Model management strategy for Hierarchical Kriging

Motivation

Many-query analysis

- Requires repeated runs of simulations
- Becomes computationally intractable when using expensive high fidelity simulations

Q1. Can we introduce cheaper low fidelity data into model training?

Q2. “How” should we allocate the samples across fidelities?

Vision

Explore budget allocation strategy for hieararchical Kriging

Building blocks

1. Hierarchical Kriging: Lack of budget allocation strategy
2. Multifidelity Monte Carlo (MFMC): Budget allocation strategy for multifidelity data

Goal

Apply MFMC budget allocation for hierarchical Kriging

Contents

Methods

- Gaussian Process (GP)
- Hierarchical Kriging
- Multifidelity Monte Carlo (MFMC) budget allocation

Results

- Ishigami function example
- Wing structural analysis problem

Gaussian Process: Notation

- $z_i \in \mathbb{R}^d$: high fidelity input
- $z^* \in \mathbb{R}^d$: test input
- $f^{(1)}: \mathbb{R}^d \to \mathbb{R}$: high fidelity model
- $k: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$: kernel covariance function

Gaussian Process: Assumptions

Models input-output relationship
by assuming a Gaussian process prior

$$f^{(1)} \sim \mathcal{GP}(0, k(\cdot,\cdot))$$

$$y_i^{(1)} = f^{(1)}(z_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_e^2)$$

$y_i^{(1)} \in \mathbb{R}$: high fidelity observation

$\varepsilon_i \in \mathbb{R}$: noise

$\sigma_e^2 \in \mathbb{R}$: noise variance

Gaussian Process: Definition

Given training data $\mathcal{D} = \{\boldsymbol{z}, \boldsymbol{y}^{(1)}\}$, GP posterior distribution

$$f^{(1)}(z^*|\mathcal{D}) \sim \mathcal{N}\left(\mathbb{E}\left[f^{(1)}(z^*|\mathcal{D})\right], \mathbb{V}\mathrm{ar}\left[f^{(1)}(z^*|\mathcal{D})\right]\right)$$

Predictor (posterior mean) $\hat{f}^{(1)}(z^*)$

$$= \mathbb{E}\left[f^{(1)}(z^*|\mathcal{D})\right] = k(\boldsymbol{z}, z^*; \theta)^\top \left(k(\boldsymbol{z}, \boldsymbol{z}; \theta) + \sigma_e^2 \boldsymbol{I}\right)^{-1} \boldsymbol{y}^{(1)}$$

Squared exponential kernel

$$k(z, z'; \boldsymbol{\theta}) = \theta_1 \exp\left(-\frac{\|z - z'\|_2^2}{2\theta_2^2}\right)$$

θ_1, θ_2: kernel hyperparameters

How to find $\boldsymbol{\theta}$, σ_e^2?

Gaussian Process: Training

Find $\boldsymbol{\theta}$, σ_e^2 that maximize log-likelihood
by gradient-based optimization methods

$$\max_{\theta,\sigma_e^2} \log p\left(\boldsymbol{y}^{(1)} \middle| \boldsymbol{\theta}, \sigma_e^2\right) = -\frac{1}{2}\left[\boldsymbol{y}^{(1)\top}\left(k(\boldsymbol{z},\boldsymbol{z};\boldsymbol{\theta}) + \sigma_e^2\boldsymbol{I}\right)^{-1}\boldsymbol{y}^{(1)}\right] + \log\left|k(\boldsymbol{z},\boldsymbol{z};\boldsymbol{\theta}) + \sigma_e^2\boldsymbol{I}\right|$$

Recall that the predictor is dependent on dataset

$$\hat{f}^{(1)}(z^*) = \mathbb{E}\left[f^{(1)}(z^* | \mathcal{D})\right]$$

If $|D|$ is small, $\mathbb{V}\mathrm{ar}\left[\hat{f}^{(1)}(z^*)\right]$ becomes large

Can we leverage abundant lower fidelity data?

