Strengthened Regression Models through Response Variable Bounds

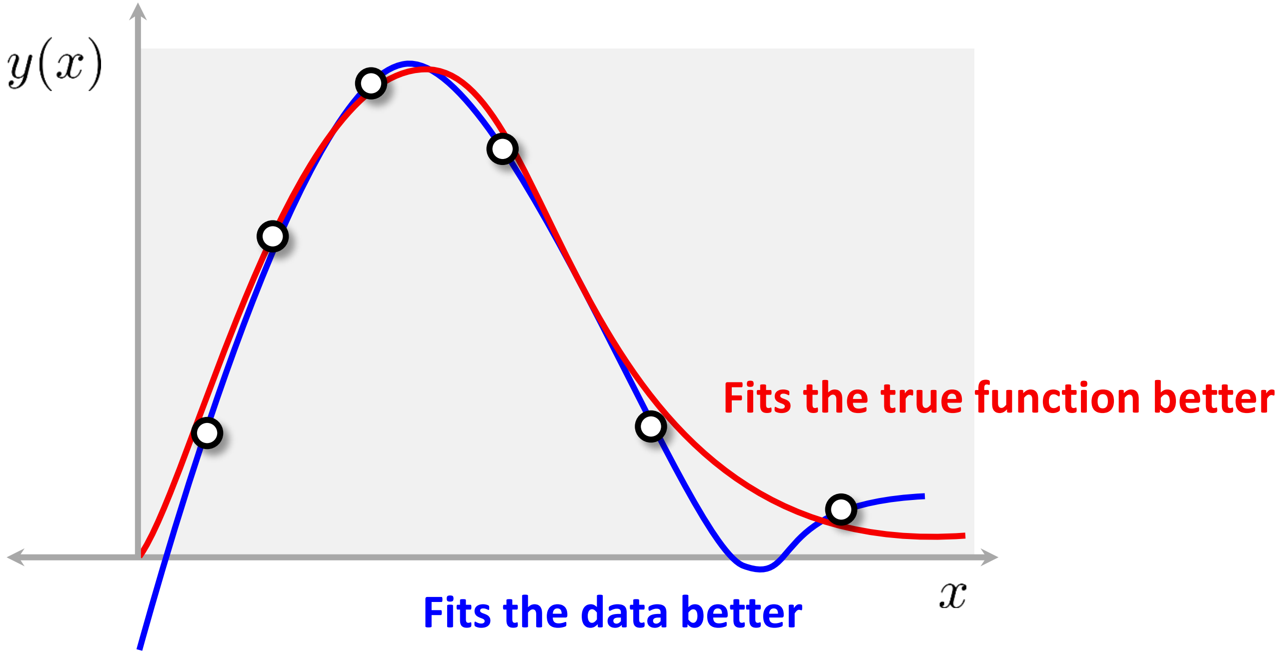
Alison Cozad[1,2], Nick Sahinidis[1,2], David Miller[1]

[1]National Energy Technology Laboratory, Pittsburgh, PA,USA

[2]Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, PA, USA

MOTIVATION

- **Leverage more information than just sampled data when building a surrogate model**



LEAST SQUARES REGRESSION

- **Ordinary least squares regression**
  - **Chooses regression coefficients based on a set of data points**

- **Generate a model for response,**

$$\hat{y}(x) = \beta_0 + \beta_1\, x + \beta_2\, x^2 + \beta_3\, \exp(x) + \ldots$$

- **Ordinary least squares regression problem**

$$\min_{\beta} \quad \sum_{i=1}^{N} \left(y_i - \hat{y}(x_i)\right)^2$$

OR

$$\min_{\beta} \quad \sum_{i=1}^{N} \left(y_i - [\beta_0 + \beta_1\, x_i + \beta_2\, x_i^2 + \beta_3\, \exp(x_i) + \ldots]\right)^2$$

Optimization variables

CONSTRAINED REGRESSION

- **What if more information is known of a less exact nature?**
  - **Leverage all information available to the modeler**
- **Explicit restrictions placed on regressors**
  - **Often times logical bounds on regressors can be found by inspection and/or analysis**
    - ***Ex: Physical constants***

$$k = \beta_0 \exp\left(\frac{\beta_1}{T}\right), \quad \beta_1 \leq 0 \rightarrow \text{Activation energy, positive}$$

  - **Relationships between parameters can be found by inspection or analysis**
    - ***Ex: Intuitive relationships***

$$H = \beta_0 + \beta_1 T^{\text{in}} + \beta_2 T^{\text{out}} + \ldots$$

$$\beta_2 - \beta_1 \geq 0 \rightarrow \text{Heat capacity, nonegative}$$

P. S. Knopov and A. S. Korkhin. *Regression analysis under a priori parameter restrictions*. Vol. 54. Springer, 2011.

ADDING EXPLICIT CONSTRAINTS

- **Adding in explicit constraints is rather straight forward,**

$$\hat{H}(T) = \beta_0 + \beta_1 T^{\text{in}} + \beta_2 T^{\text{out}} + \ldots$$

$$\beta_2 - \beta_1 \geq 0 \rightarrow \text{Heat capacity, nonegative}$$

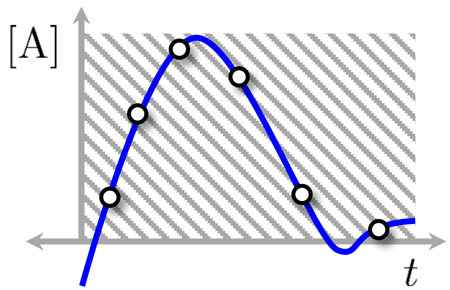
$$\min_{\beta} \quad \sum_{i=1}^{N} \left(H_i - \left[\beta_0 + \beta_1 T_i^{\text{in}} + \beta_2 T_i^{\text{out}} + \ldots\right]\right)^2$$

$$\text{s.t.} \quad \beta_2 - \beta_1 \geq 0$$

CONSTRAINED REGRESSION

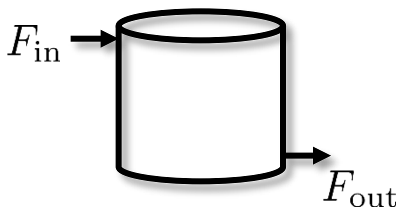
- **What if more information is known of a less exact nature?**
  - **Leverage all information available to the modeler**

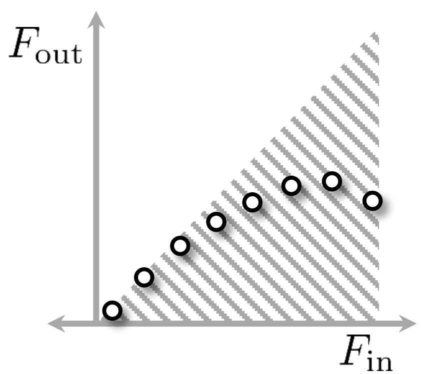
- **Restrictions implied by constraints on dependent variables**
  - **Implied by bounds on dependent variable**
    - ***Ex:***



$$0 \leq [\mathrm{A}]_t \leq [\mathrm{A}]^{\max}$$

  - **Implied by constraints on dependent variable**
    - ***Ex:***





$$F_{\text{out}}(F_{\text{in}}) \leq F_{\text{in}}$$

ADDING IMPLIED CONSTRAINTS

- **Adding in constraints implied by dependent variable bounds is less straight forward,**

$$\text{Generate a model for } y \text{ given that } y^l \leq y \leq y^u$$

$$\hat{y}(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots$$

$$\min_{\beta} \quad \sum_{i=1}^{N} \left(y_i - [\beta_0 + \beta_1 x + \beta_2 x^2 + \ldots]\right)^2$$

$$\text{s.t.} \quad y^l \leq \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots \leq y^u \qquad \forall x$$

ADDING IMPLIED CONSTRAINTS

- **Adding in constraints implied by dependent variable bounds is less straight forward,**

Generate a model for y given that $y^l \leq y \leq y^u$

$$\hat{y}(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots$$

$$\min_{\beta} \quad \sum_{i=1}^{N} \left(y_i - [\beta_0 + \beta_1 x + \beta_2 x^2 + \dots]\right)^2$$

$$\text{s.t.} \quad y^l \leq \beta_0 + \beta_1 x + \beta_2 x^2 + \dots \leq y^u \qquad \forall x$$

Semi-infinite programming

ADDING IMPLIED CONSTRAINTS

- **Adding in constraints implied by dependent variable bounds is less straight forward,**

$$\text{Generate a model for } y \text{ given that } y^l \leq y \leq y^u$$

$$\hat{y}(x) = \beta_0 + \beta_1\, x + \beta_2 x^2 + \ldots$$

$$\min_{\beta} \quad \sum_{i=1}^{N} \left(y_i - \left[\beta_0 + \beta_1\, x + \beta_2 x^2 + \ldots\right]\right)^2$$

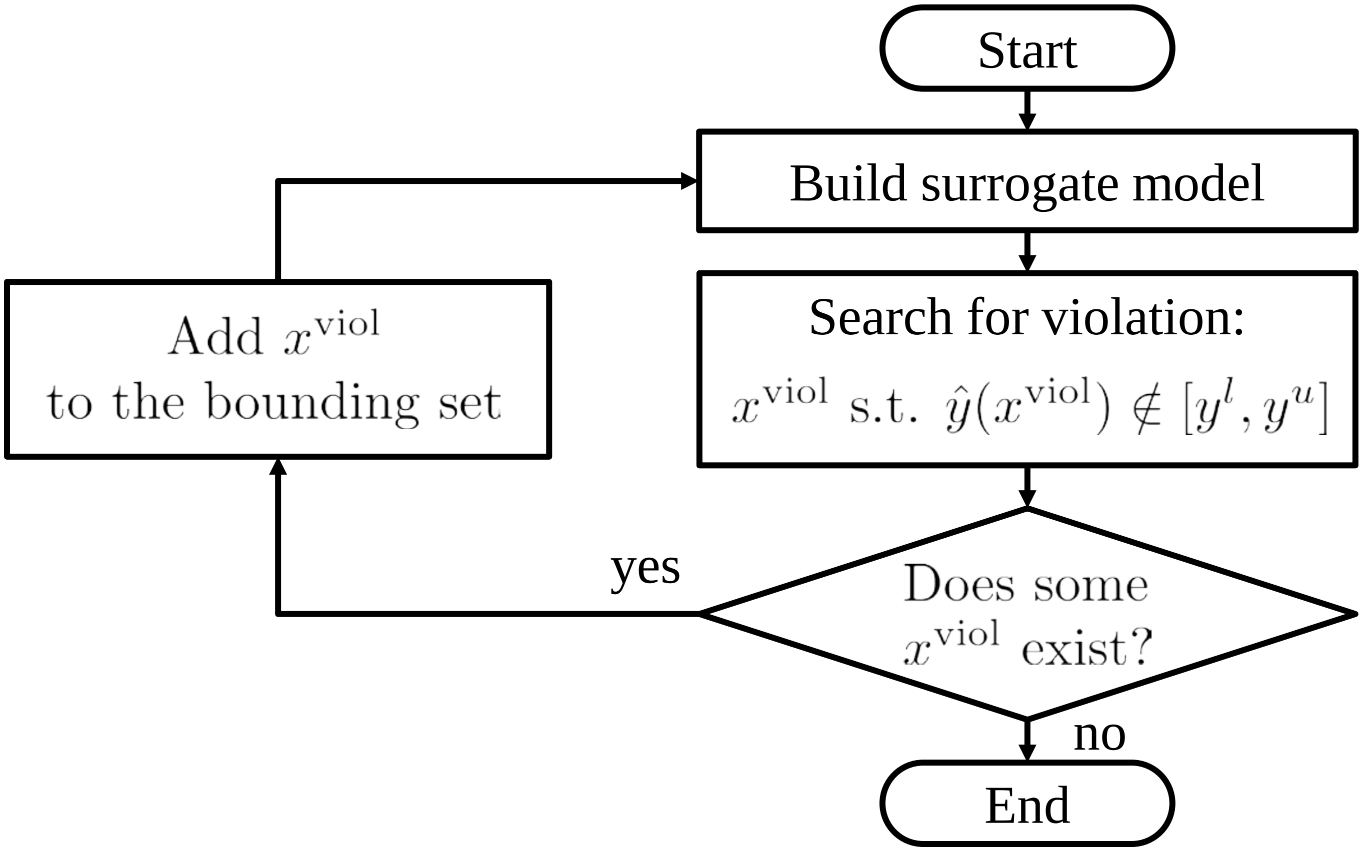
$$\text{s.t.} \quad y^l \leq \beta_0 + \beta_1\, x_j + \beta_2 x_j^2 + \ldots \leq y^u$$

