Parallel Design of Experiments for Kinetic Parameter Estimation to Support Product Stability Predictions

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

Ben Weinstein, Robert Johnson, Steve Hodson
Procter & Gamble
West Chester, OH 45069

Fall EWO Meeting
30 September – 1 October 2015

Carnegie Mellon

Motivation & Objective

- Consumer products may be subject to prolonged storage and transportation times
  - Ensuring Product Quality and Stability is a critical and time-consuming activity
  - Shelf-lives typically measured in years
- Need to understand impact of formula design on product performance and stability
  - Undesirable chemical reactions may lead to product degradation
- Develop a suite of modeling tools to support the development of new consumer products:
  - Simulate product chemistry
  - Design efficient experimental campaigns
  - Discriminate among alternative models and estimate model parameters
  - Understand the uncertainty associated with model predictions

Framework

- CheK is a model library in gPROMS, developed and maintained by P&G
  - CheK allows for any combination of 8 kinetic models to form a complex reaction network
  - The reactor model is an isothermal batch reactor
  - For Product Stability applications, optimization for Parameter Estimation has been challenging in gPROMS
- The goal is to extend the current CheK capability by interfacing it with enhanced optimization tools and methods:
  - Integration using orthogonal collocation on finite elements
  - Simultaneous NLP solution using IPOPT
  - Open source modeling package PyOMO within Python platform
- The deliverable will be a flexible and robust application that can be maintained by P&G

Historical Usage Compared to Extension

- gPROMS has been successfully used for Process Engineering problems at P&G for approximately 16 years
  - More recent use has extended to Product Stability (PS) applications
- Initiated work on an actual Product Stability test problem
  - Validated gPROMS simulation results with optimization-based simulation
  - While results are similar, the PS problem is not well-scaled for optimization purposes using gPROMS internal units:
    - Concentrations are in μmol/L vs mol/L
    - Time units are in days vs sec
    - Some constraints were merely satisfied due to small variable numeric values (constraint feasibility tolerance 4E-10)
    - Small concentration values were also causing function errors for exponentiation in rate expressions

Automatic Scaling & Reformulation

- It is <u>desirable</u> to have variables and equations scaled to O(1)
  - $0.01 \leq$ scaled variable ≤ 100
  - Median of scaled equation coefficients = 1
- It would be <u>undesirable</u> to maintain two separate versions of CheK with different internal units to provide scaling
  - Implemented an automatic scaling that is indifferent to units of measure (Process Engineering vs Product Stability)
  - Scaling factors adjusted based on the input data
    - The input data, of course, must be dimensionally consistent
- Solve the problem in terms of ln_concentration and ln_reaction_rate *as variables* of optimization
  - $reaction\_rate = k \prod_{i \in Reactant} concentration_i^{\nu_i}$ becomes
  - $\ln\_reaction\_rate = \ln\_k + \sum_{i \in Reactant} \nu_i \ln\_concentration_i$
    - Reformulated expression is linear in the case of constant ν

Initialization & Performance

- After scaling and reformulation, could not get the “full-space” (all time periods simultaneously) model to solve
- Used an optimization to initialize the concentration profile for fixed rate of change
  - Initialization ~500 iterations
  - Simultaneous solution ~2000 iterations
- Developed a procedural initialization based on 1st-degree collocation approximation between collocation points (Euler)
  - No optimization now needed for initialization
  - “Full-space” solution ~13 iterations
- Implemented Finite Element by Finite Element initialization (integration) to handle more complex kinetic networks and multiple experiments
  - ~6 iterations per Finite Element (x365 total Finite Elements)
  - Initialized “full-space” model then solves in 6 iterations