Hierarchical Kriging: Notation

- $z_i^{(2)} \in \mathbb{R}^d$: low fidelity input
- $f^{(2)}: \mathbb{R}^d \to \mathbb{R}$: low fidelity model
- $y_i^{(2)} \in \mathbb{R}$: low fidelity observation
- $\boldsymbol{y}^{(1)} \in \mathbb{R}^{n}$: high fidelity output vector
- $\boldsymbol{y}^{(2)} \in \mathbb{R}^{m}$: low fidelity output vector

Typically $n < m$

Hierarchical Kriging: Definition

Based on Kennedy O'Hagan approach

$$f^{(1)}(z^*) = \alpha f^{(2)}(z^*) + \delta(z^*)$$

$\alpha \in \mathbb{R}$: scaling factor

- Assumes $f^{(2)}$ is a low fidelity GP model
- δ is a discrepancy GP model

Predictor (posterior mean)

$$\hat{f}^{(1)}(z^*) = \alpha \hat{f}^{(\mathbf{2})}(z^*) + \hat{\delta}(z^*)$$

$\hat{f}^{(\mathbf{2})}(z^*)$: posterior mean of $f^{(2)}$

$\delta(z^*)$: posterior mean of δ

Hierarchical Kriging: Definition

Predictor (posterior mean)

$$\hat{f}^{(1)}(z^*) = \alpha \hat{f}^{(2)}(z^*) + \hat{\delta}(z^*)$$

$\hat{f}^{(2)}(z^*)$: posterior mean of $f^{(2)}$ trained with $y^{(2)} \in \mathbb{R}^m$

$\hat{\delta}(z^*)$: posterior mean of δ trained with discrepancy data

$$\boldsymbol{y}_d = \boldsymbol{y}^{(1)} - \alpha \hat{f}^{(2)}(\boldsymbol{z}), \quad y_d \in \mathbb{R}^n$$

$$\begin{aligned}
&\hat{f}^{(1)}(z^*) \\
&= \alpha \left(k\left(\boldsymbol{z}^{(2)}, z^*; \theta\right)^\top \left(k\left(\boldsymbol{z}^{(2)}, \boldsymbol{z}^{(2)}; \theta\right) + \sigma_e^2 \boldsymbol{I} \right)^{-1} \boldsymbol{y}^{(2)} \right) \\
&+ k(\boldsymbol{z}, z^*; \theta)^\top (k(\boldsymbol{z}, \boldsymbol{z}; \theta) + \sigma_e^2 \boldsymbol{I})^{-1} \boldsymbol{y}^{(1)}
\end{aligned}$$

How to find $\boldsymbol{\theta}$, σ_e^2 for δ?

Hierarchical Kriging: Training

Find $\boldsymbol{\theta}$, σ_e^2 that maximize log-likelihood
by gradient-based optimization methods

$$\log p(\boldsymbol{y}_d|\boldsymbol{\theta}, \sigma_e^2) =$$

$$\max_{\theta, \sigma_e^2} -\frac{1}{2}\left[\boldsymbol{y}_d^\top (k(\boldsymbol{z}, \boldsymbol{z}; \boldsymbol{\theta}) + \sigma_e^2 \boldsymbol{I})^{-1} \boldsymbol{y}_d\right] + \log|k(\boldsymbol{z}, \boldsymbol{z}; \boldsymbol{\theta}) + \sigma_e^2 \boldsymbol{I}|$$

Recall $\boldsymbol{y}_d = \boldsymbol{y}^{(1)} - \alpha \hat{f}^{(2)}(\boldsymbol{z})$

How to find α?

Hierarchical Kriging: scaling factor

By solving generalized least squares

$$\left\|\boldsymbol{y}^{(1)} - \alpha \hat{f}^{(2)}(\boldsymbol{z})\right\|^2_{k(\boldsymbol{z},\boldsymbol{z};\boldsymbol{\theta})+\sigma_e^2\boldsymbol{I}}$$

$$\alpha = \left(\hat{f}^{(2)}(\boldsymbol{z})^\top\left(k(\boldsymbol{z},\boldsymbol{z};\boldsymbol{\theta}) + \sigma_e^2\boldsymbol{I}\right)\hat{f}^{(2)}(\boldsymbol{z})\right)^{-1} \left(\hat{f}^{(2)}(\boldsymbol{z})^\top\left(k(\boldsymbol{z},\boldsymbol{z};\boldsymbol{\theta}) + \sigma_e^2\boldsymbol{I}\right)^{-1}\boldsymbol{y}^{(1)}\right)$$

