Adaptive Computing and multifidelity learning

Juliane Mueller
Computational Science Center

With contributions from Marc Day, Hilary Egan, Ryan King, Kevin Griffin, Jibo Sanyal
Scale-up of complex systems and associated risks

**Scale-up:** Extending systems and processes that were developed in the laboratory to function in the real world

Device and process scale-up comes with significant technical challenges and risk

Typical challenges:

- Data-driven models perform best when interpolating, *extrapolation is inherently uncertain*, and therefore risky
- Increasing ranges of scale (spatial, temporal) often lead to new/enriched physics
- High-fidelity physics-based models may capture new physics, but are typically too expensive for design/optimization work
- Operational regimes of existing experiments are limited, and new experiments are expensive
Goal: reduce scale-up challenges by integrating multi-fidelity modeling and optimal compute resource use

**Adaptive Computing**
Orchestration of a multi-fidelity model hierarchy and/or experiment campaign to arrive at the best goal-based solution with well-characterized uncertainty given finite resources

**Multi-fidelity models and real-time experiment synergy**
Connect models with experiments to drive experiment design and data acquisition needs

**Goal-oriented solutions**
Optimize the use of finite resources to achieve a specific science goal

**Optimization and uncertainty management**
Control of extrapolation uncertainty through targeted active learning
Most applications feature an assortment of models of widely varying fidelities, developed for different purposes:

- **Experiment**: “Truth”, but limited operational regime
- **High-fidelity simulations**: Physics-based (PDE/ODE), costly
- **Lower fidelity levels**: reduced physics, coarser meshes, less costly
- **Data-driven surrogates**: AI/ML, PINNs, Gaussian Processes (GPs), really cheap

**Fig**: Exploiting information from multiple fidelity levels can increase surrogate accuracy
Black-box expensive optimization: high fidelity

\[
\min f(x) \\
\text{s.t. } g_i(x) \leq 0, \ i = 1, \ldots, I \\
x \in \Omega
\]
Surrogate models approximate the expensive function

\[ f(x) = m_f(x) + e_f(x) \quad \text{Surrogate of the objective function} \]
\[ g_i(x) = m_{g_i}(x) + e_{g_i}(x) \quad \text{Surrogate of the constraint function} \]

Different types of surrogate models exist:
- Radial basis functions
- Gaussian Process models
- Multivariate adaptive regression splines
  - Polynomials
  - …..

Throughout the optimization, we let the surrogate model guide the search for improved solutions
Surrogate model based optimization loop

1. Initial experimental design
2. Evaluate expensive objective function
   - Stop?
     - Yes: Return best solution found
     - No: Fit/Update surrogate model
6. Select new evaluation point
   - Guided by surrogate model

RBF, GP, polynomial..

Stop when compute budget used up

Needs adaptation for problems with constraints, integers, failed evaluations, noise, multiple conflicting objective functions
Surrogate model guided sampling

Initial experimental design
Compute surrogate surface
Select and evaluate new point
Update surrogate surface
Select and evaluate new point
Update surrogate surface
Exploiting multiple fidelity levels

Correcting the low-fidelity model:

• Multiplicative: \( \hat{y}_{hf}(x) = \rho(x) \cdot y_{lf}(x) \)

• Additive: \( \hat{y}_{hf}(x) = y_{lf}(x) + \delta(x) \)

• Hybrid:
  - \( \hat{y}_{hf}(x) = \rho(x) \cdot y_{lf}(x) + \delta(x) \) (\( \rho \) const.)
  - \( \hat{y}_{hf}(x) = w(x) \cdot \rho(x) \cdot y_{lf}(x) + (1 - w(x)) \cdot (y_{lf}(x) + \delta(x)), w \in [0,1] \)

How do we make use of multiple fidelity levels during active learning?
Exploiting multi-fidelity information

Build a surrogate model for the low(er) fidelity function

- Allow more samples than for high-fidelity function
- Use this surrogate to decide where to focus the search in the high-fidelity function
- *Low fidelity model does not have to be accurate*

Build a surrogate model for the high-fidelity function

- Fewer samples are affordable
- Surrogate is less accurate (built on less data)
- Surrogate can be used to make (final) sample decisions
Gaussian Process: Using multiple fidelity information in one model

Red = high-fidelity evaluations
Black = Lower fidelity evaluations

Combining high and lower fidelity information can lead to better approximation surface (compare to true contours)

Python package surrogate modeling toolbox (SMT)
GP-based sampling methods

Expected improvement balances local and global search

\[ E(I)(x) = s(x) (\nu \Phi(\nu) + \phi(\nu)) , \quad \nu = \frac{f^{\text{best}} - m_{\text{GP}}(x)}{s(x)} \]

Probability of improvement – mostly local search

\[ P(I)(x) = \Phi \left( \frac{f^{\text{best}} - m_{\text{GP}}(x)}{s(x)} \right) \]

\[ LCB(x) = m_{\text{GP}}(x) - \kappa s(x) \]

- \( m_{\text{GP}} \) is the GP prediction
- \( s \) is the standard deviation of the GP predictions
- \( f^{\text{best}} \) is the best function value found so far
- \( \phi, \Phi \) are the normal density and cumulative distribution

\( \kappa \) – adjustable parameter
GP based sampling in 1d

Maximization problem

Expected improvement

Upper confidence bound

Probability of improvement

ParBayesianOptimization in Action (Round 1)
Maximize the expected improvement to select a new point

Expected improvement surface is multimodal and can become flat – making it difficult to find the global maximum...