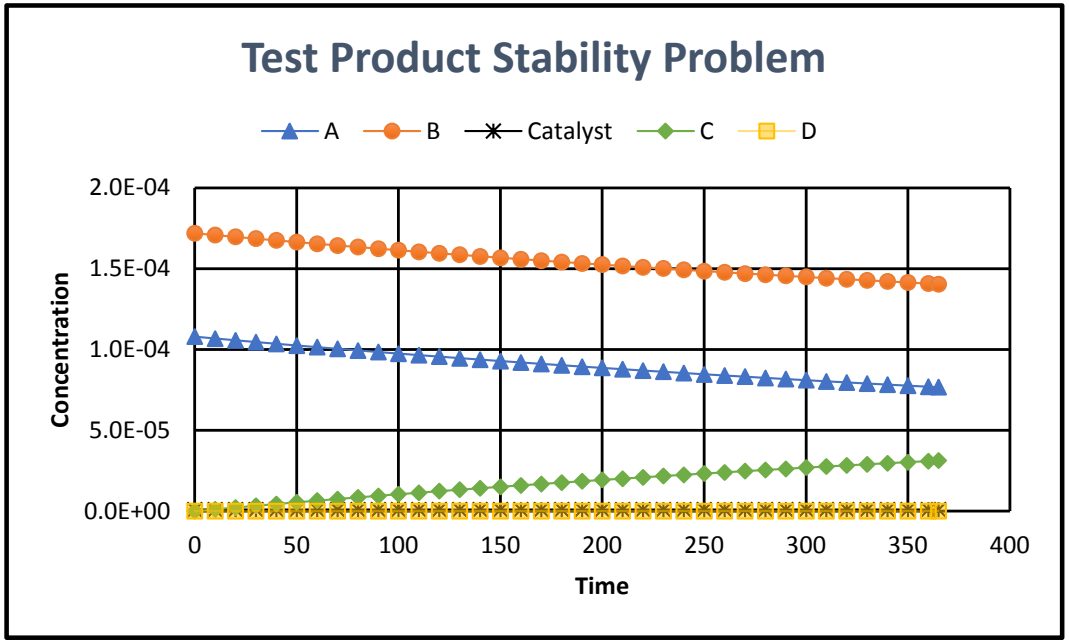
Results of Test Problem

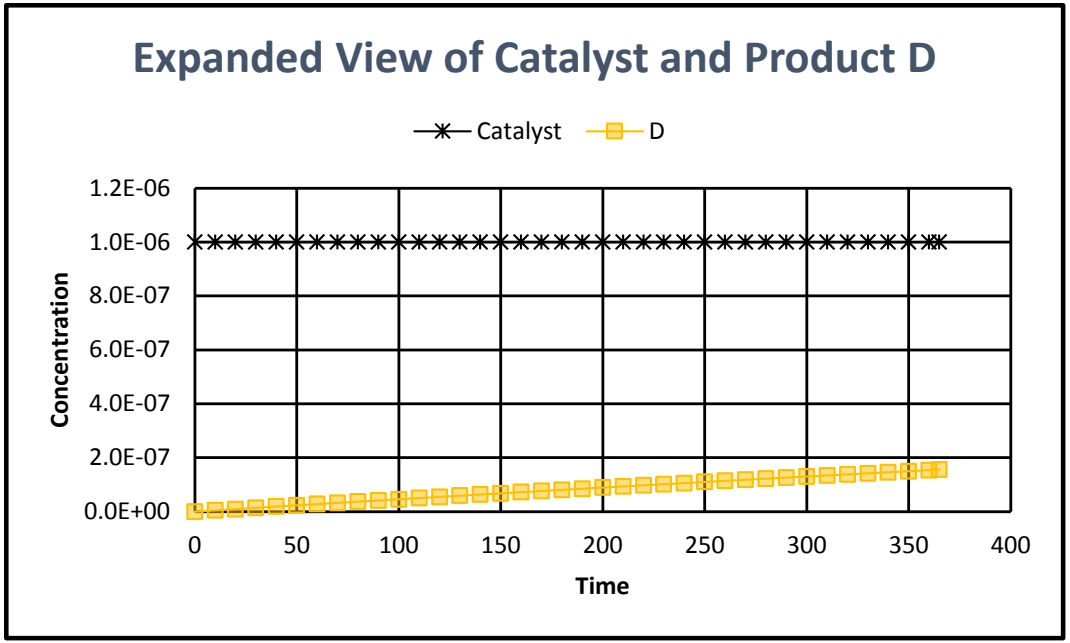
$$A + B \rightleftharpoons C$$

$$A + Catalyst \rightarrow D$$

"Catalyst" is a ubiquitous component in the solution matrix.

It is not an added component to accelerate the reaction.