Predictor $\hat{f}^{(1)}(z^*)$

$$= \alpha\left(k\left(\boldsymbol{z}^{(2)}, z^*; \boldsymbol{\theta}\right)^\top\left(k\left(\boldsymbol{z}^{(2)}, \boldsymbol{z}^{(2)}; \boldsymbol{\theta}\right) + \sigma_e^2\boldsymbol{I}\right)^{-1}\boldsymbol{y}^{(2)}\right) + k(\boldsymbol{z}, z^*; \boldsymbol{\theta})^\top(k(\boldsymbol{z},\boldsymbol{z};\boldsymbol{\theta}) + \sigma_e^2\boldsymbol{I})^{-1}\boldsymbol{y}_d$$

$$\boldsymbol{y}^{(2)} \in \mathbb{R}^m, \boldsymbol{y}_d \in \mathbb{R}^n$$

How to allocate n, m?

MFMC budget allocation: MFMC estimator

More robust way to estimate mean

- Monte Carlo estimator

$$\mathbb{E}\left[f^{(1)}(Z)\right] \approx \frac{1}{n}\sum_{i=1}^{n} y_i^{(1)}$$

$Z \in \mathbb{R}^d$: high fidelity input random variable

- MFMC estimator adds low fidelity data
    - Assumes nested samples $Z \subset Z^{(2)}$

$$\mathbb{E}\left[f^{(1)}(Z)\right] \approx \frac{1}{n}\sum_{i=1}^{n} y_i^{(1)} + \alpha\left(\frac{1}{m}\sum_{i=1}^{m} y_i^{(2)} - \frac{1}{n}\sum_{i=1}^{n} y_i^{(2)}\right)$$

$Z^{(2)} \in \mathbb{R}^d$: low fidelity input random variable

MFMC budget allocation: Variance

Obtain optimal n and m that minimizes variance of the estimator

$$\frac{\sigma_1^2}{n} + \left(\frac{1}{n} - \frac{1}{m}\right)\left(\alpha^2\sigma_2^2 - 2\alpha\rho_{1,2}\sigma_1\sigma_2\right)$$

σ_1: standard deviation of high fidelity data

σ_2: standard deviation of low fidelity data

$\rho_{1,2}$: correlation coefficient of high and low fidelity data

MFMC budget allocation

$$n = \frac{c}{w_1 + w_2 \tau}, \qquad m = \tau n, \qquad \tau = \sqrt{\frac{w_1 \rho_{1,2}^2}{w_2 (1 - \rho_{1,2}^2)}}$$

n: number of high fidelity samples

m: number of low fidelity samples

c: computational budget

w_1: high fidelity model evaluation cost

w_2: low fidelity model evaluation cost

τ: ratio of number of low fidelity samples to high fidelity samples

Connection between MFMC and hierarchical Kriging

Claim: Hierarchical Kriging predictor is analogous to MFMC-based ridge regression predictor

Consider ridge regression problem

$$\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \mathbb{E}\left\|y^{(1)} - \phi(\boldsymbol{z})^\top \boldsymbol{\beta}\right\|_2^2 + \sigma_e^2 \|\boldsymbol{\beta}\|_2^2$$

$\phi: \mathbb{R}^d \to \mathbb{R}^p$: feature map

$\boldsymbol{\beta} \in \mathbb{R}^p$: regression coefficients

$$\boldsymbol{\beta}^* = \left(\mathbb{E}[\phi(\boldsymbol{z})\phi(\boldsymbol{z})^\top] + \sigma_e^2 \boldsymbol{I}\right)^{-1} \mathbb{E}\left[\phi(\boldsymbol{z}) y^{(1)}\right]$$

Apply MFMC to estimate $\mathbb{E}\left[\phi(\boldsymbol{z}) y^{(1)}\right]$

$$\mathbb{E}\left[\phi(\boldsymbol{z}) y^{(1)}\right] \approx \frac{1}{n} \boldsymbol{\Phi}_n \boldsymbol{y}_n^{(1)} + \alpha \left(\frac{1}{m} \boldsymbol{\Phi}_m \boldsymbol{y}_m^{(2)} - \frac{1}{n} \boldsymbol{\Phi}_n \boldsymbol{y}_n^{(2)}\right)$$

