Initial Design of Experiments to Determine Feasibility to Meet Parameter Estimation Objectives *Prior to* Data Collection

Mark Daichendt, Pooja Bhalode, Lorenz Biegler
Carnegie Mellon University
Pittsburgh, PA 15217

Ben Weinstein, Robert Johnson, Steve Hodson
Chemical Systems Modeling & Simulation
Procter & Gamble
West Chester, OH 45069

Spring EWO Meeting

14-15 March 2017

Motivation

- The Modeling & Simulation Group frequently is first engaged to perform a parameter estimation analysis ***only after*** all the experimental data has been collected ("inherited data")
    - Range of experimental conditions may have been insufficient
        - Limited to the ***Formulation Design Space***
    - Some of the responses may be noisier than others
        - ***Low Signal-to-Noise Ratio***
    - Potential responses may not have been measured in order to reduce costs
        - ***Information content*** may not have been considered
    - Preferable for Business to ***first consult*** with the Modeling & Simulation Group to determine:
        - How the ***limits of the Experimental Design Space*** impact
            - The ***precision*** of the parameter estimates
            - The ***uncertainty*** in the prediction
        - How ***different responses*** contribute to the ***business objective***

Problem Statement

- **Given** a dynamic model of a reaction network generated from the CheK Library in gPROMS
- **Find** an optimal set of experimental design points
  - Initial concentrations, temperature
  - To be ***run in parallel due to the long campaign length***
- That **optimize** a selected criterion of the Information Matrix
  - D/E/A-optimal Designs (maximization)
  - Covariance (minimize absolute value)
- **While** not adversely impacting the other criteria (parallel design)
  - Avoid designing experiments at cross purposes
    - E.g. improving one covariant element while making others worse
  - Calculate all criteria simultaneously, ***one optimized, others constrained***
    - Underlying Pareto-optimal aspect to the approach
- In a ***"Robust" manner*** and ***provide "Recourse"***
  - For future sampling times and selected responses as data is collected

Accomplishments & Obstacles

- Initialize multiple (>100) experiments one temporal finite element (FE) at a time while determining optimal finite element length
  - Uses current parameter estimate and experimental conditions. NEED TO:
    - Perform for worst-case initial conditions for Design of Experiments
    - Perform for worst-case parameter estimates for Robust Designs
    - **Solution:** Find combination that produces the largest absolute value of any component's concentration rate of change (smallest FE size)
- ***Generate*** the information matrix (IM) for the next-best design point (to be extended to multiple design points) ***and optimize***
  - Integrating the composition profiles along with their sensitivities
  - ☹ Optimizer does not move away from the initial guess
    - Multiple local optima
    - "Egg carton" surface