Parallel Design of Experiments

- The goal of a Product Stability experimental campaign is to predict the amount of product degradation after a long time period (e.g. two years) as the product makes its way through the supply chain
  - Predicted concentration of degradation product(s)
  - Confidence limits on prediction(s)
- The reactions are slow and the experimental temperatures have a limited range, based on anticipated supply chain temperatures.
  - Vials of various concentrations (initial conditions) are stored in one of a limited-number of temperature-controlled rooms
- Depending on the initial conditions and temperatures, a single experiment my run from three to six months (desired) or more
- The duration of the experimental campaign necessitates parallel Design of Experiments (DoE)

Parallel Design Objectives

- A parallel design may incorporate combinations of design objectives
- Currently, Cholesky and Schur matrix decompositions are available to perform the following design objectives:
  - E-optimal: Maximize the minimum eigenvalue of the Information Matrix
    - In parallel DoE, sequentially generate design points to maximize the value of each of the eigenvalues, starting with the smallest
  - D-optimal: Maximize the determinant of the Information Matrix (minimizes the volume of the joint confidence region)
  - A-optimal: Maximize the trace of the Information Matrix
  - Anti-Correlation: Minimize the covariance between selected parameter pairs
- Optimized design objective values can be constrained in subsequent designs, so that one obtains the best value for a new objective that does not worsen previous objective values

Robust Optimal Design

- Typically, there is not an abundance of kinetic data for the reactions in the Consumer Products area
  - Little or no experimental data is initially available
  - Wide bounds on permissible parameter values around initial estimates
- Desire to make efficient usage of Analytical Chemistry resources and limited experimental space
- Robust optimization (optimization under uncertainty) identifies the worst combination of parameter values for the desired experimental design objective for each experiment
  - Parameter values only constrained by lower and upper bounds
- Robust Optimal Design thus seeks to find the best experimental design point for each experiment based on the most conservative set of parameter values for the particular design objective

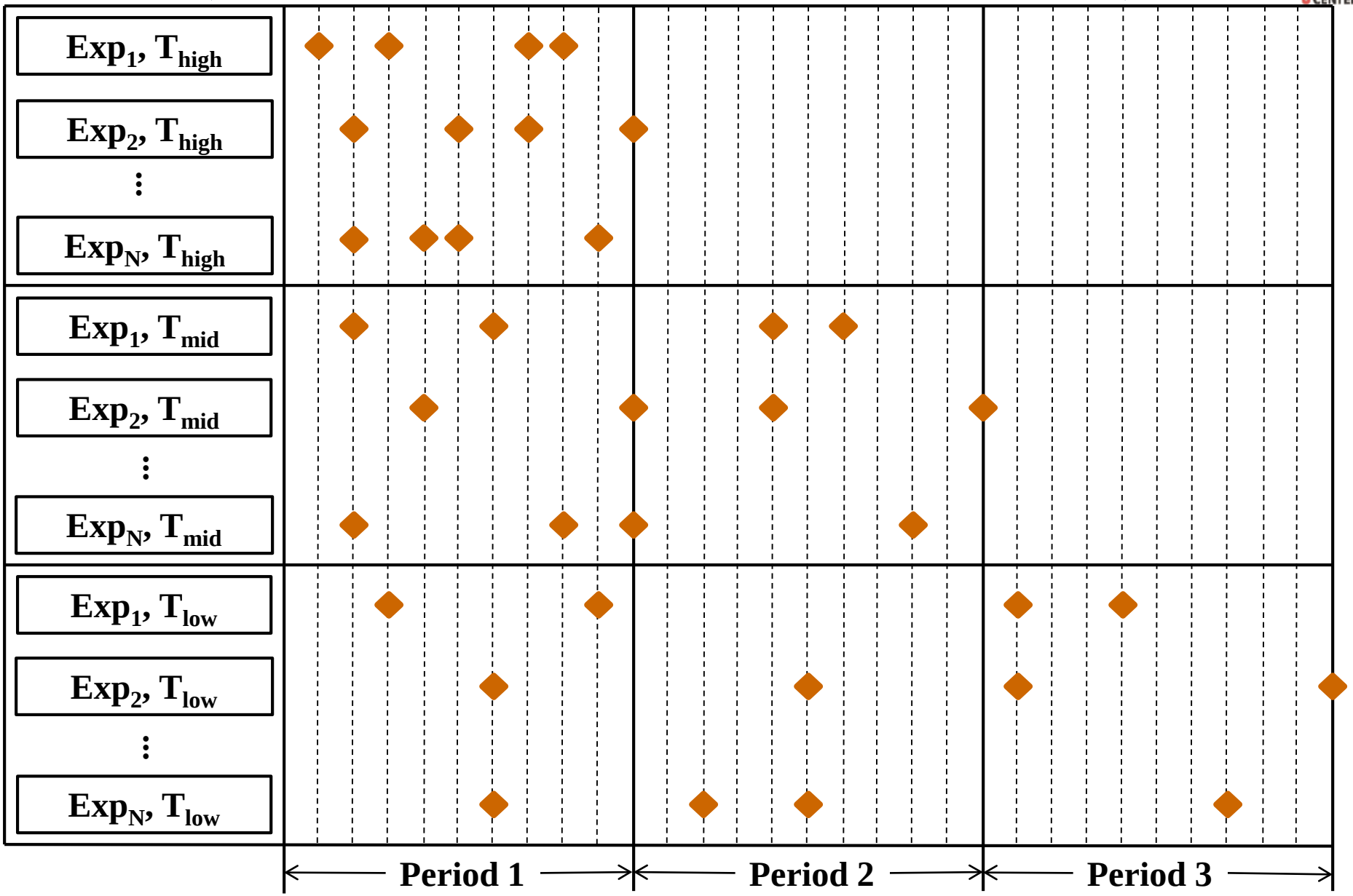
Opportunity for Parallel-Sequential Design