$\boldsymbol{\Phi}_n \in \mathbb{R}^{p\times n}$: feature matrix

Connection between MFMC and hierarchical Kriging

Applying Woodbury Identity,

$$\hat{\beta} = \frac{1}{n}\boldsymbol{\Phi}_n\left(\frac{1}{n}k(\boldsymbol{z},\boldsymbol{z};\boldsymbol{\theta}) + \sigma_e^2\boldsymbol{I}\right)^{-1}\boldsymbol{y}_n^{(1)}$$

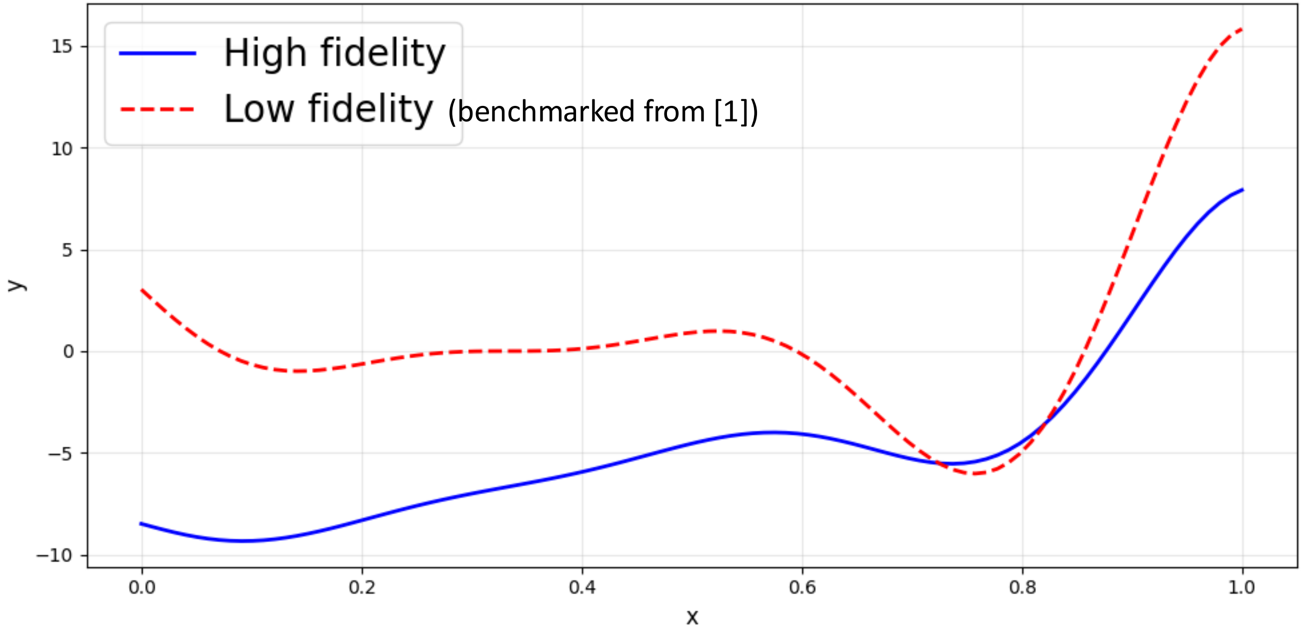
$$+\alpha\left(\frac{1}{m}\boldsymbol{\Phi}_m\left(\frac{1}{m}k\left(\boldsymbol{z}^{(2)},\boldsymbol{z}^{(2)};\boldsymbol{\theta}\right) + \sigma_e^2\boldsymbol{I}\right)^{-1}\boldsymbol{y}_m^{(2)} - \frac{1}{n}\boldsymbol{\Phi}_n\left(\frac{1}{n}k(\boldsymbol{z},\boldsymbol{z};\boldsymbol{\theta}) + \sigma_e^2\boldsymbol{I}\right)^{-1}\boldsymbol{y}_n^{(2)}\right)$$

$$\begin{aligned}\hat{f}^{(1)}(z^*) &= \phi(z^*)^\top\hat{\beta} \\ &= \alpha k\left(\boldsymbol{z}^{(2)},z^*;\boldsymbol{\theta}\right)^\top\left(k\left(\boldsymbol{z}^{(2)},\boldsymbol{z}^{(2)};\boldsymbol{\theta}\right) + m\sigma_e^2\boldsymbol{I}\right)^{-1}\boldsymbol{y}_m^{(2)} \\ &\quad + k(\boldsymbol{z},z^*;\boldsymbol{\theta})^\top\left(k(\boldsymbol{z},\boldsymbol{z};\boldsymbol{\theta}) + n\sigma_e^2\boldsymbol{I}\right)^{-1}\left(\boldsymbol{y}_n^{(1)} - \alpha\hat{f}^{(2)}(\boldsymbol{z})\right)\end{aligned}$$