Leverage Analogous Approach

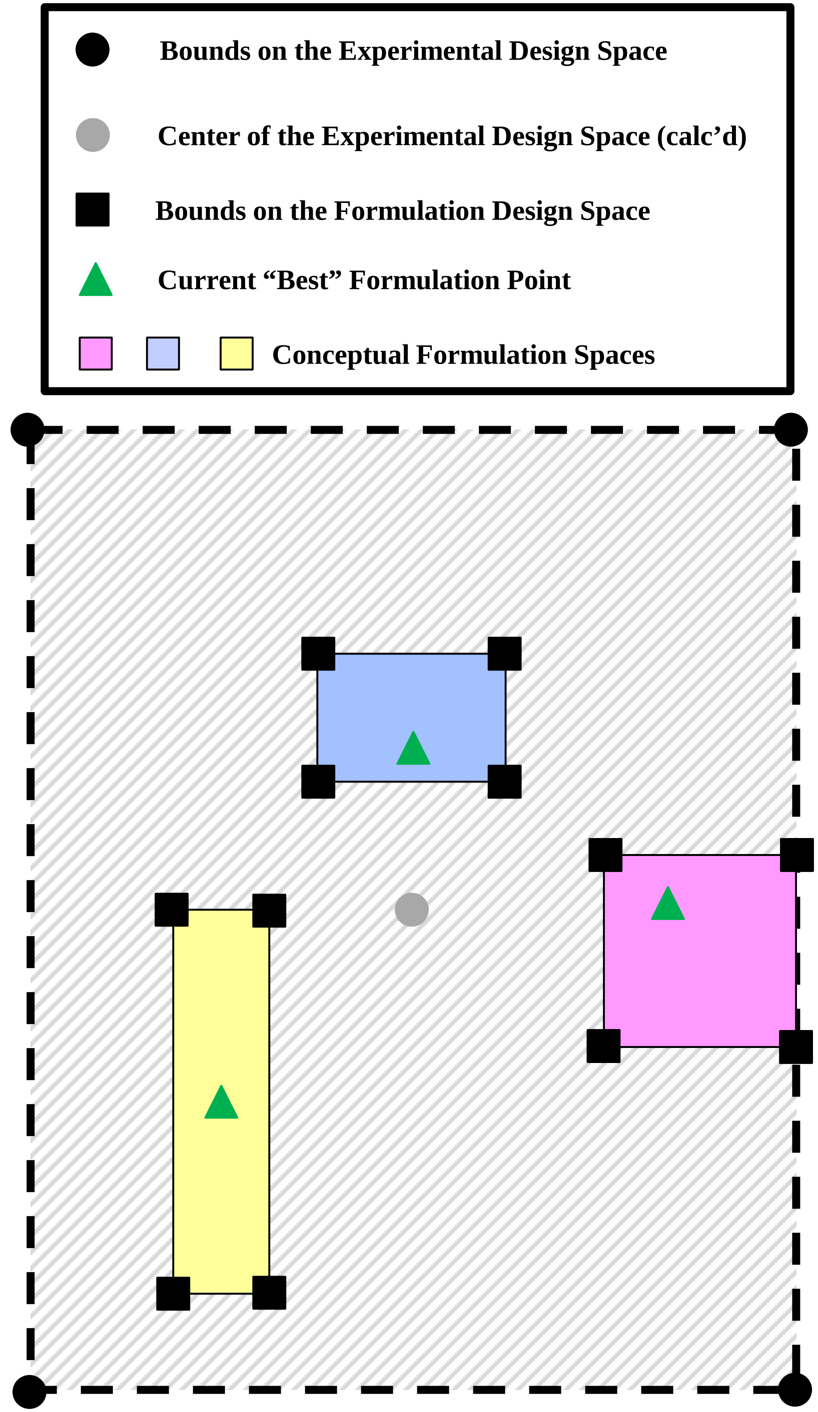
- Highly nonlinear high-fidelity adsorption models optimized using a low-order model (LOM)
  - Choose starting initial condition estimates based on factorial Design of Experiments approaches, denoted by the set U
    - Initial guesses for parameters, $\boldsymbol{\theta}^{k=0}$
    - Initial guesses for data variance-covariance matrix $\boldsymbol{\sigma}^{k=0}$
      - Instrument precision based on manufacturer's specifications
      - In-house determination (e.g. following IUPAC guidelines)
  - Demonstrated capability to initialize multiple experiments simultaneously
    - High-fidelity model results
  - Determine which combination of experimental points $u \in \hat{U} \sqsubseteq U$ leads to a good $IM^{\Sigma} = \sum_{u \in \hat{U} \sqsubseteq U} IM_u$, where IM_u is the IM for experiment u
  - Fit points to a quadratic response surface model (LOM "~")
    - Predicts how the elements of the $\widetilde{IM}^{\Sigma}$ change with initial conditions $\widetilde{IM}_u$
  - Optimize the initial conditions using the LOM, denoted by the set U^*
  - Update $IM^{\Sigma *} = \sum_{u \in \hat{U}^* \sqsubseteq U^*} IM_u^*$ using the high-fidelity model

Design Space and Measured Responses

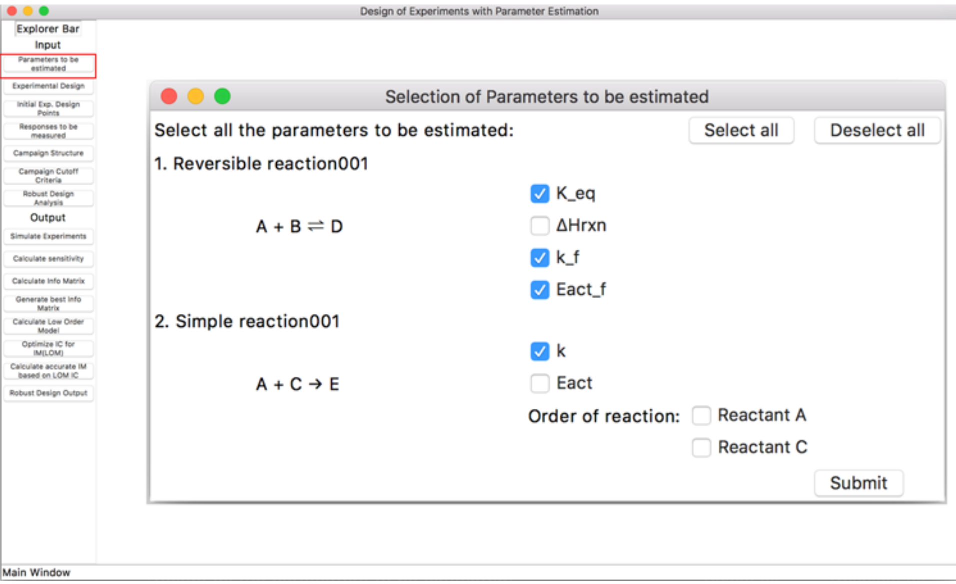
- Business/Technical partnership to develop a good experimental campaign
  - Size of the design space is important:
    - Effects the precision of the parameter estimates
    - Experiments can be run more quickly at more aggressive conditions
      - Early initial data $\rightarrow$ early parameter estimates
        - Recourse for future sampling times and measurements
        - Design new experiments for remaining horizon
        - Early adjustments to formulation
  - Which responses are measured brings different information content
    - Understand from the onset how including or excluding particular responses affects the information matrix
      - Impacts the rank of the matrix
      - Impacts the precision of the parameter estimate