- Obervation 1: For a single reaction, the profiles of identical experiments (initial concentrations) at different temperatures trace identical “scaled profiles”
  - The profile at a higher temperature is a “condensed” (in time) version of the profile at a lower temperature
- Observation 2: Optimal sampling time profile will be shorter at higher temperatures, so higher temperature experiments can be designed and run sequentially
  - Within the overall experimental campaign duration
  - Using increasingly tighter parameter estimates based on data accumulated in the earlier phase of the experimental campaign

Recourse in Robust Parallel Design

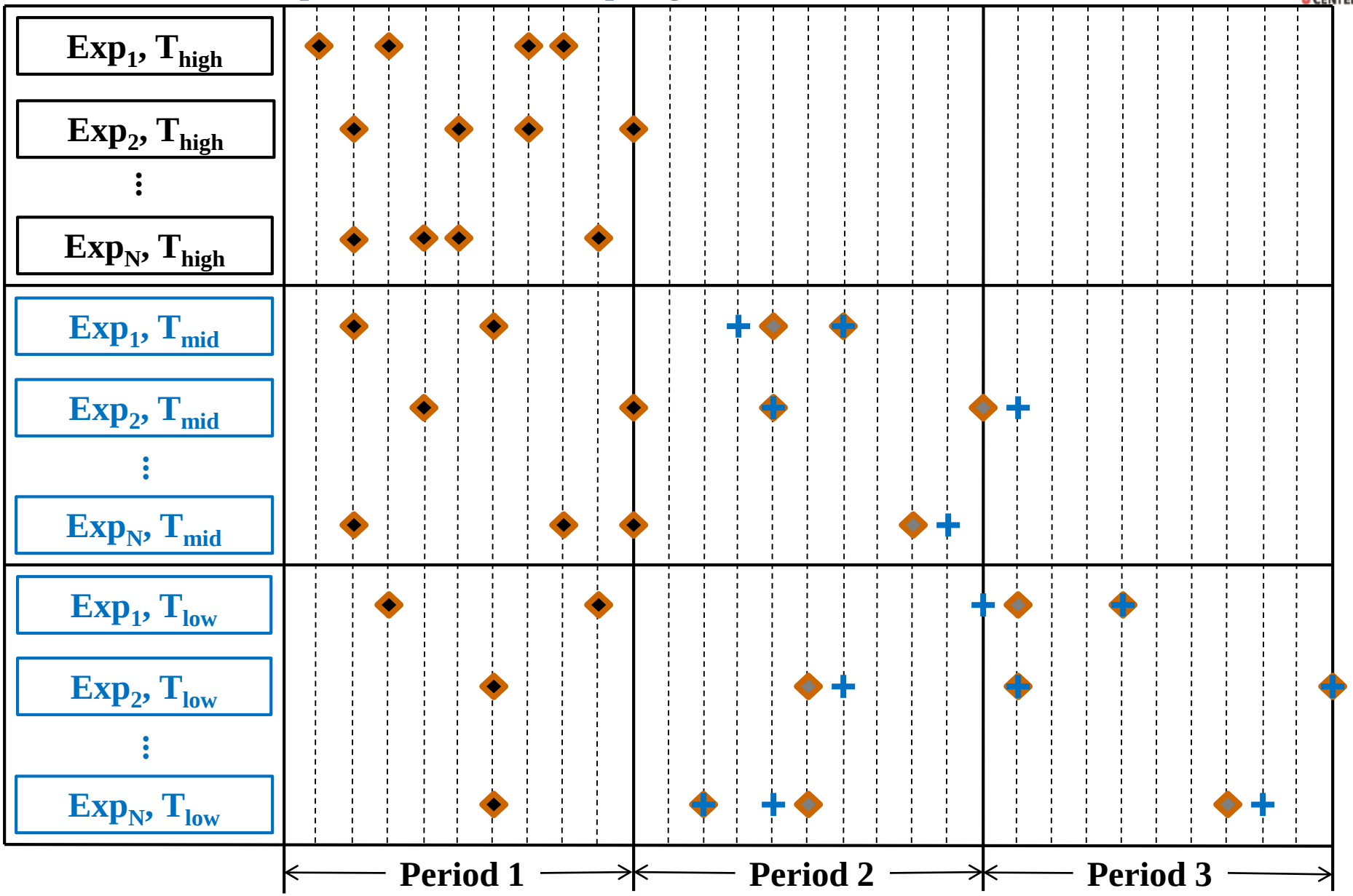
- Perform intermediate parameter regression based on data as it becomes available on a period by period basis
  - Multiperiod Design of Experiments with sequentially improved parameter estimates
  - Recourse 1: Adjust *optimal future sampling times* for experiments already in progress
  - Recourse 2: Design *additional (new) experiments* at intermediate and higher temperatures
    - Limited by the overall campaign duration
    - Sequentially improved designs due to incorporation of new information from accumulated measurements data
- For illustrative purposes in the following "cartoon" slides, assume that an experiment proceeds
  - Two times as fast at T_{high} compared to T_{mid}
  - Three times as fast at T_{high} compared to T_{low}