$$\forall j \in \text{Bounding set}$$

IMPLEMENTATION

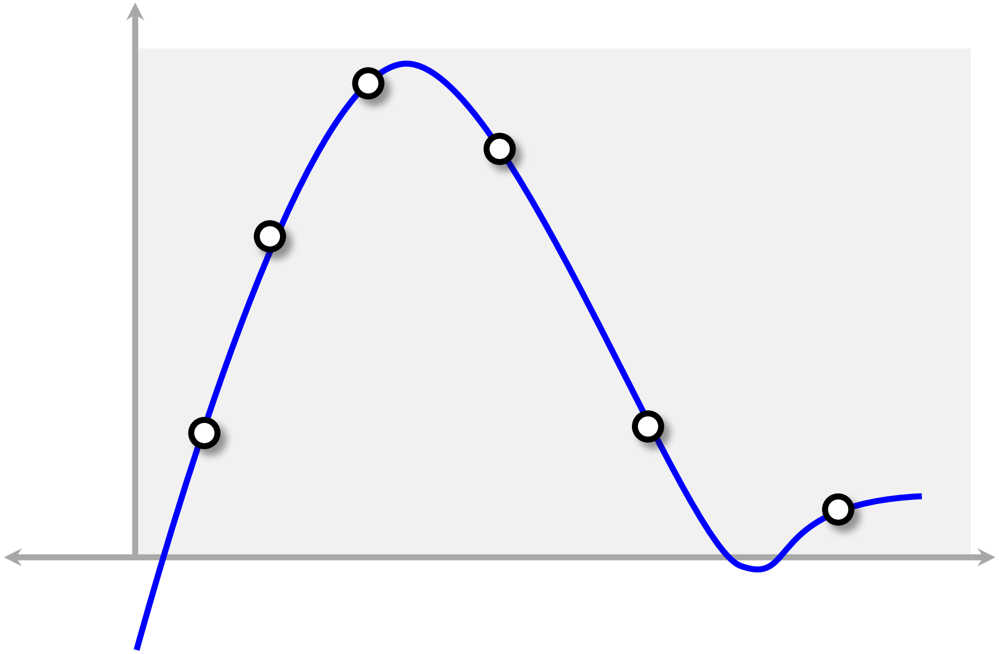
Generate a model for y given that $y^l \leq y \leq y^u$

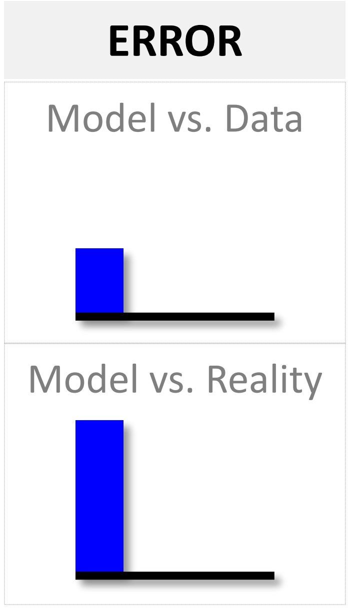
$$\hat{y}(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots$$



ILLUSTRATIVE EXAMPLE

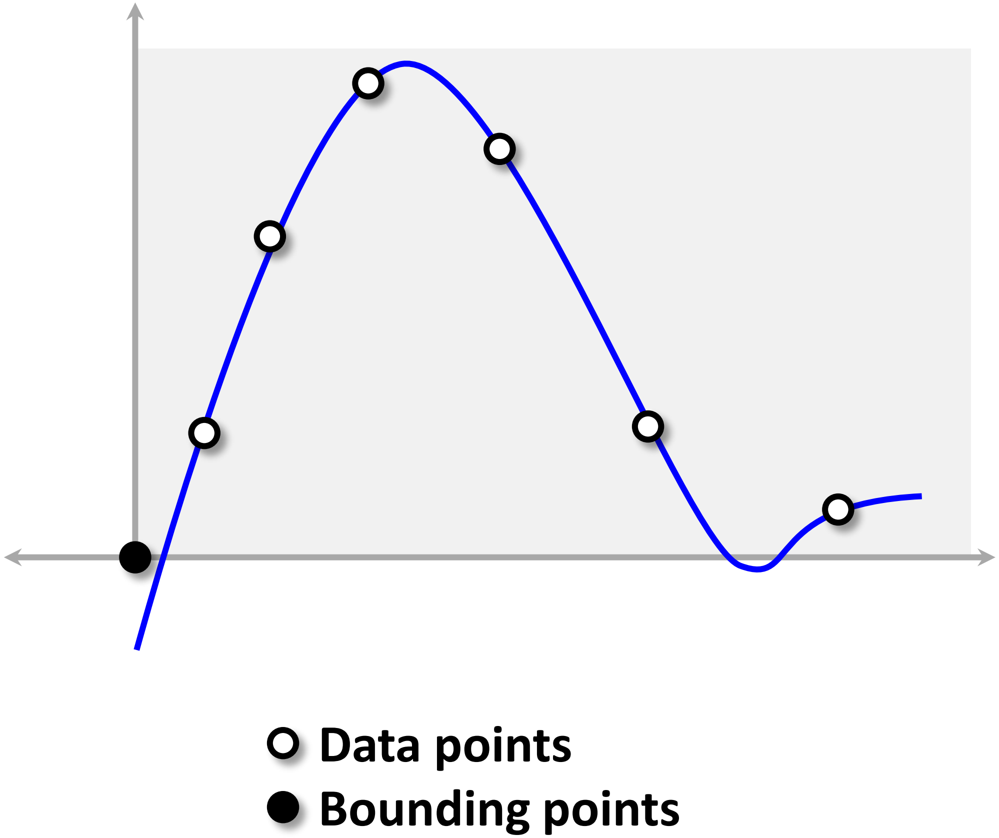
- **Build an initial model**

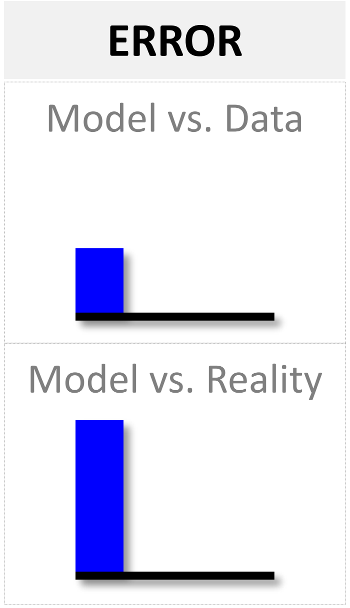




ILLUSTRATIVE EXAMPLE

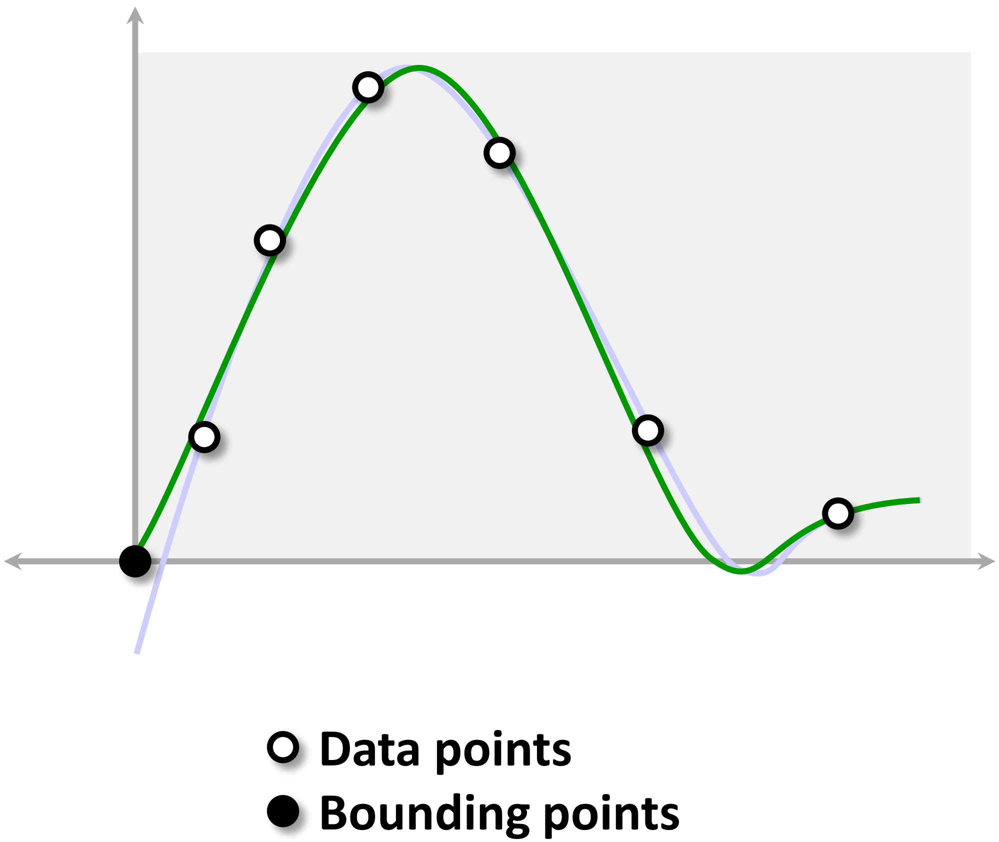
- **Build an initial model**
- **Locate areas that violate output bounds**