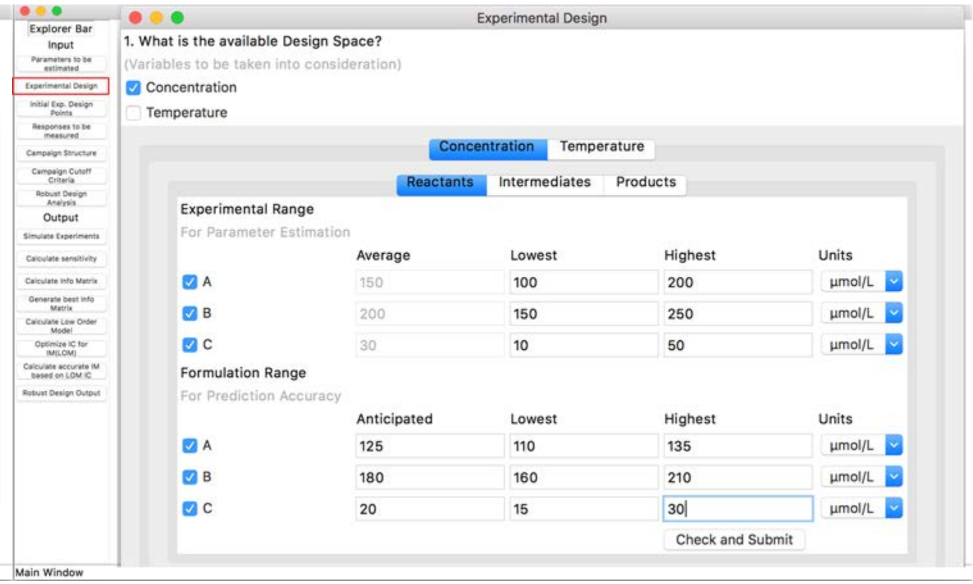
Design Spaces

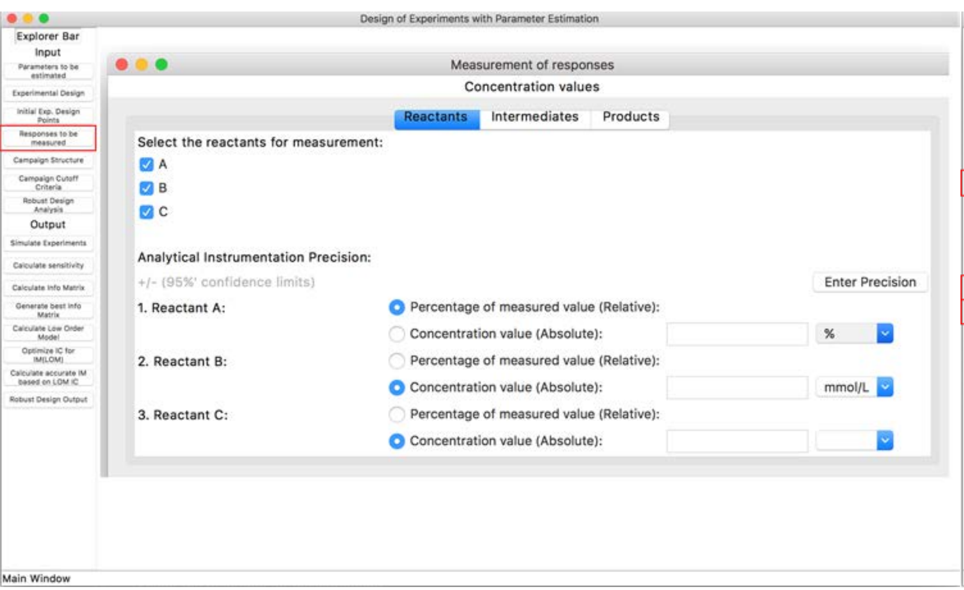
- Distinguish between
  - ***Experimental Design Space***
  - ***Formulation Design Space***
- Experimental Design Space depends on
  - Equipment limitations
  - Short-term thermal degradation, etc.
- Desire largest possible Experimental Design Space
  - (Bard) Optimal Design Points usually have at least one variable at a bound
- Variance-Covariance matrices of both the ***parameters*** and the ***data*** impact the prediction uncertainty
- Sample data within the Formulation Design Space to
  - ***Assess prediction accuracy***
    - Hold out data
  - ***Fine tune the parameter estimates***