N Experiments at Different Temperatures
(For Illustration Assume Same Initial Conditions for each *n*)



◆ Initial Sampling Times

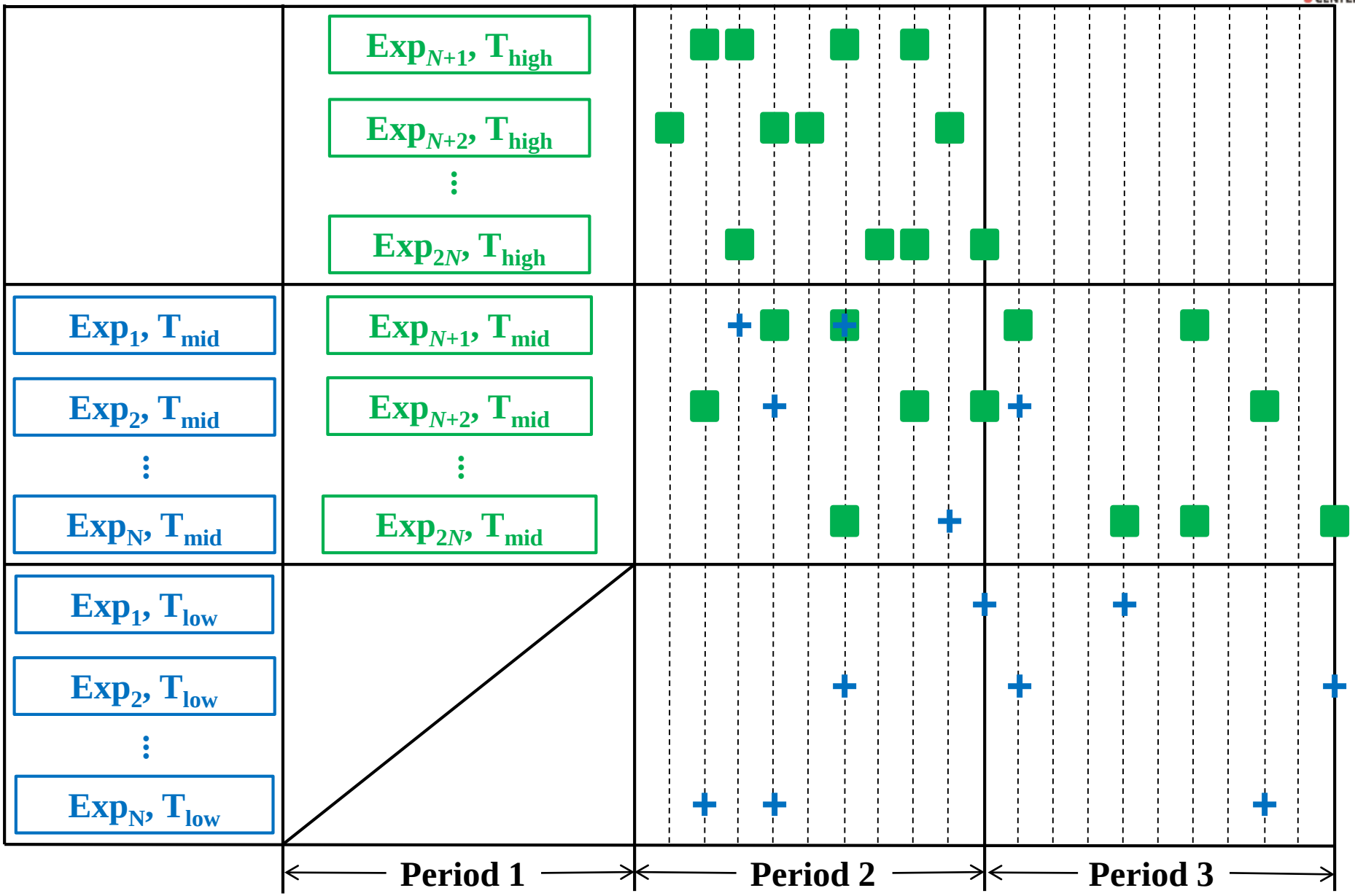
Update Parameter Estimates after Period 1 and Re-optimize Future Sampling Times with New Estimates