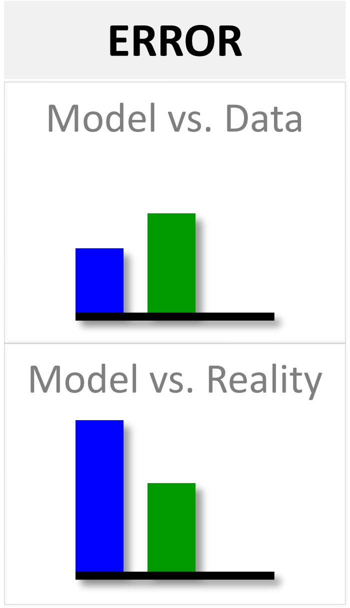




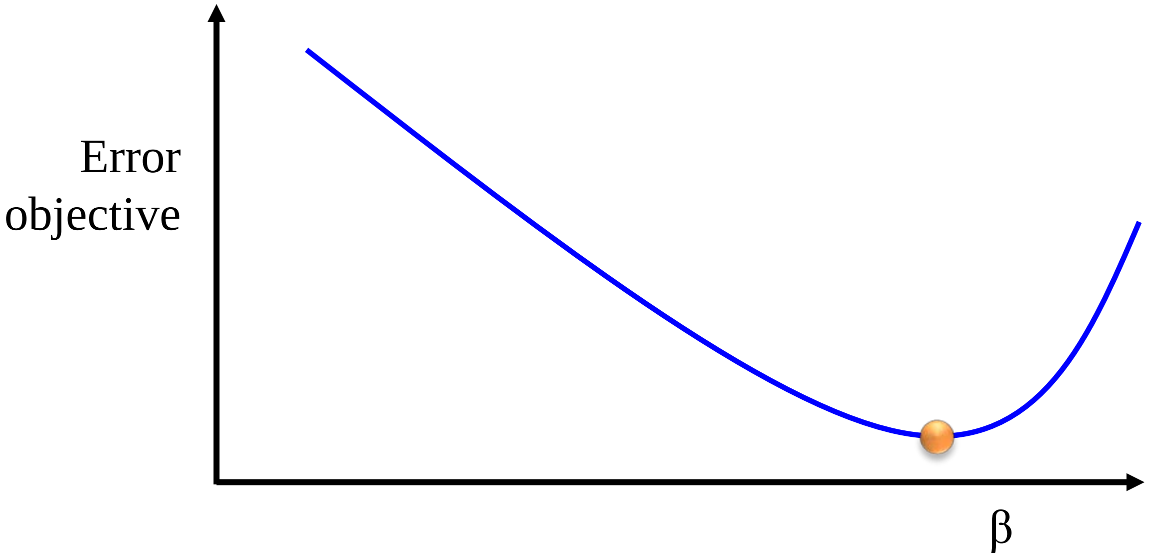
ILLUSTRATIVE EXAMPLE

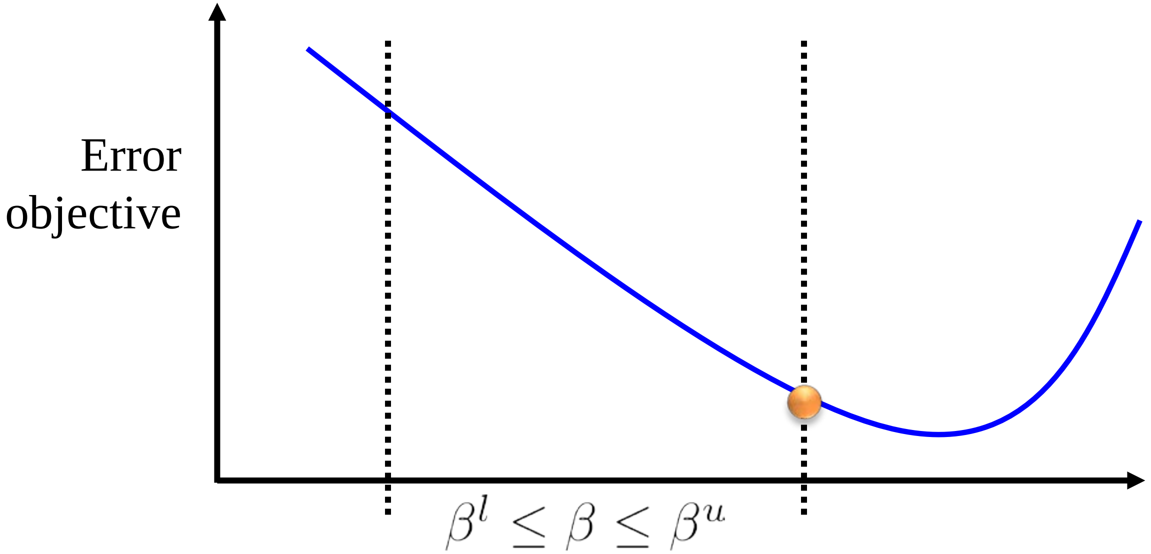
- **Rebuild the model ensuring the output is within bounds at the bounding points**





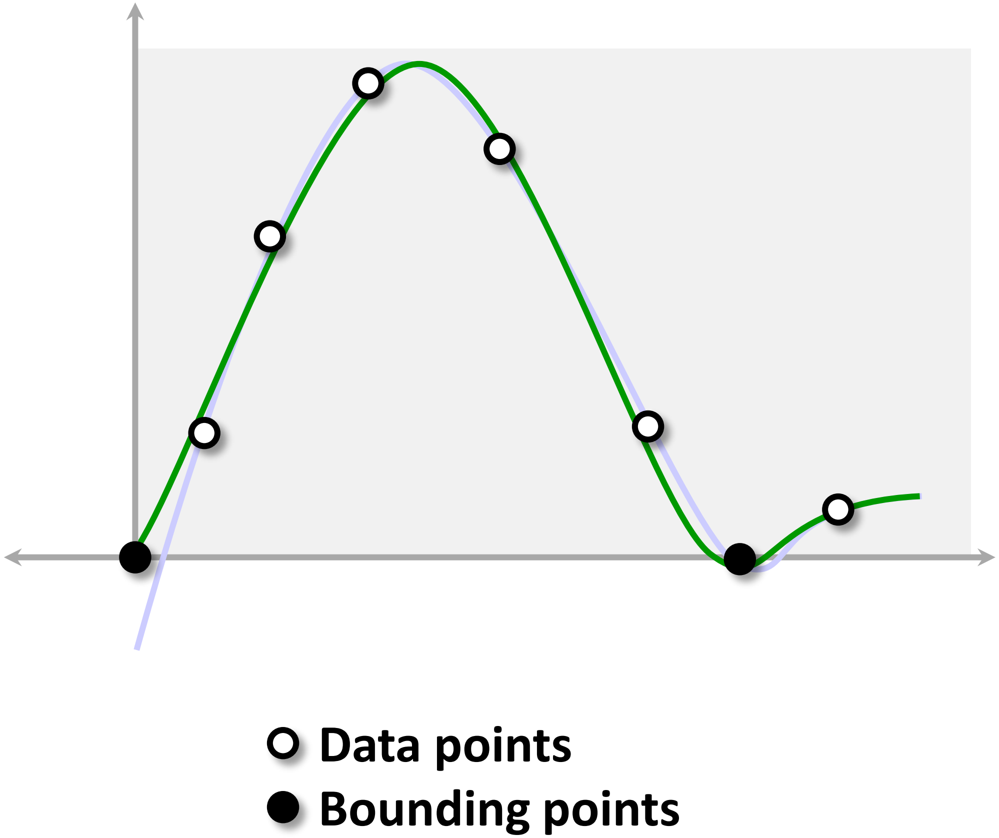
NON STATIONARITY

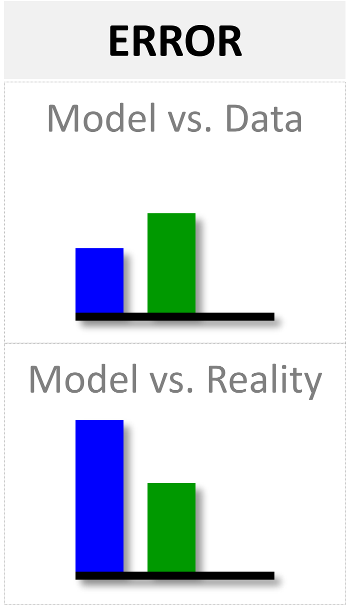




ILLUSTRATIVE EXAMPLE

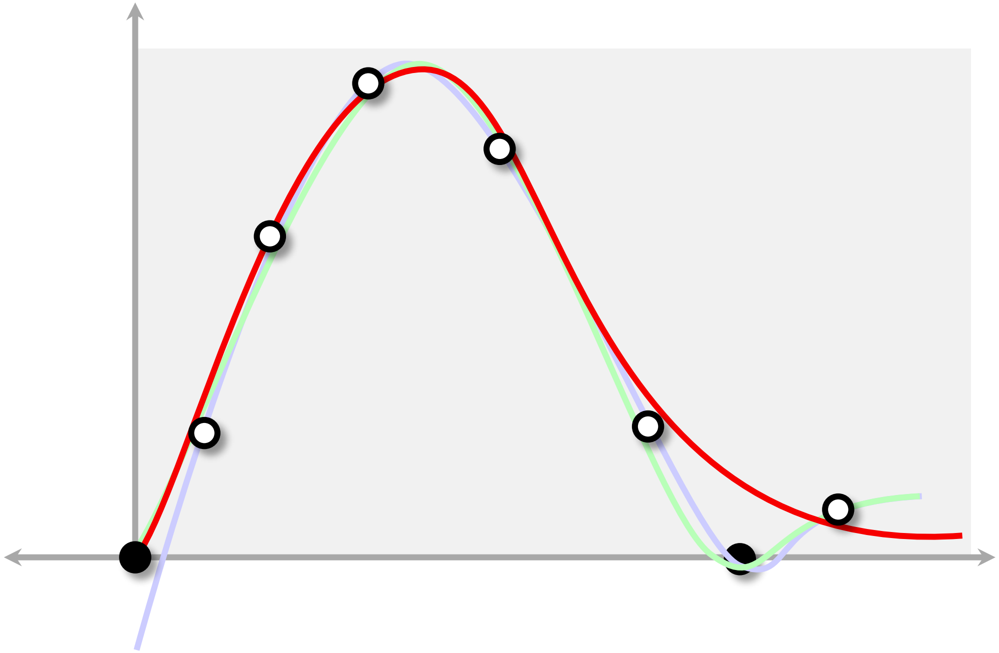
- **Rebuild the model ensuring the output is within bounds at the bounding points**
- **Search for additional violation points**

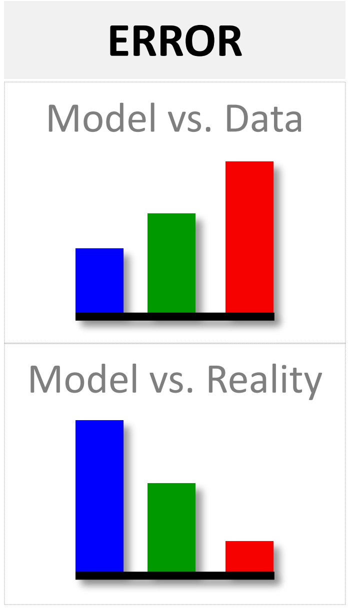




ILLUSTRATIVE EXAMPLE

- Rebuild the model ensuring the output is within bounds at the bounding points
- Ensure that no violation points remain



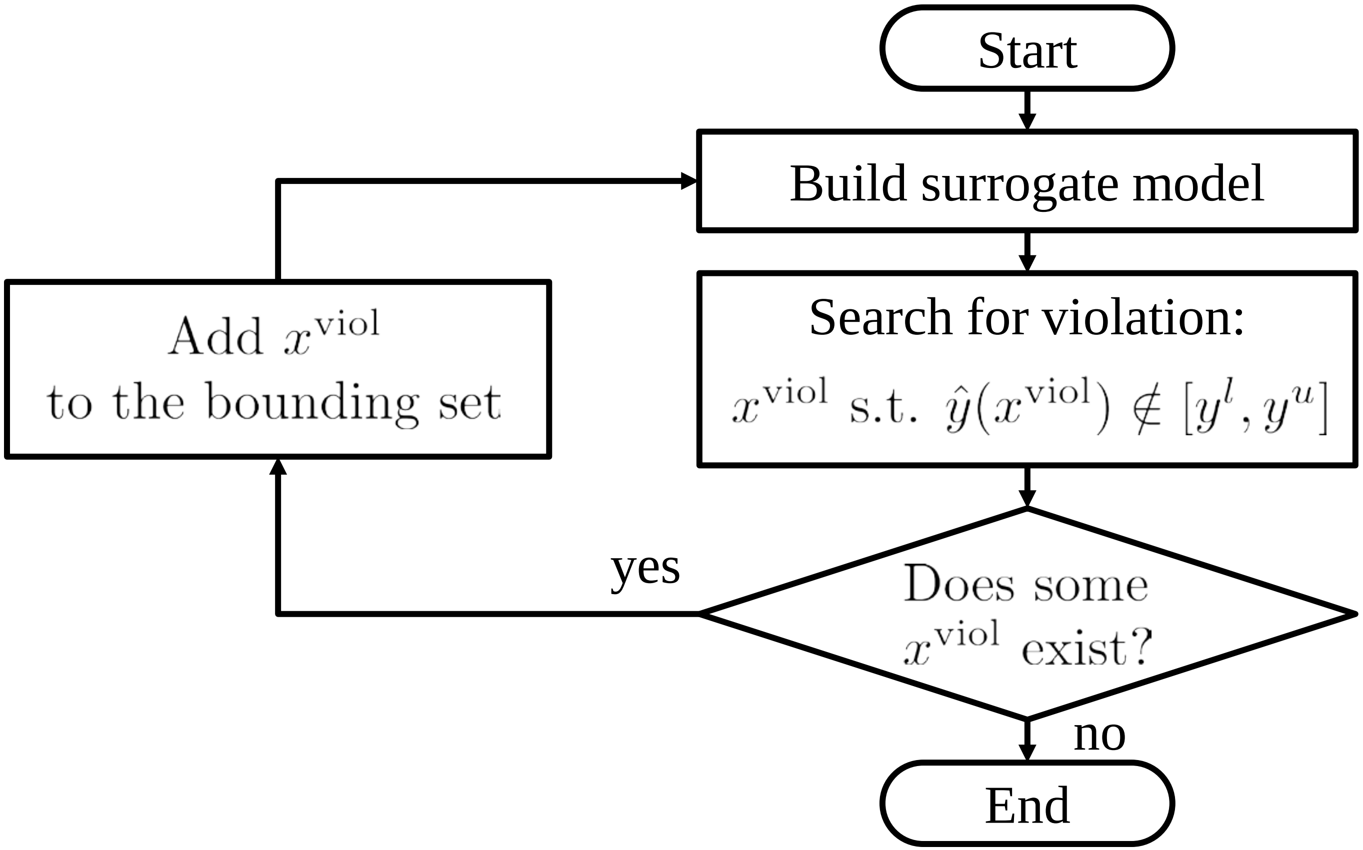


○ Data points
● Bounding points

IMPLEMENTATION

Generate a model for y given that $y^l \leq y \leq y^u$

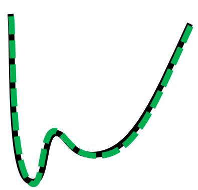
$$\hat{y}(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots$$