...requires development of other sampling strategies, guided by low-fidelity model
Sampling with candidate points

Add random perturbations to (select) variables of the best point(s) found so far

- Maximize a merit function that trades off predicted function value and distance to already evaluated points
  - Low function value -> local search
  - Large distance -> global search
- Select $N$ new points for potential evaluation
Multi-fidelity sampling: when to ignore the low-fidelity model

- Make use of as much information as is available
  - Surrogate of high-fidelity model
  - Low-fidelity (cheap) information – what if this one is very inaccurate/uncorrelated?
  - Surrogate of the difference as a selection constraint

1. Define auxiliary function $a(x)$ using the surrogate model predictions
2. Optimize $a(x)$ to find $x_{new}$
3. If $-\delta \leq d(x_{new}) \leq \delta$ probe with low fidelity model first, otherwise ignore and evaluate high-fidelity model
Goal: reduce scale-up challenges by integrating multi-fidelity modeling and optimal compute resource use

Adaptive Computing
Orchestration of a multi-fidelity model hierarchy and/or experiment campaign to arrive at the best goal-based solution with well-characterized uncertainty given finite resources

Goal-oriented solutions
Optimize the use of finite resources to achieve a specific science goal

Multi-fidelity models and real-time experiment synergy
Connect models with experiments to drive experiment design and data acquisition needs

Optimization and uncertainty management
Control of extrapolation uncertainty through targeted active learning
Key capability: diverse compute resources

Edge

Cloud

Datacenter

Resource Management

Resource manager: how much and what kind of resources do I have available?

Solve stochastic discrete optimization problem

Optimal computing strategy driven by specific output quantity of interest
Compute resource optimization problem

- What resources are available when?
- Formulate as optimization problems with stochasticity
- Implement solutions as constraints for multifidelity sampling
- Eventually must exploit asynchronous parallel computations
- Enumerate the user-defined simulation types (fidelity levels)
- Possible hardware configurations (# of CPUs, GPUs)
- Corresponding calculation duration
- Measurement noise estimate (aleatoric uncertainty)
Compute resource allocation with stochasticity

1. Select sample points (e.g., maximize EI with multi-start; candidate point approach)
   \[ \{x_1, x_2, \ldots, x_N \} \in \arg \max_{x \in \mathcal{X}} a(x) \]

2. Get total resource limit \( T \) and per level resource limit \( T_j \) and allocate compute resources

\[
\max_{b_{ji} \in \{0,1\}^k} \sum_{j=1}^{J} r_j(x_i) \times b_{ji} \]

\[
\sum_{j=1}^{J} \sum_{i=1}^{N} b_{ji} \times t_j(x_i, \zeta_{ji}) \leq T
\]

\[
\sum_{i=1}^{N} b_{ji} \times t_j(x_i, \zeta_{ji}) \leq T_j \quad \forall j
\]

- \( r_j \) the benefit of evaluating \( x_i \) at fidelity level \( j \), e.g.,
- \( r_j \) captures accuracy or other QoI

\[
b_{ji} = \begin{cases} 
1 & \text{if } x_i \text{ evaluated with fidelity level } j \\
0 & \text{else}
\end{cases} \quad J \times N \text{ binaries to optimize}
\]

- Total resource restriction
- \( t_j \) resource consumption at level \( j \)
- Resource restriction on fidelity level \( j \)
AC framework

Check resource availability
- Data center
- Edge
- Cloud
- Experiment

Resource management

Application specific code

AC common software stack

Surrogate models

Sampling methods

Statistics, analysis, visualization

- Optimization
- Sensitivity analysis
- Fidelity lvl 1
- Fidelity lvl 2
- ... Fidelity lvl M
ARIES – Advanced Research on Integrated Energy Systems

Various RE generation leading to grid needs

Data store and Control Center

Generation

AC Challenge: How do we build a transferable framework to assess the available flexibility?

Community scale real and emulated buildings, equipment, electric vehicles, etc.

