mass Transport Through Sub-2 Nanometer Carbon Nanotubes


Measurement of the Rate of Water Translocation through Carbon Nanotubes
Xingcai Qin, et al., NanoLetters, 11, 2173, 2011
DATA HETEROGENEITY

“What we observe is not a Nature itself, but a Nature exposed to our method of questioning” - Werner Heisenberg

Different:
Equipment, Labs, Conditions, Patients, Doses etc.

Structure for predictions based on data

Argon viscosity
How to achieve multiple goals

As stated above, the goal was to develop the simplest potential function, which reproduces well the density anomaly of liquid water, while simultaneously yielding good thermodynamic and structural properties near 25°C and 1 atm. Additional studies of three- and four-site models, in-

- Multiobjective Optimization
  - Pareto front
  - Weighted sum of objective functions

Mahoney & Jorgensen, J. Phys. Chem. 2000
Hierarchical Bayesian Models

Graph based models of a joint probability distribution

A way to describe **conditional (in)dependencies in data**

\[
p(\tilde{\theta}_1, \tilde{\theta}_2|\tilde{\psi}) = p(\tilde{\theta}_1|\tilde{\psi}) \cdot p(\tilde{\theta}_2|\tilde{\psi})
\]

- **Hierarchical Bayesian (HB) Model**
- **Classical Bayesian (CB) Model** + Hierarchical Prior
Hierarchical Bayesian Framework

CB: \[ p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} \]

HB: \[ p(\theta|D) = \int p(\theta|\psi)p(\psi|D) \, d\psi \]

Sampling \( \theta \Rightarrow \) multiple evidence evaluation

Very High Computational cost

HPC

\[ p(D|\psi) = \prod_{i=1}^{N} p(D_i|\psi) \]

Evidence of individual data set

\[ = \prod_{i=1}^{N} \int p(D_i|\theta_i, \psi)p(\theta_i|\psi) \, d\theta_i \]

Derived from probability axioms and graph models

RECALL: Evidence is a tradeoff between data-fitting and information gain from the data.
Example 1: Inconsistent Posteriors - Argon

Parameters:
Lennard-Jones Potentials $\varepsilon_{LJ}$, $\sigma_{LJ}$

Krypton viscosity
Kestin et al. 1984
Effect of stochastic model on predictions

Classical Bayesian: One set of parameters fit all

Hierarchical Bayesian: Varying parameters

Parameters:
Lennard-Jones Potentials $\varepsilon_{\text{LJ}}, \sigma_{\text{LJ}}$

Krypton viscosity
Kestin et al. 1984

$\text{MCB log-Evidence : -77.02}$

$\text{M}_{\text{HB}} \text{ log-Evidence : -20.64}$
Example 2: **Calibrating MD simulation for water**

- Heterogeneous Data Sets
  - Diffusion coefficient
  - Density
  - Radial Distribution Function (RDF)

*Data sources: Holz et al. 2000; Jones & Harris 1992; Soper 2013*
Bayesian Inference using Individual Data Set

Calibrate for each data set individually...

* Calibrate $q$ and $\epsilon_{LJ}$ only, $\sigma_{LJ}$ highly correlated with $\epsilon_{LJ}$
Hierarchical Bayesian model vs Classical Bayesian model

Hierarchical Bayesian

Classical Bayesian

Model Selection

| Model | P(Model|D) |
|-------|-------|
| HB    | 0.98  |
| CB    | 0.02  |

(3) Range from diff. water models

* q ~ [0.4, 0.65]

(1) Mahoney & Jorgensen 2000
(2) Rick et al. 2004
(3) http://www1.lsbu.ac.uk/water/water_models.html

NEXT: Discrete Element Methods, Dissipative Particle Dynamics
HPC Challenges

- **Simulations**
  - variable computational cost
  - can generate large volume of data
- **Algorithms**
  - include multiple levels of parallelism
  - introduce load imbalance
- **Hardware**
  - variety of computing resources, highly heterogeneous
- **Programming methodology**
Π4U: HPC Framework for Bayesian UQ

- Platform agnostic task-based parallelism
- Multi-level parallelism
- Transparent load balancing
- Extensible/ built upon the TORC tasking library

https://github.com/cselab/pi4u
SUMMARY

- **first-principle knowledge** describes universal causal relationships. These relationships (laws) are deterministic and they have both the predictive and explanatory value, as they involve just a few concepts (variables).

- **empirical knowledge** describes statistical dependencies (derived from observed data) that have predictive properties.

  Explanation of such empirical models is outside the scope of predictive modeling.

- **beliefs** are statistical dependencies (derived from observed data) that have explanatory (or descriptive) value but little predictive value. In philosophical terms, beliefs can be regarded as a form of psychological induction.
CLASSICAL OR EMPIRICAL KNOWLEDGE?

“I have called this principle, by which each slight variation, if useful, is preserved, by the term Natural Selection.”

Charles Darwin - *The Origin of Species*
Thank You!