TESTING PLATFORM

- **We will test this implementation on an existing software: ALAMO**

- **ALAMO (Automated Learning of Algebraic Models for Optimization)**
  - **Iteratively sample and model black box systems as algebraic model that**
    - ***Accurate***
      - **We want to reflect the true nature of the simulation**
    - ***Simple***
      - **Low-complexity models**



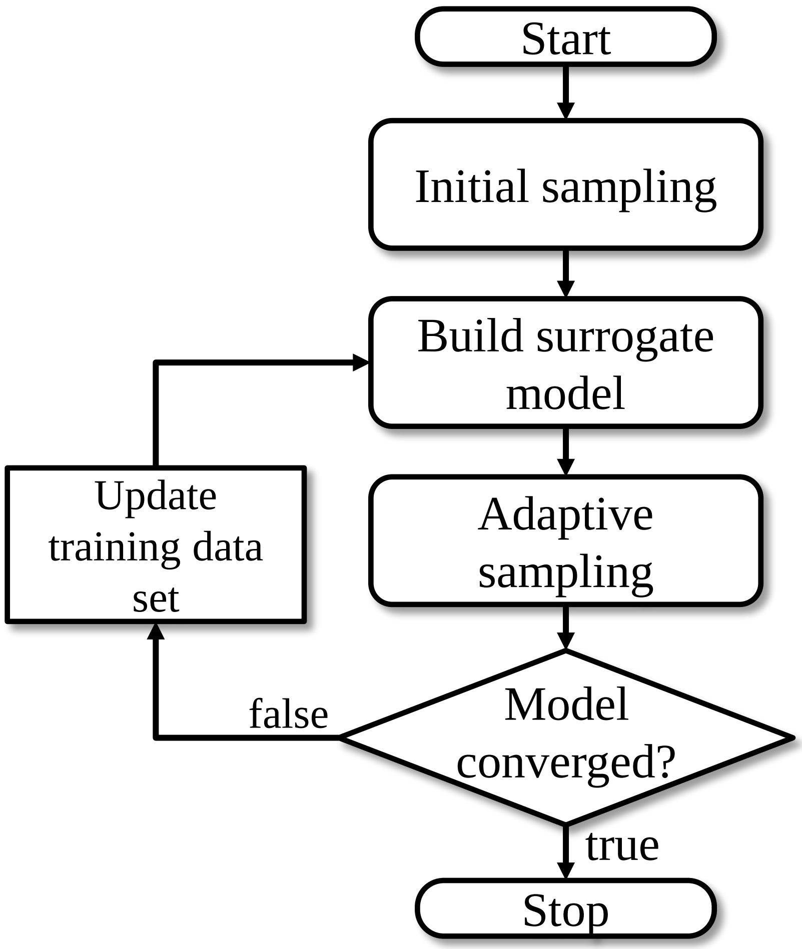
$$\hat{f}(x) = \sum_{i=1}^{n} \gamma_i \exp\left(\frac{\|x\|}{\sigma^2}\right) + \beta_0 + \beta_1\, x + \ldots$$

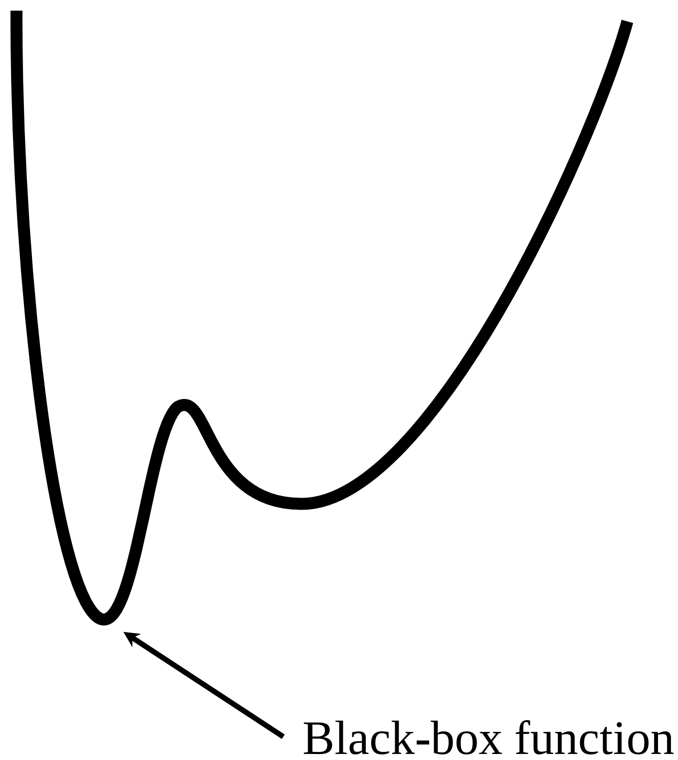
$$\hat{f}(x) = \beta_1\, x + \beta_2\, x^2 + \beta_3\, x^3 + \beta_4\, e^x$$

    - ***Generated from a minimal data set***
      - **Reduce experimental and simulation requirements**

ALAMO

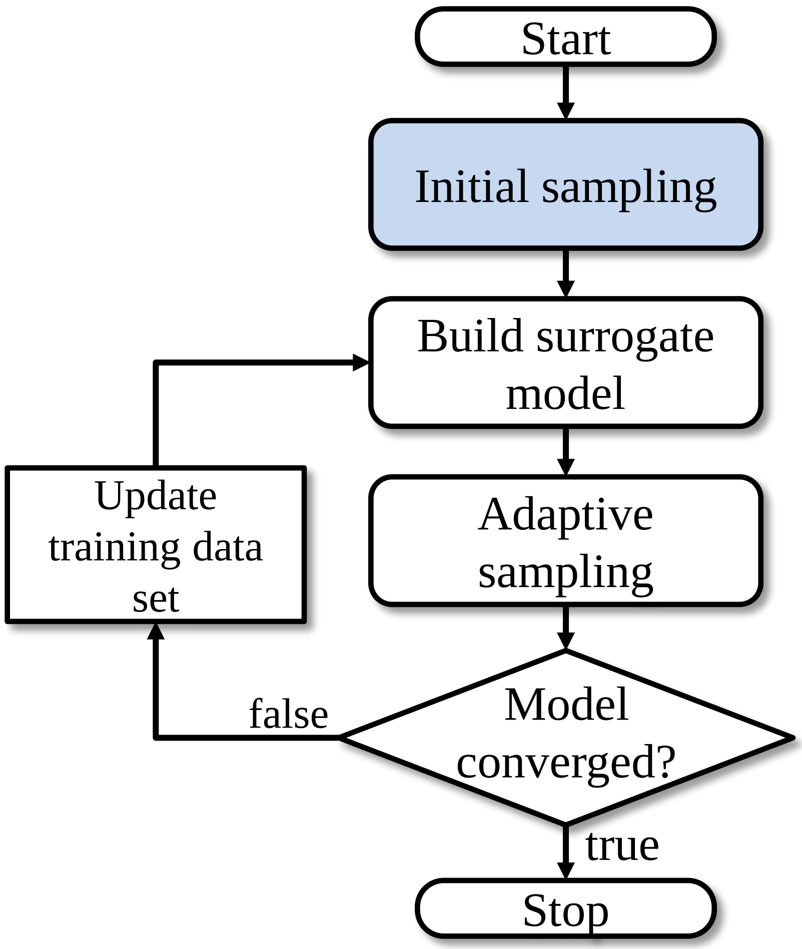
Automated Learning of Algebraic Models for Optimization

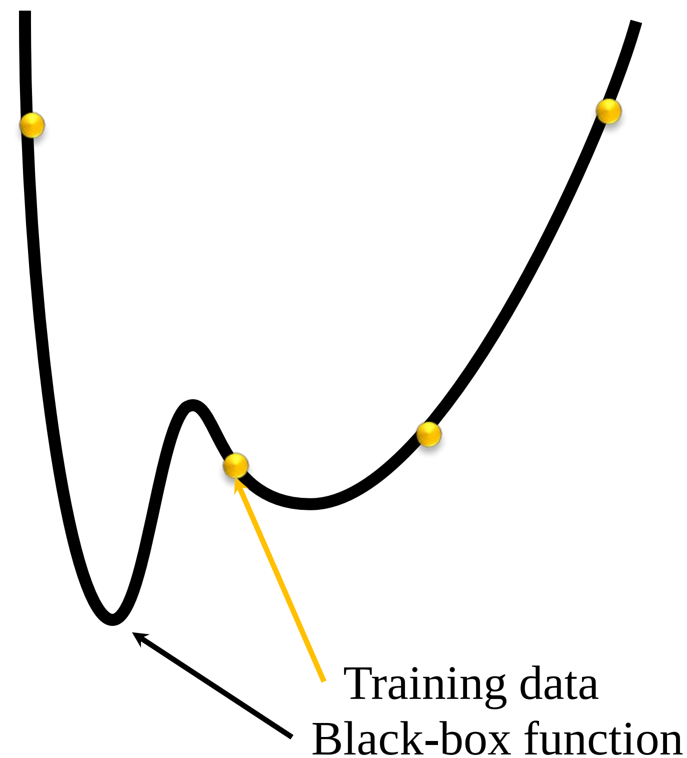




ALAMO

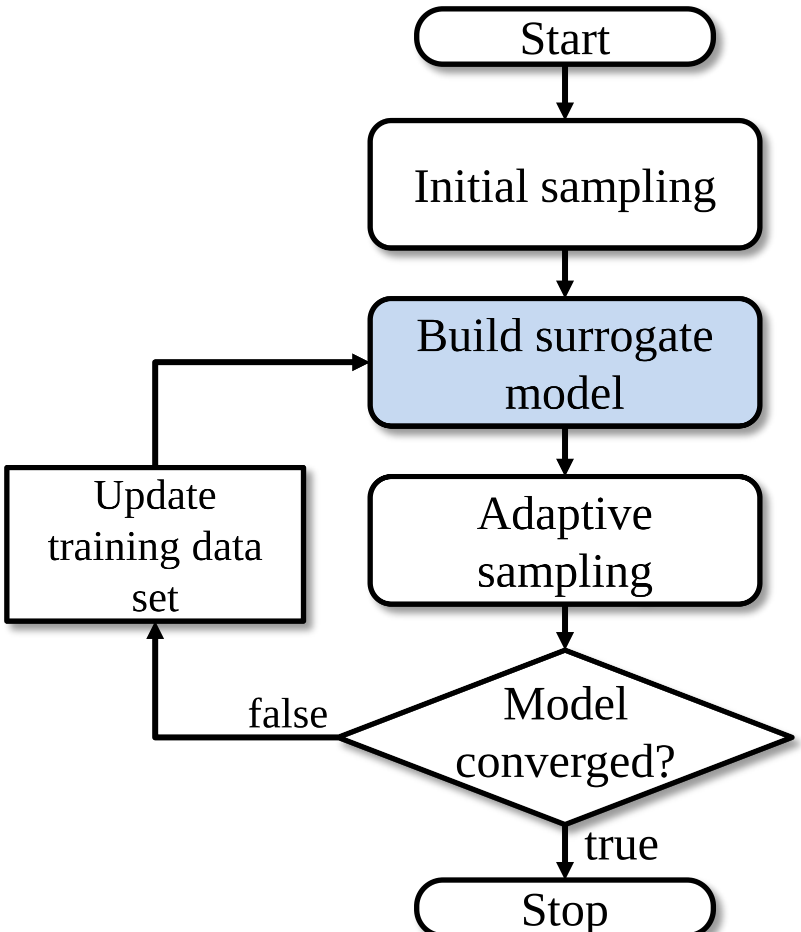
Automated **L**earning of **A**lgebraic **M**odels for **O**ptimization