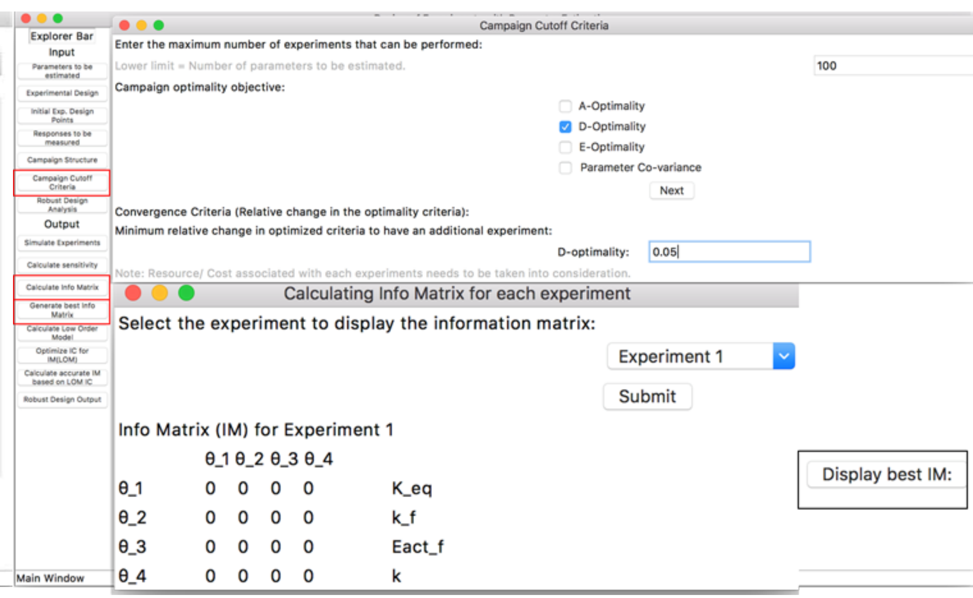


Selected Interface Menus Guide Consult









Novelty/Significance of Work

- Large-scale optimization capability of IPOPT
- Port to open-source PyOMO / Python platform
- Dynamic optimization using Orthogonal Collocation on Finite Elements
- Simultaneous direct evaluation of sensitivities
- Simultaneous calculation (to optimize or constrain) the various DoE objectives of the information matrix
- Parallel experimental campaign structure due to long duration
- Design for multiresponse data with “missing observations”
  - Account for both cost and information content of a measured response
- Incorporate “Robust” experimental design methodologies
  - Reduce dependence on initial parameter guess
  - Best design for worst-case parameters values
- Provide “Recourse” by
  - Performing parameter estimates throughout the campaign
  - Updating sampling times for specific responses
  - Designing additional experiments for remaining campaign horizon

Impact for Industrial Applications

- Create the incentive for the Business Area to ***first consult*** with the Modeling & Simulation Group to determine, ***before any experiments are conducted and measurements are taken***:
  - Whether or not the contemplated ***experimental campaign*** can achieve the desired ***business objective***
  - If not, utilize the tool to determine the impact of:
    - Extending the limits on the ***Experimental Design Space***
    - Including ***responses*** that may be ***less noisy*** and/or provide significantly ***better information content***
- ***Reduce cost*** of running an "ad-hoc" (versus "model-based" or "systematic") campaign with the associated cost and time that
  - May generate ***data with limited information content***
  - ***Cannot be analyzed*** to determine if the business objective can be met
- Paradigm shift to start with a parallel Design of Experiments before gathering any data by leveraging the reaction network model