Analogous to hierarchical Kriging predictor, except that regularization parameters are scaled by n and m

Forrester function example: Set up

- Input: $Z \sim \mathcal{U}(0,1)$

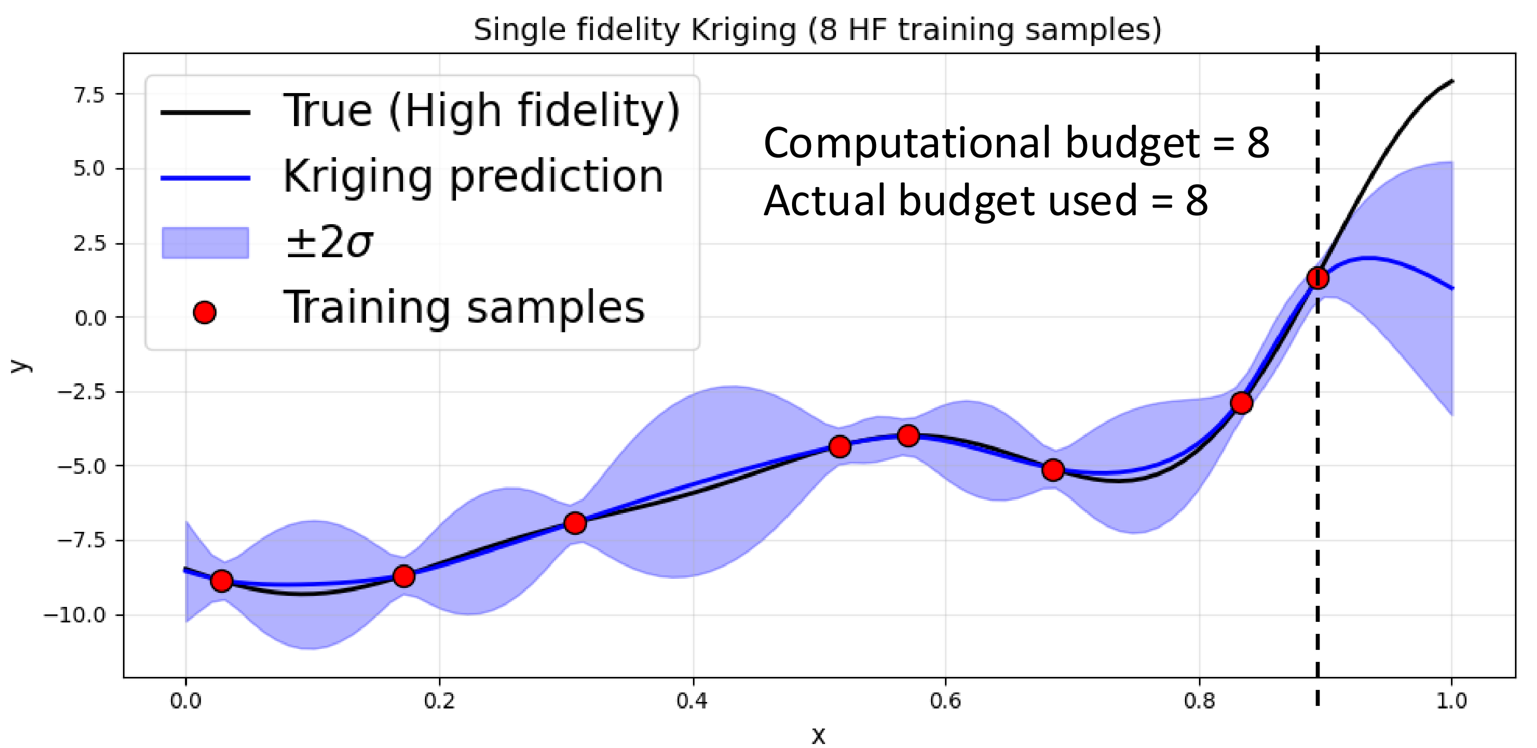


[1] Meng, X., & Karniadakis, G. E. (2020). A composite neural network that learns from multi-fidelity data: Application to function approximation and inverse PDE problems. Journal of Computational Physics, 401, 109020.