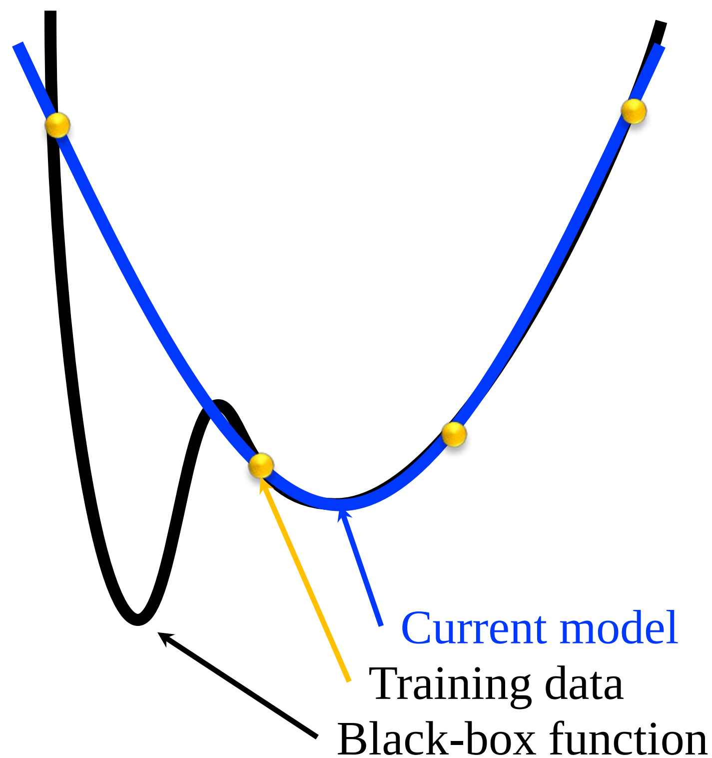




ALAMO

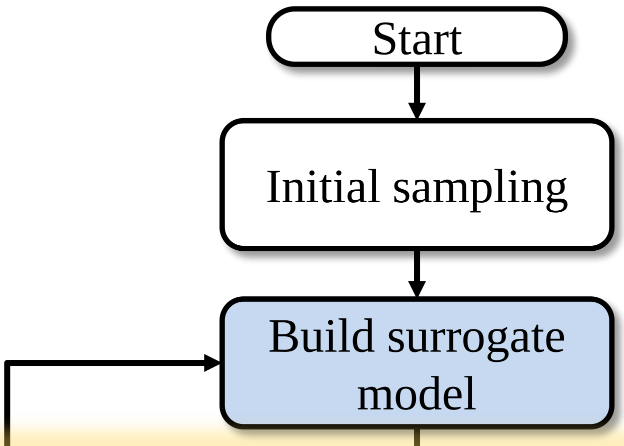
Automated Learning of Algebraic Models for Optimization





ALAMO

Automated Learning of Algebraic Models for Optimization



Determining the unknown functional form

Step 1: Define a large set of potential basis functions

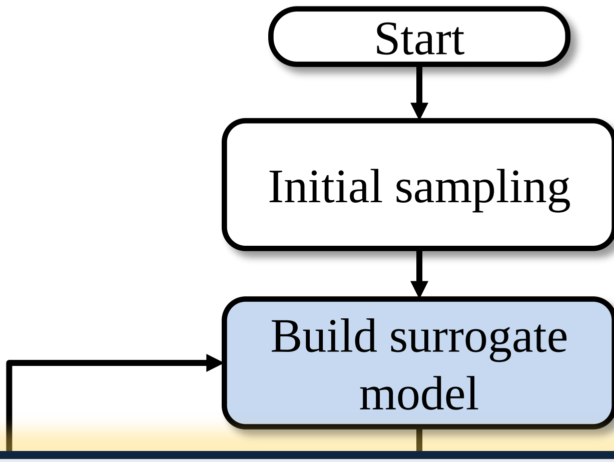
$$\hat{z}(x_1) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 \frac{x_1}{x_2} + \beta_5 \frac{x_2}{x_1} + \beta_6 e^{x_1} + \beta_7 e^{x_2} + \ldots$$

Step 2: Model reduction

$$\hat{z}(x) = \beta_0 + \beta_2 x_2 + \beta_5 \frac{x_2}{x_1} + \beta_7 e^{x_2}$$

ALAMO

Automated Learning of Algebraic Models for Optimization



Using the new method, we **guarantee** that the model does not violate output variable bounds

Determining the unknown functional form

Step 1: Define a large set of potential basis functions

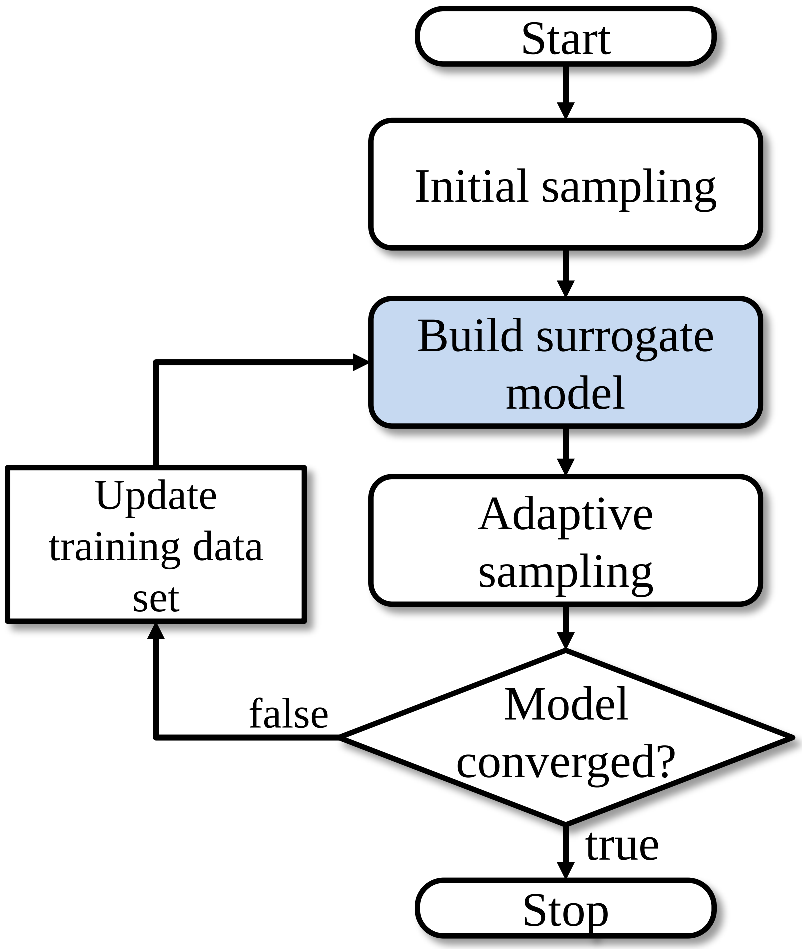
$$\hat{z}(x_1) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 \frac{x_1}{x_2} + \beta_5 \frac{x_2}{x_1} + \beta_6 e^{x_1} + \beta_7 e^{x_2} + \ldots$$

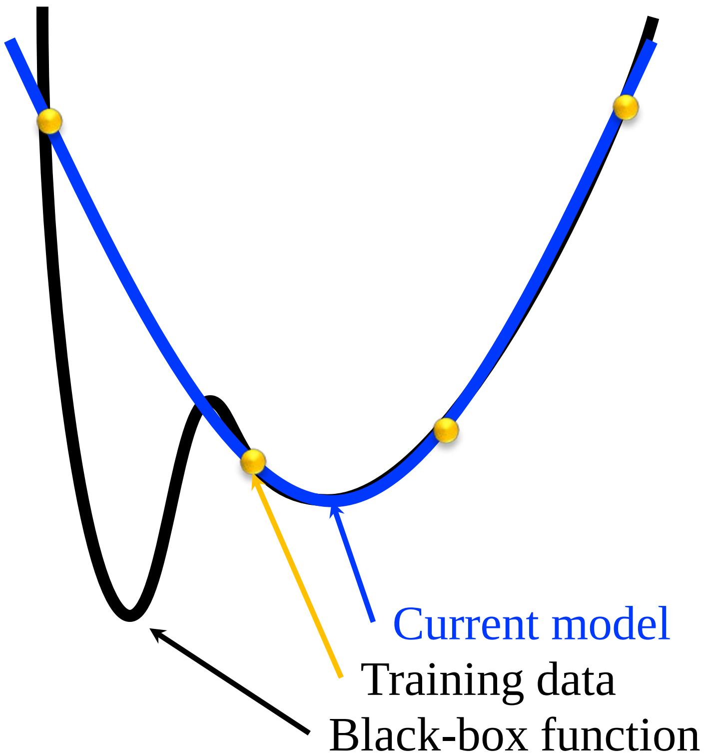
Step 2: Model reduction

$$\hat{z}(x) = \beta_0 + \beta_2 x_2 + \beta_5 \frac{x_2}{x_1} + \beta_7 e^{x_2}$$

ALAMO

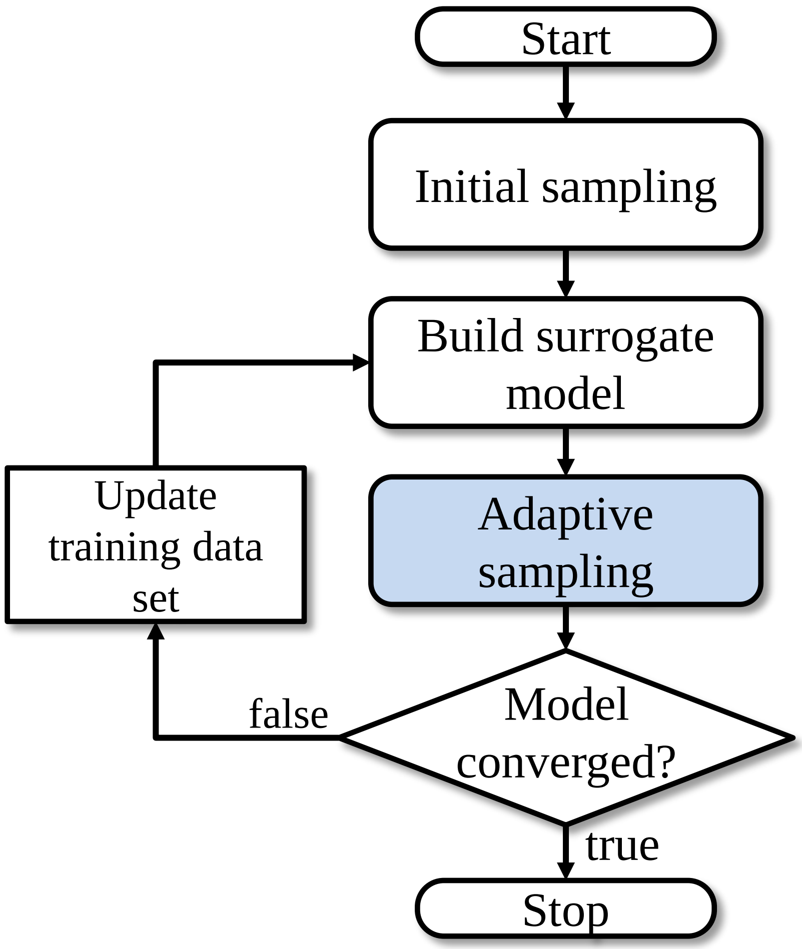
Automated Learning of Algebraic Models for Optimization