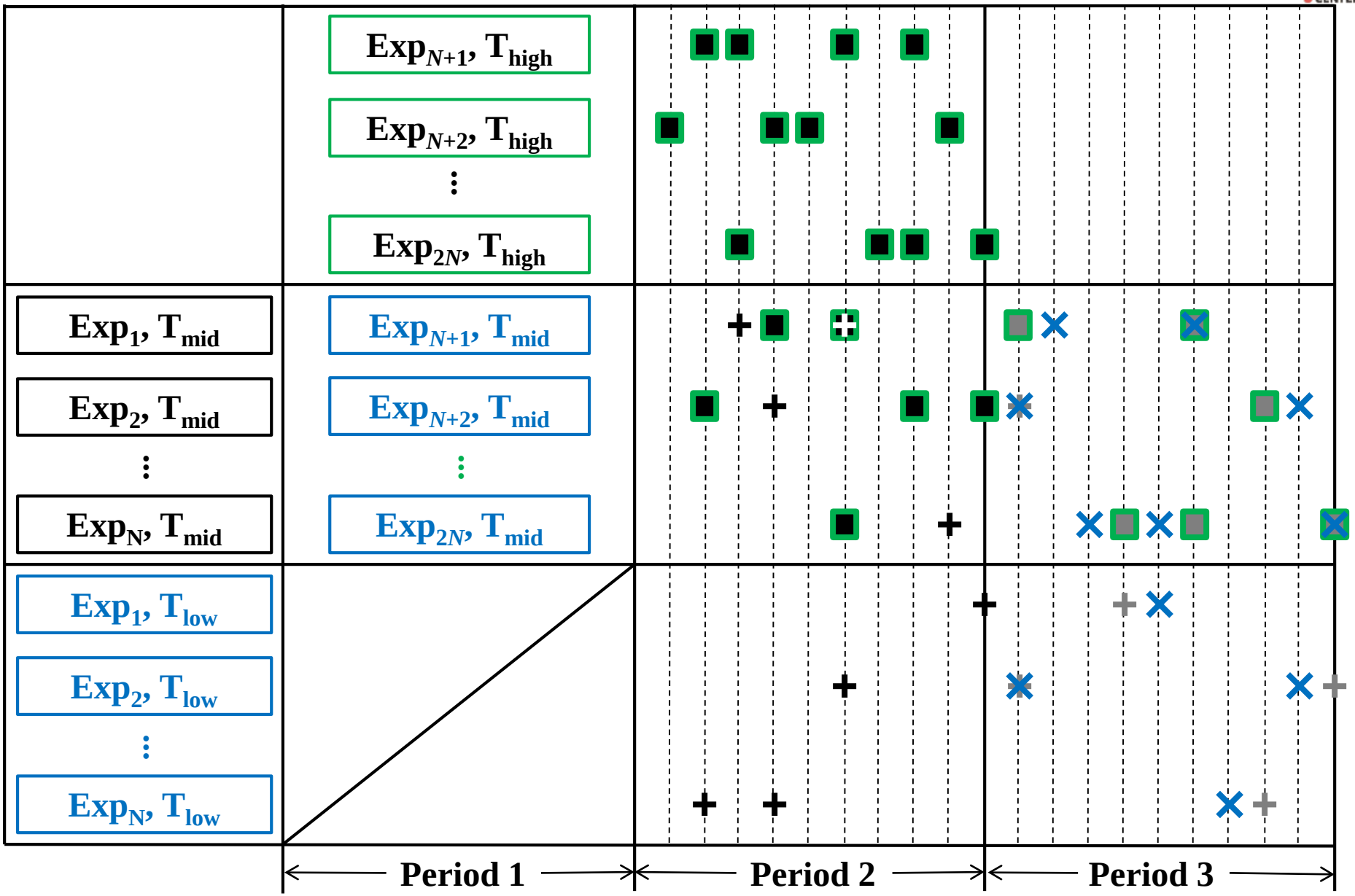
Initial Sampling Times

Re-optimized Sampling Times

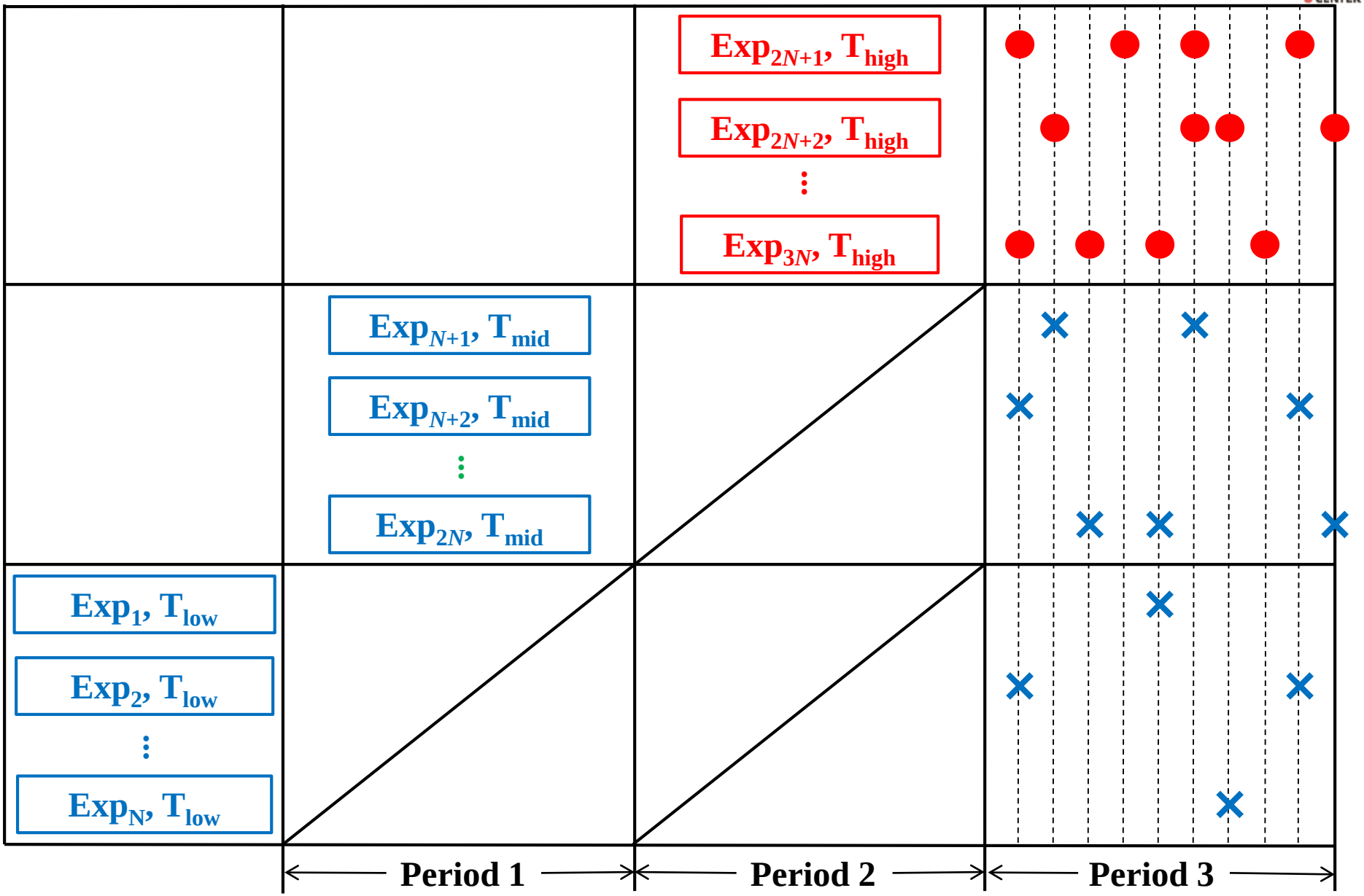
Update Parameter Estimates after Period 1 and Design Additional Experiments at Mid and High Temp with New Estimates