- Cost = [1, 0.1]
- Statistics: $\sigma_1 = 4.05$, σ_1 =4.34, $\rho_{1,2} = 0.7332$

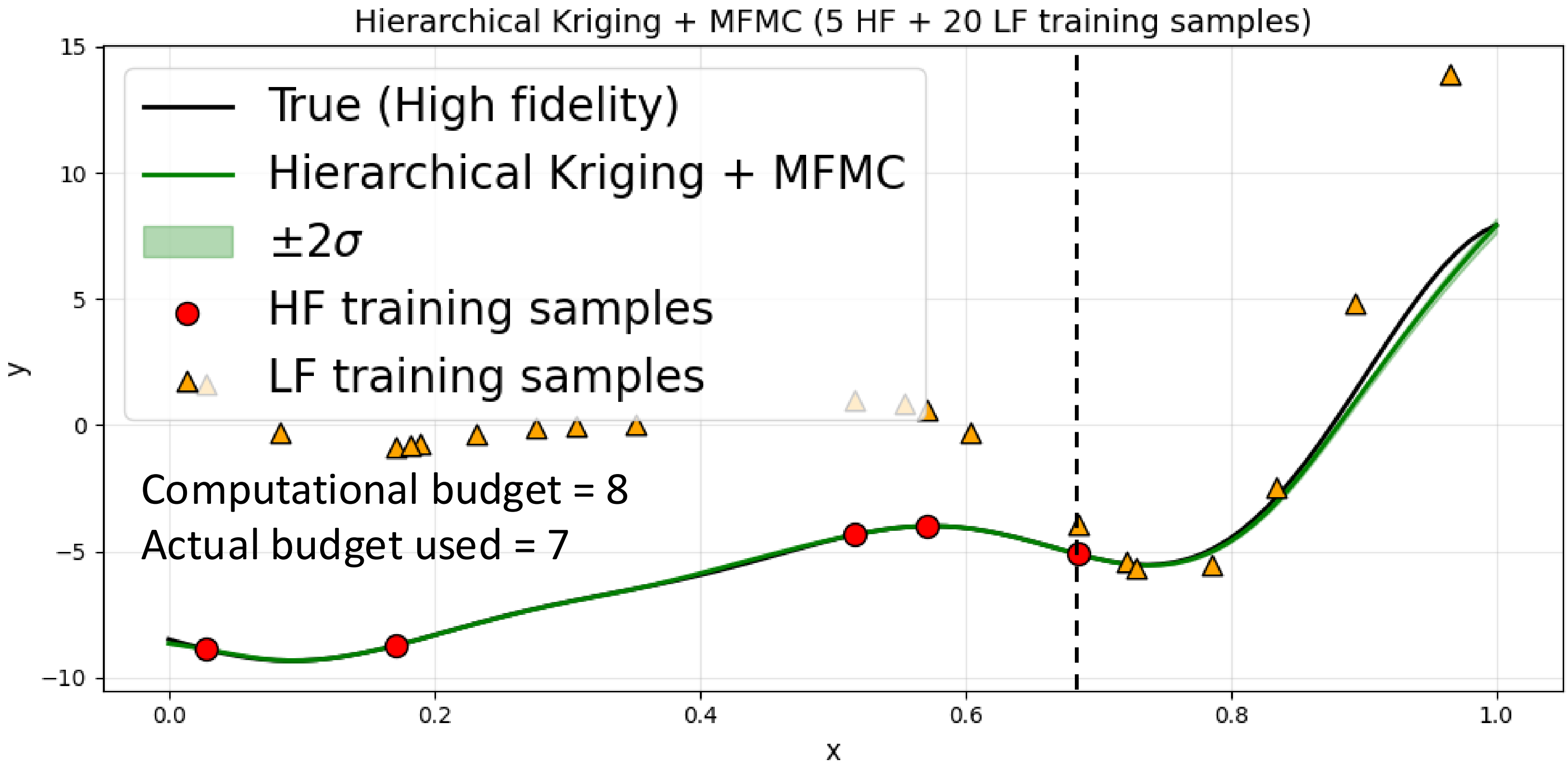
Forrester function example: Results

- No high fidelity samples available at $x > 0.9$
- Single fidelity fails to capture trend after $x > 0.9$