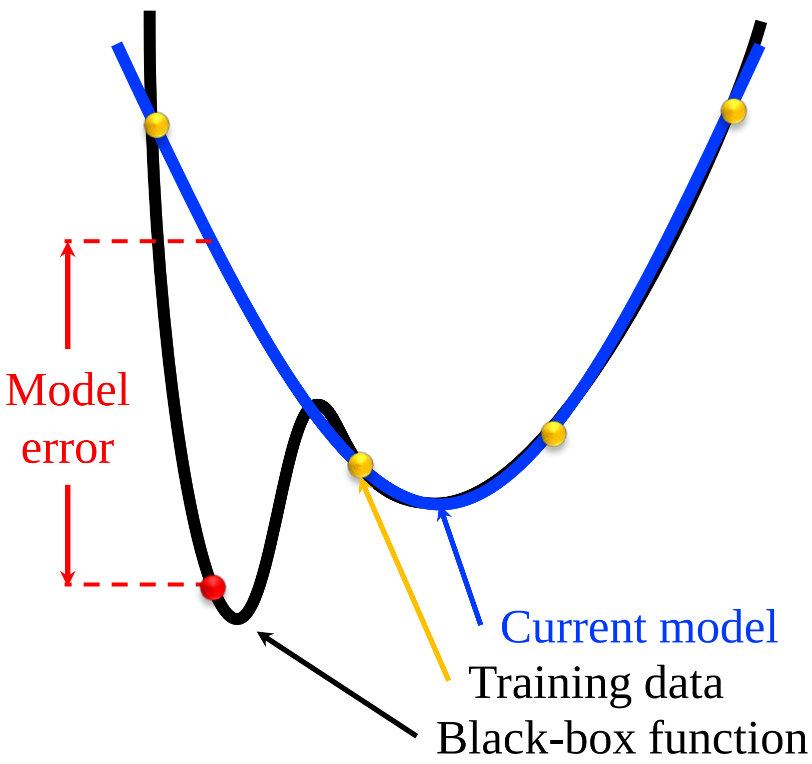




ALAMO

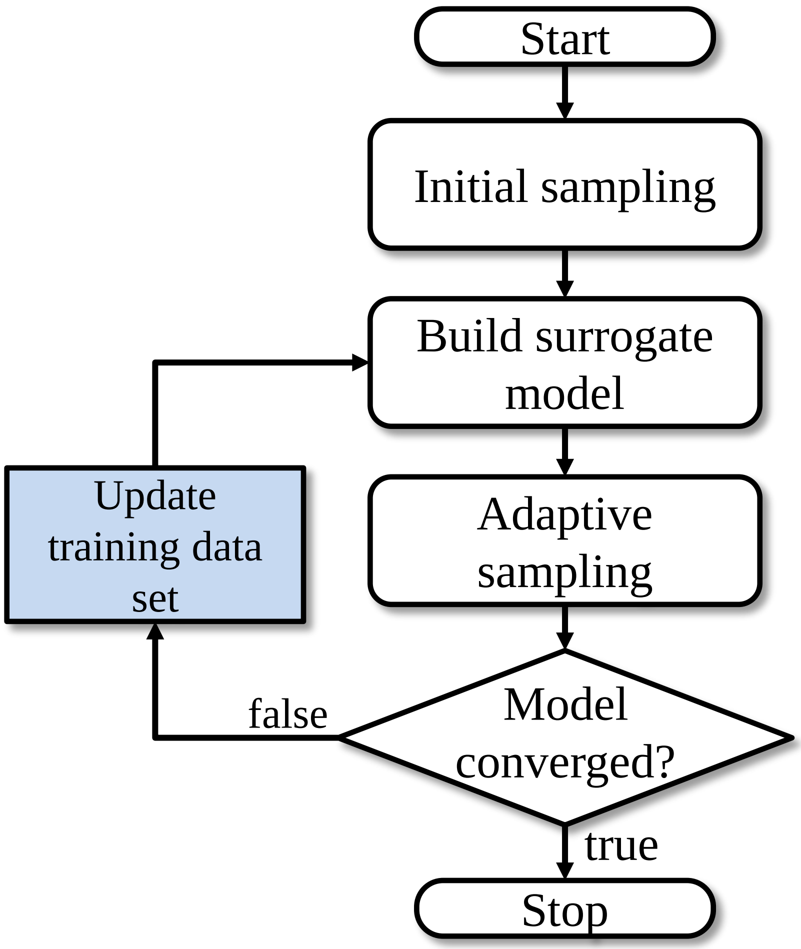
Automated **L**earning of **A**lgebraic **M**odels for **O**ptimization

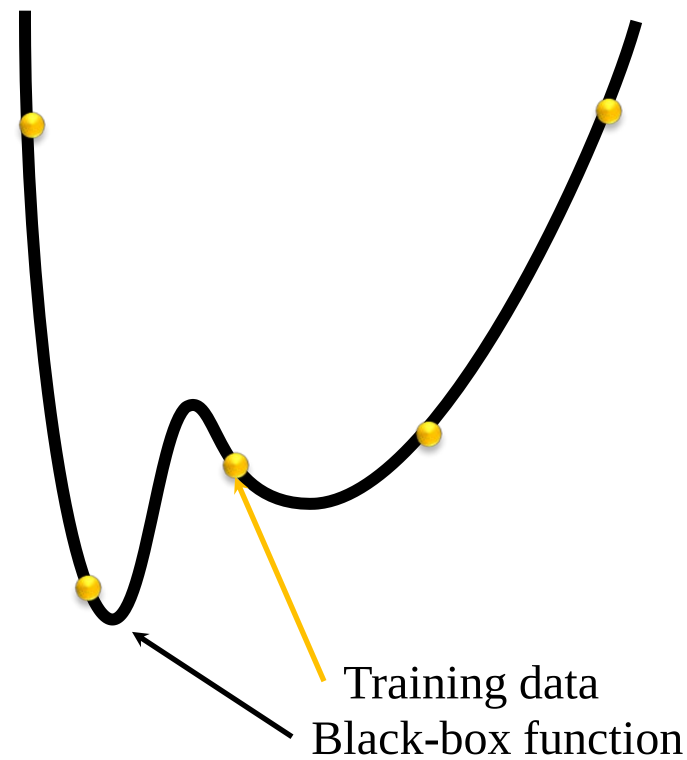




ALAMO

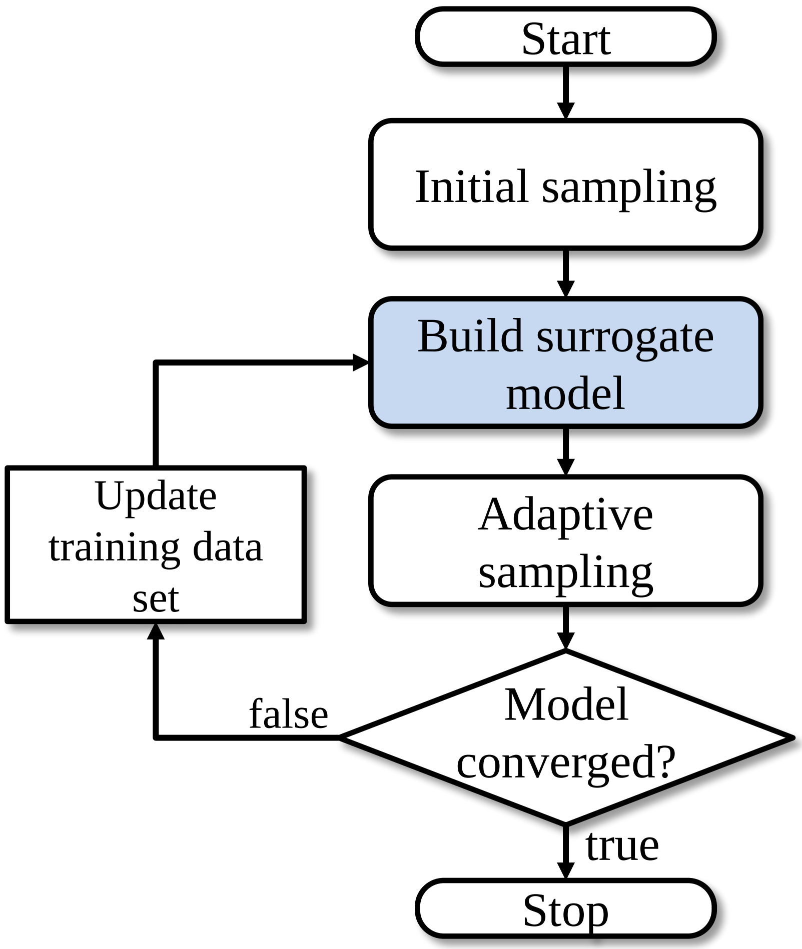
Automated Learning of Algebraic Models for Optimization

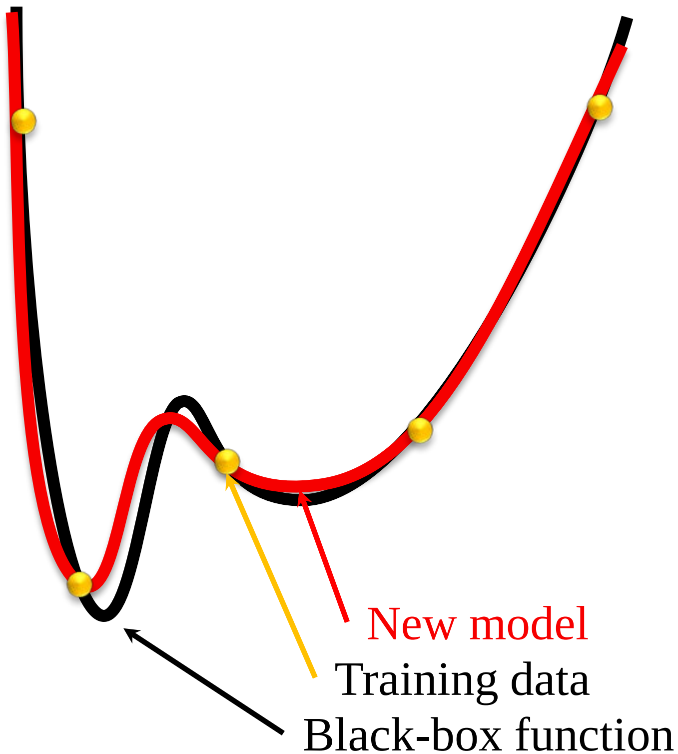




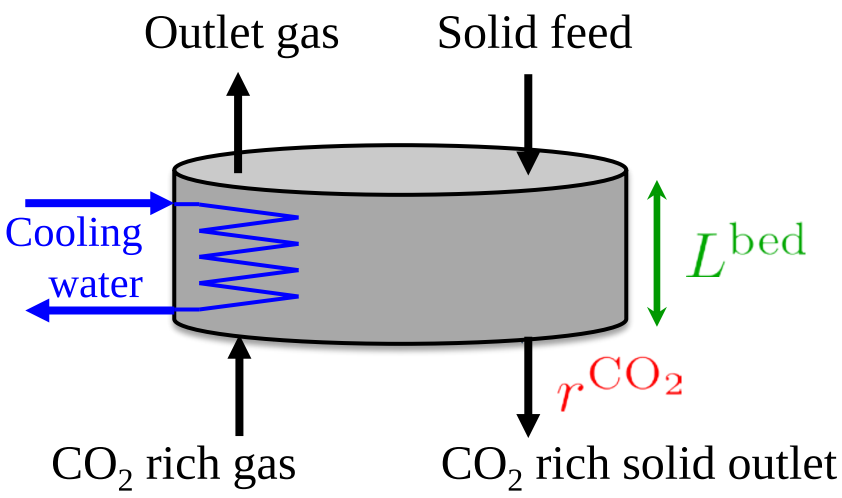
ALAMO

Automated **L**earning of **A**lgebraic **M**odels for **O**ptimization





SMALL EXMAPLE – CARBON CAPTURE



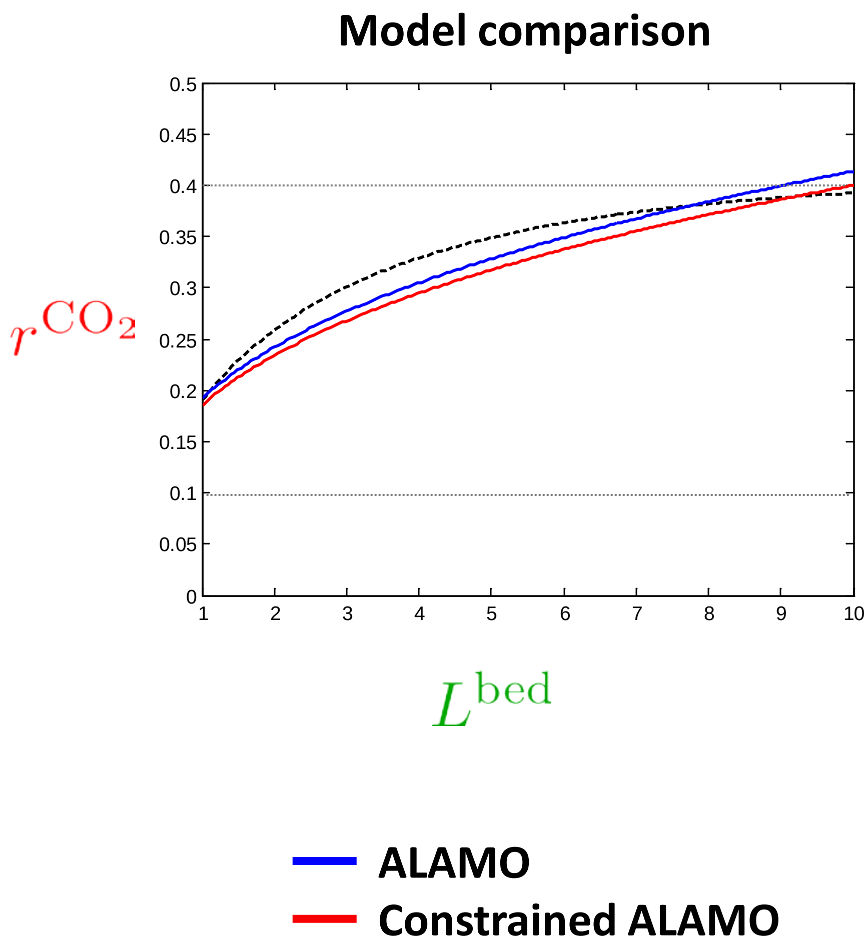
- **Dependent variable and range**
  - **Fraction of CO2 remove from the gas stream**

$$r^{\mathrm{CO_2}}(L^{\mathrm{bed}}, F^{\mathrm{cw}}) = f_1(L^{\mathrm{bed}}, F^{\mathrm{cw}}) \qquad 0.10 \leq r^{\mathrm{CO_2}} \leq 0.4$$