Update Parameter Estimates after Period 2 and Re-optimize Future Sampling Times with New Estimates



Update Parameter Estimates after Period 2 and Design Additional Experiments at High Temp with New Estimates

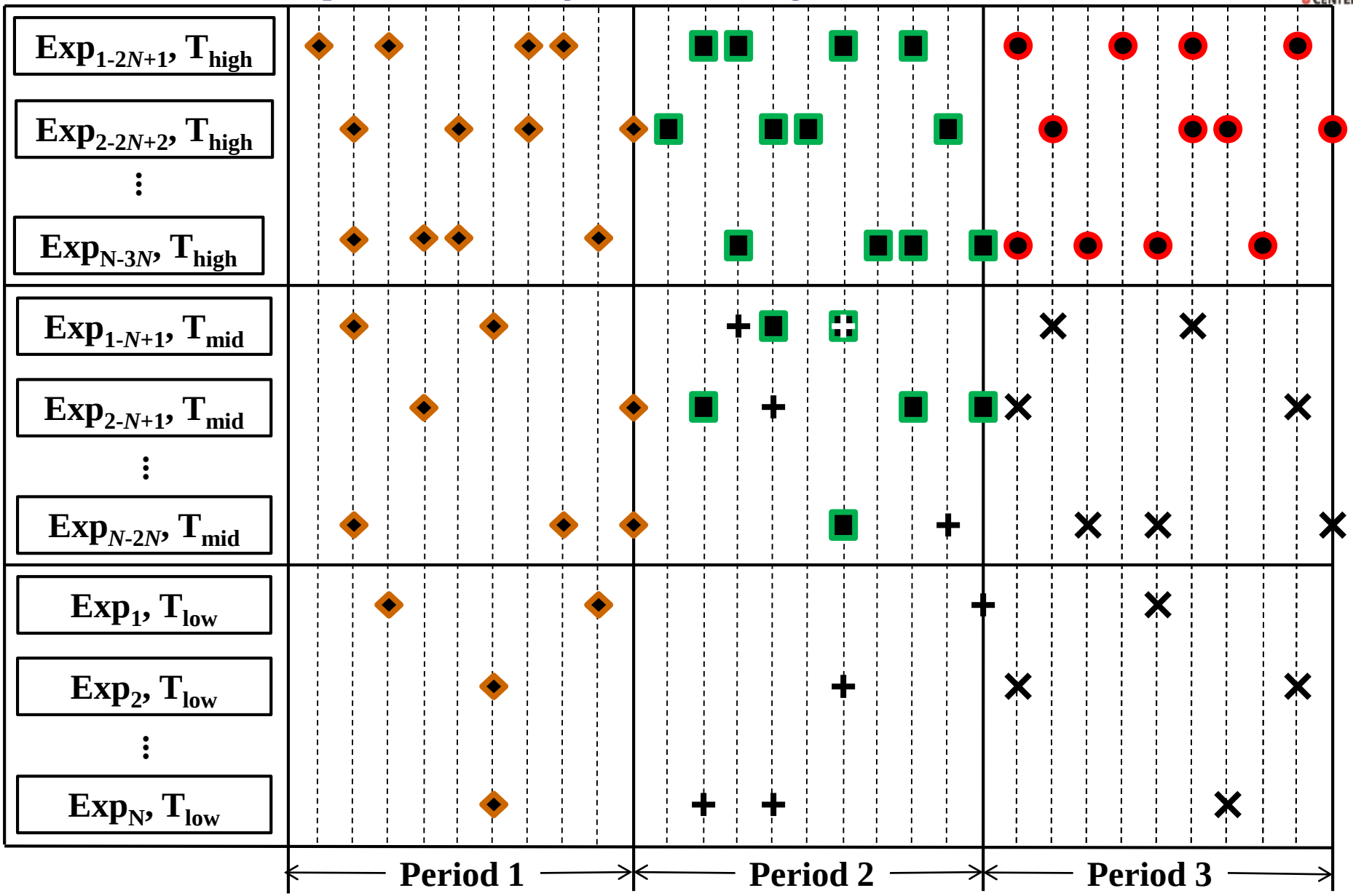


Carnegie Mellon

× Re-optimized Sampling Times

● New Experiments at T$_{mid/high}$

Summary of Adjusted Sampling Times and Sequential Experiments Using Parallel Design with Recourse



Algorithm for Parallel Design with Recourse