Demand

Real House

Virtual House

Neighborhood
How much demand flexibility is available to support the grid

Goal: Understand, evaluate, and predict the demand flexibility a device, building, or a community can provide:
• Optimizing or shaping the energy demand to support grid conditions
• Providing a way to tap into extra demand flexibility in extreme situations

Challenge: Computing the anticipated energy demand and the opportunities to exploit optimal control is computationally expensive
• Data from real-world observations and complex building simulation models to evaluate the available flexibility in the community thru:
  • Control of HVAC, lighting, water heating, etc. individually or in combination
  • Aggregating the demand profile to the distribution network and grid
• Use surrogate models and understand where they are too inaccurate and additional real-world data and/or complex full-scale simulation data must be collected
• Computational cost in deployment settings is substantially more acute than in the lab setting
Sources and characteristics of uncertainty

Limitations of surrogate models: surrogate models can only be created for a subset of model parameter distributions, number of dimensions, etc.
- Building energy models have 1000s of parameters with 100s that are important
- Will not represent all possible building types/configurations/equipment classes
- Sampling of only a subset of weather conditions is possible
- Limited transferability across geographic locations

Trust in the surrogate requires uncertainty quantification
Objectives: controls-oriented questions around energy and heating/cooling
  • Peak shaving, load shifting, preheating/precooling
  • Assess impact of higher efficiency equipment (e.g., LED lighting)
    • Some component changes come in from community questions, renewable energy mandates, prioritize based on carbon footprint
    • ResStock /ComStock libraries for components
  • Widespread EV charging
Modeling ecosystem

- *EnergyPlus* is high fidelity model (1-30 minutes per building for a year simulation w/ 1 min timestep, heavy on I/O)
- *OCHRE*, struggles in some locations, residential only (E+ does residential and commercial)
- *3R2C* lumped capacitance ROM (4-5 hrs to train) for model predictive control w/ 15 min to 1 hour timestep (works for residential and commercial)
- Models are *deterministic and highly nonlinear* (issues with using ROM outside of intended regime)

Parameterizations

- Need to preserve explicit component representation
- Building thermodynamics should be independent of component selection
Challenges
- Scaling simulation and control to larger communities
- Discrete optimization
- ROM accuracy is a function of component choice

Test case going forward
- Pena station with ~150 buildings
  - 50 defined so far
  - Multifamily buildings are challenging, EnergyPlus extra expensive
- EnergyPlus and 3R2C ROM already built
- UrbanOpt usually used to generate model inputs, but have a little more info for Pena station
Example application: virtual engineering of biofuels

- Process lignocellulose-rich biomass into biofuel
- 3 step chemical processes
  - Pretreatment: fast simulation
  - Enzymatic hydrolysis: surrogate or CFD calculation
  - Bioreaction: surrogate or CFD calculation
AC specifics for virtual engineering

- **Objective:** maximum reactor-averaged oxygen uptake rate
- **Inputs:** order 10 chemical and processing design parameters

- **Fidelity Lvl 1:** HF simulation (pretreatment, enzymatic hydrolysis, bioreactor)
  - 32 CPU-cores @ 57 hours
- **Fidelity Lvl 2:** HF pretreatment, LF Lignocellulose model, and HF bioreactor
  - {72 CPU-cores @ 4 hours, or 32 CPU-cores @ 9 hours}
- Could add simulation type that varies the time to steady state/grid resolution

<table>
<thead>
<tr>
<th>Parameter name</th>
<th>Double or Int*</th>
<th>Default</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraction of solids that is xylan</td>
<td>Double</td>
<td>0.263</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Fraction of solids that is glucan</td>
<td>Double</td>
<td>0.4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Porous fraction of the biomass particles</td>
<td>Double</td>
<td>0.8</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Initial concentration of acid</td>
<td>Double</td>
<td>0.0001</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Steam temperature (C)</td>
<td>Double</td>
<td>150</td>
<td>3.8</td>
<td>250.3</td>
</tr>
<tr>
<td>Fraction of insoluble solids</td>
<td>Double</td>
<td>0.745</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Enzymatic load</td>
<td>Double</td>
<td>30</td>
<td>0</td>
<td>1000</td>
</tr>
<tr>
<td>FIS_0 target</td>
<td>Double</td>
<td>0.05</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Gas velocity (m/s)</td>
<td>Double</td>
<td>0.08</td>
<td>0.01</td>
<td>0.1</td>
</tr>
<tr>
<td>Column height (m)</td>
<td>Double</td>
<td>40</td>
<td>10</td>
<td>50</td>
</tr>
<tr>
<td>Column diameter (m)</td>
<td>Double</td>
<td>5</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Bubble diameter (m)</td>
<td>Double</td>
<td>0.006</td>
<td>0.003</td>
<td>0.008</td>
</tr>
<tr>
<td>OUR_max (mol/m^3/hr)</td>
<td>Double</td>
<td>88.71</td>
<td>5</td>
<td>100</td>
</tr>
</tbody>
</table>

*Integers require adjustment of the MF methods*
Preliminary tests on virtual engineering app

- 8 parameters, 9 random samples from LHS, 10 iterations
- Run for 5 minutes on 1 core, low fidelity models only
- Max Oxygen Uptake Rate = 0.06723323
- For some parameter settings, we obtained NaNs
  - The low fidelity models used may not be valid across the entire parameter space (hidden constraints – we know how to deal with these)

(No multi-fidelity business yet)
Adaptive computing: optimizing the use of computational resources to target deficiencies and challenges related to scale-up

- Efficient use of computational resources and reduction of associated costs
- Reliable scale-up of power systems
- Reliable scale-up of laboratory experiments

Adaptive computing: Novel generalized framework

enables

enables

enables