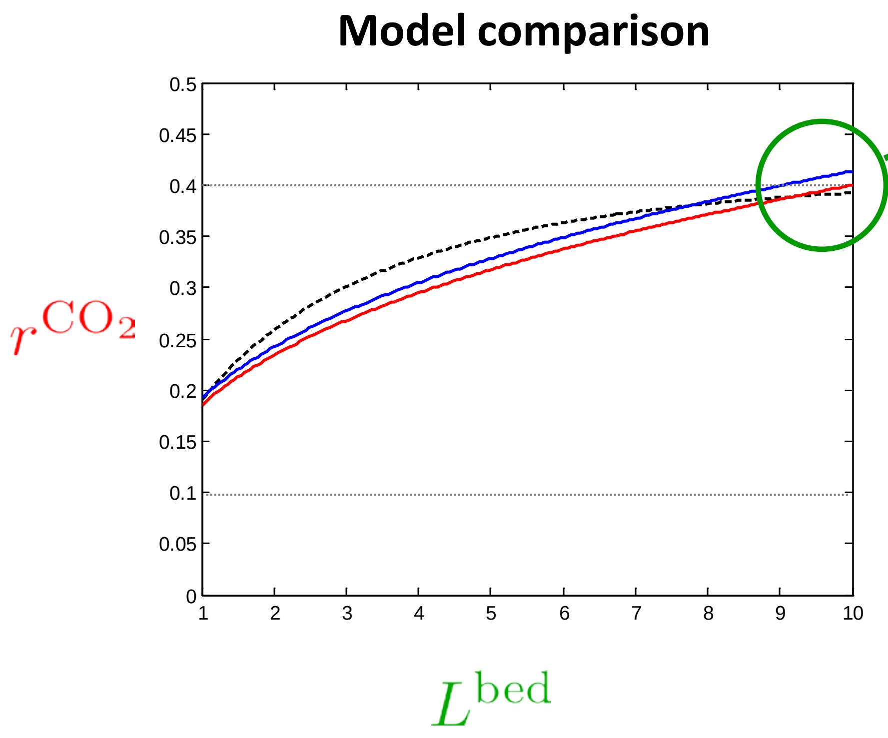
- **Independent variable and range**

$$1 \text{ m} \quad \leq L^{\mathrm{bed}} \quad \leq 10 \text{ m}$$

COMPARISON – ITER 1



COMPARISON – ITER 1



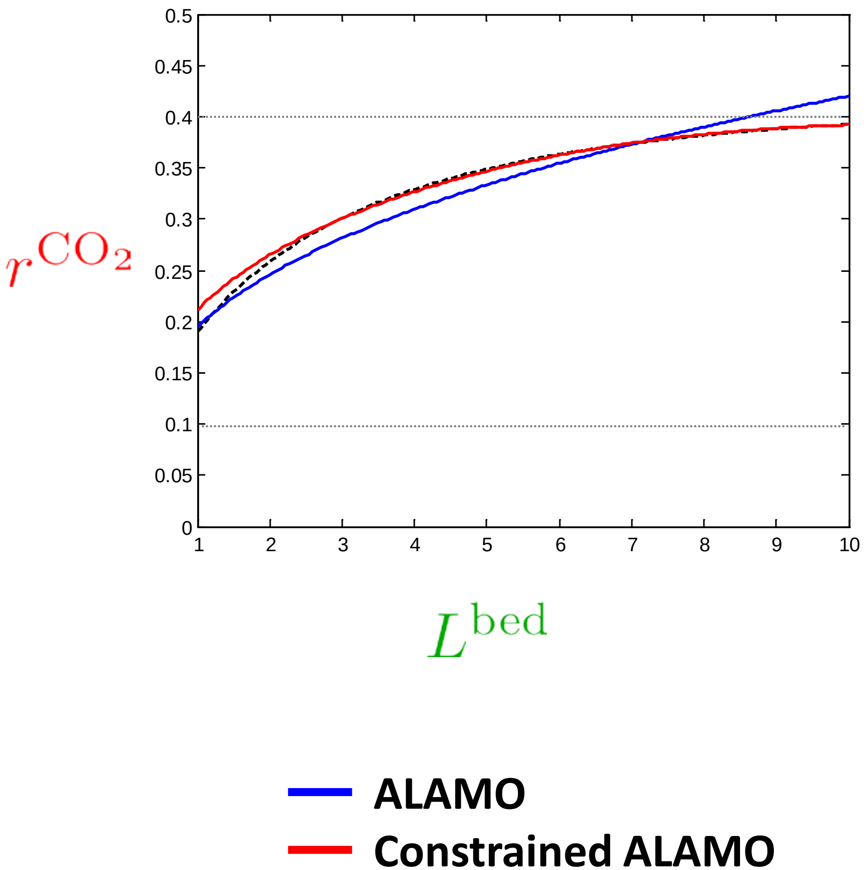
The constrained ALAMO model is able to stay within the bounds

ALAMO

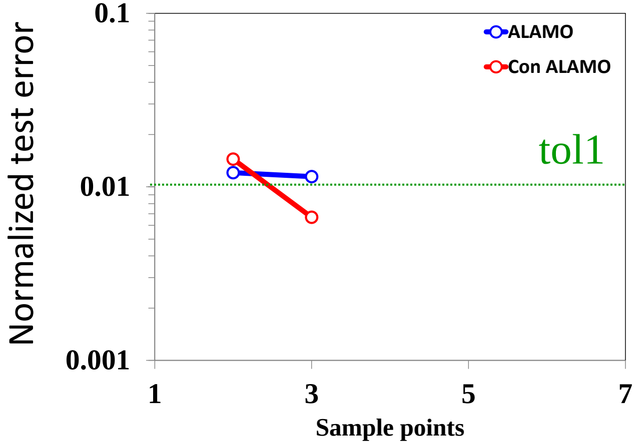
Constrained ALAMO

COMPARISON – ITER 2

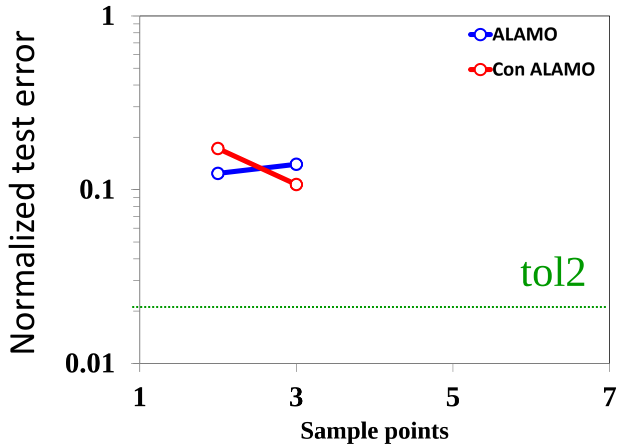
Model comparison



Mean

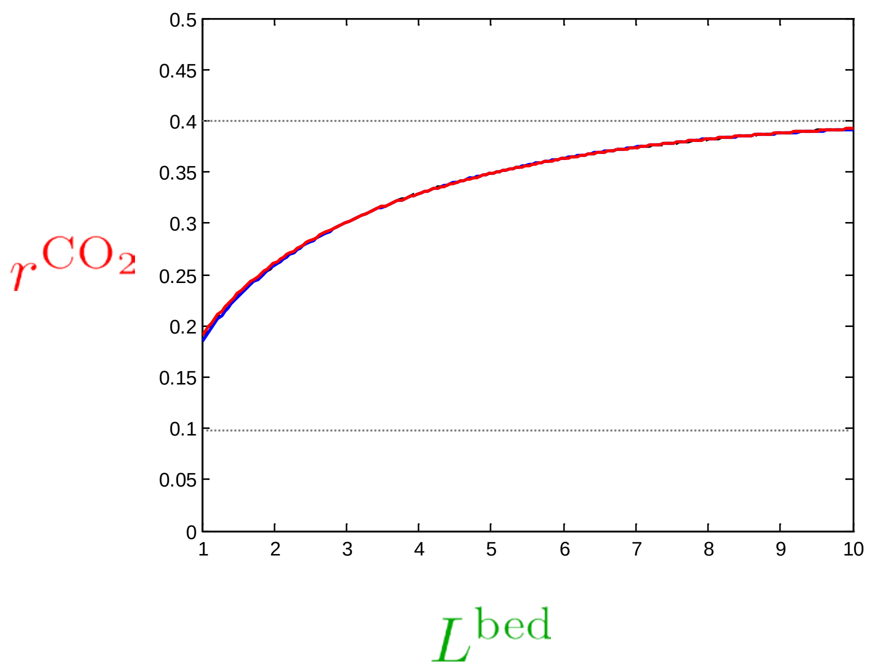


Maximum



COMPARISON – ITER 3

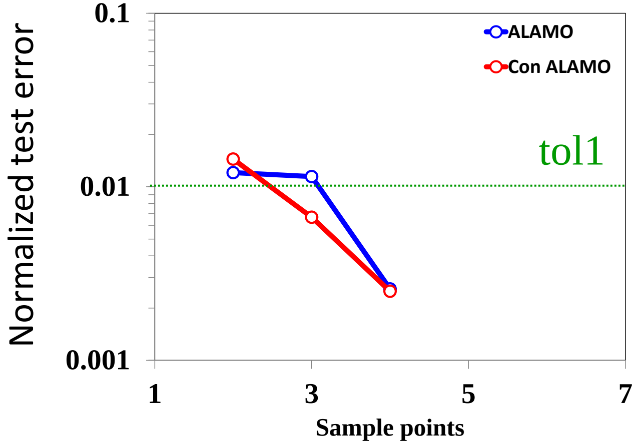
Model comparison



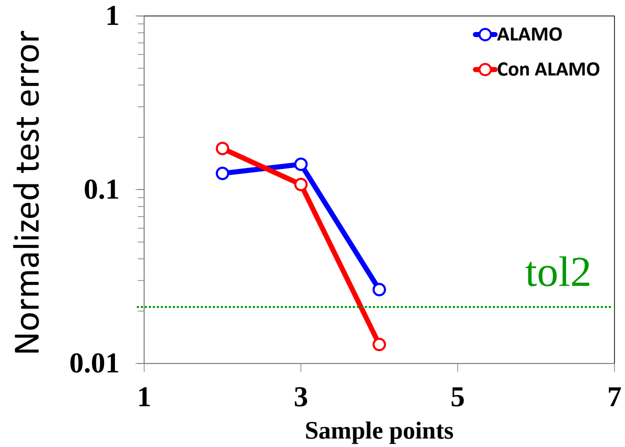
ALAMO

Constrained ALAMO

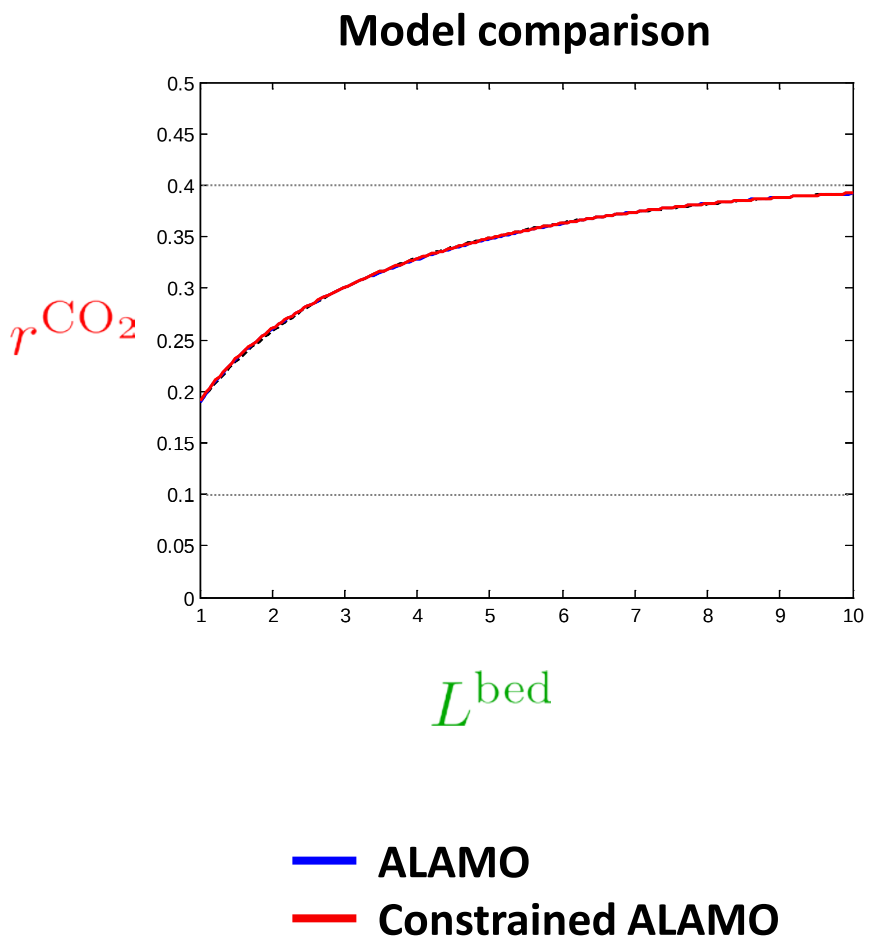
Mean

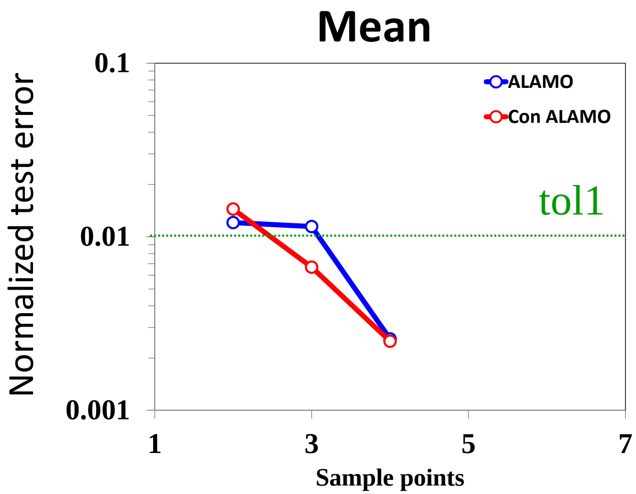


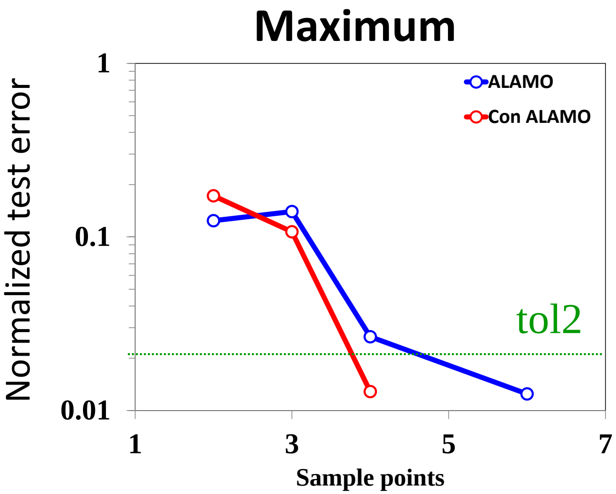
Maximum



COMPARISON – ITER 4

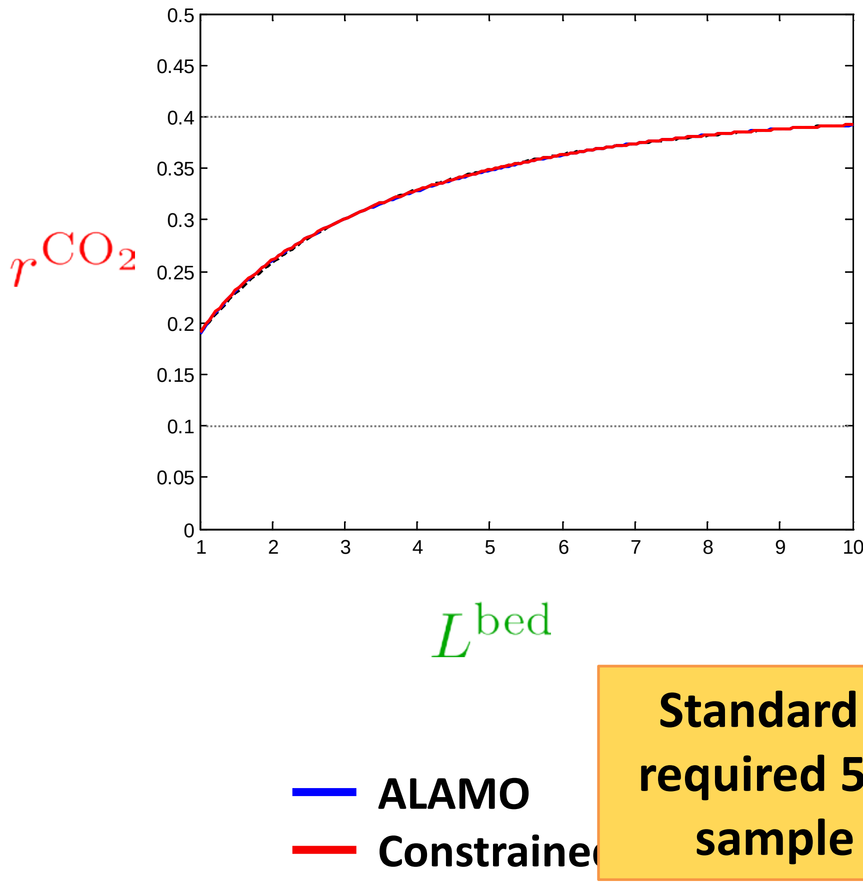




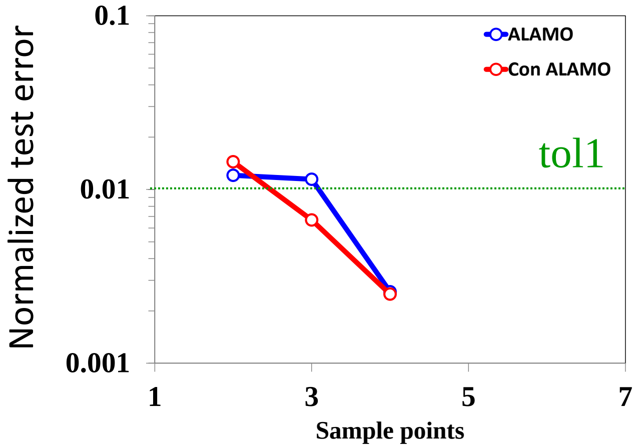


COMPARISON – ITER 4

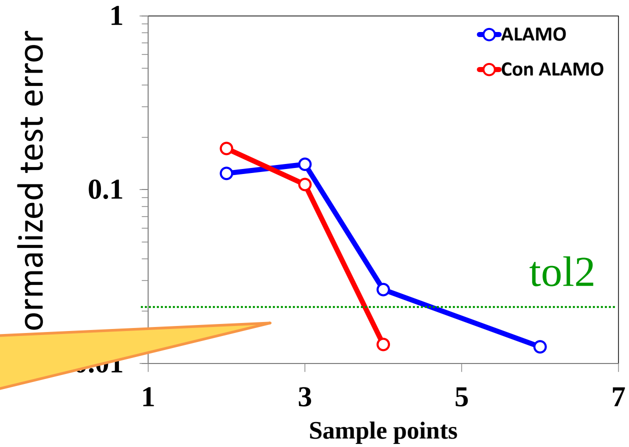
Model comparison



Mean



Maximum



Standard ALAMO required 50% more sample points

CONCLUSIONS

- **Constrained optimization provides a new avenue to provide a model with a priori information**
  - **Include more information in your model without additional sampling**
  - **Reduced the sampling required for an accurate model**

- **Ensure a more robust model by using output bounds as a "reality" check on the model**

- **Future work,**
  - **More complex solution manifolds**
    - ***Nonlinear constraints on regressors***
    - ***Nonlinear feasible domain for output variables***
  - **Simultaneous model generation and constraints**
    - ***Restrictions implied by constraints on multiple outputs***
      - **Ex: Sum-to-one constraints**

STANDARD BASIS FUNCTION SELECTION

$$\min \quad SE = \sum_{i=1}^{N} \left| z_i - \sum_{j \in \mathcal{B}} \beta_j X_{ij} \right|$$

Find the model with the least error

$$\text{s.t.} \quad \sum_{j \in \mathcal{B}} y_j = T$$

$$-U(1 - y_j) \leq \sum_{i=1}^{N} X_{ij} \left(z^i - \sum_{j \in \mathcal{B}} \beta_j X_{ij} \right) \leq U(1 - y_j) \qquad j \in \mathcal{B}$$

$$\beta^l y_j \leq \beta_j \leq \beta^u y_j \qquad j \in \mathcal{B}$$

$$y_j = \{0, 1\} \qquad j \in \mathcal{B}$$

BASIS FUNCTION SELECTION

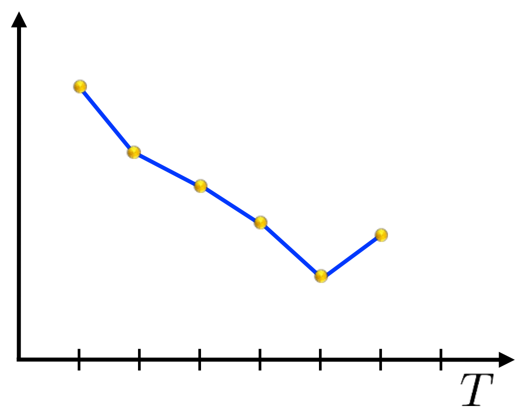
$$\min \quad SE = \sum_{i=1}^{N} \left| z_i - \sum_{j \in \mathcal{B}} \beta_j X_{ij} \right|$$

$$\text{s.t.} \quad \sum_{j \in \mathcal{B}} y_j = T$$

$$-U(1 - y_j) \leq \sum_{i=1}^{N} X_{ij} \left(z^i - \sum_{j \in \mathcal{B}} \beta_j X_{ij} \right) \leq U(1 - y_j) \qquad j \in \mathcal{B}$$

$$\beta^l y_j \leq \beta_j \leq \beta^u y_j$$

$$y_j = \{0, 1\}$$



We will solve this model for increasing T until we determine a model

BASIS FUNCTION SELECTION

$$\min \quad SE = \sum_{i=1}^{N} \left| z_i - \sum_{j \in \mathcal{B}} \beta_j X_{ij} \right|$$

$$\text{s.t.} \quad \sum_{j \in \mathcal{B}} y_j = T$$

$$-U(1-y_j) \leq \sum_{i=1}^{N} X_{ij} \left(z^i - \sum_{j \in \mathcal{B}} \beta_j X_{ij} \right) \leq U(1-y_j) \qquad j \in \mathcal{B}$$

$$\beta^l y_j \leq \beta_j \leq \beta^u y_j \qquad j \in \mathcal{B}$$

$$y_j = \{0, 1\} \qquad j \in \mathcal{B}$$

$$\left(\begin{array}{c} \mathbf{y_j = 1} \\ \\ \textbf{Basis function used in the model} \\ \beta_j \text{ is chosen to satisfy a least} \\ \text{squares regression} \\ (\textit{assumes loose bounds on } \beta_j) \end{array} \right) \bigvee \left(\begin{array}{c} \mathbf{y_j = 0} \\ \\ \textbf{Basis function NOT used} \\ \textbf{in the model} \\ \\ \beta_j = 0 \end{array} \right)$$