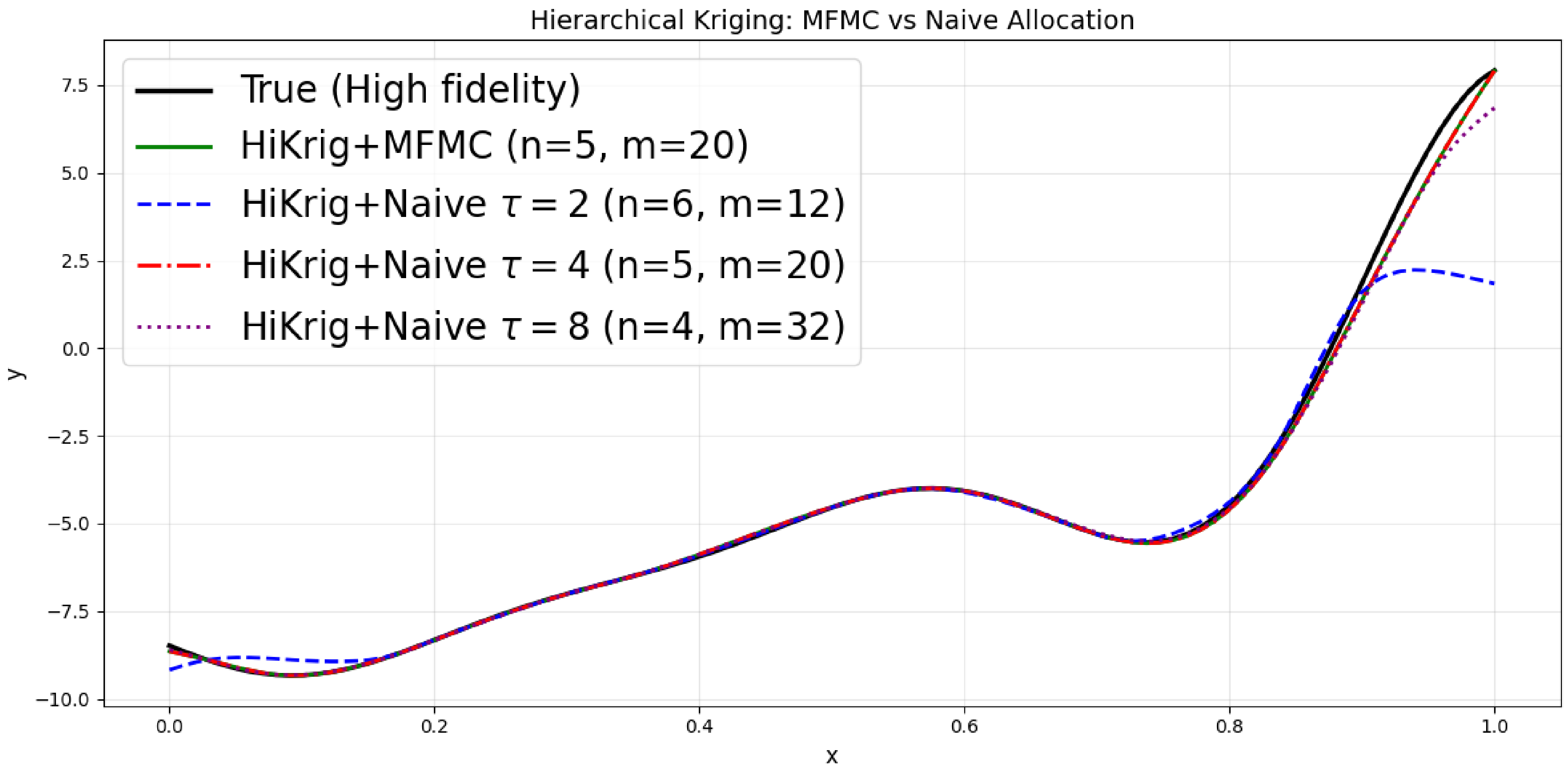
Forrester function example: Results

- Accurately predicts overall trend even in regions that lack high fidelity samples



Forrester function example: Results

- Hierarchical Kriging using MFMC allocation achieves comparable accuracy with Naïve allocation with $\tau = 4$ and higher accuracy compared to other naïve allocations



Ishigami function example: Set up

- Input: $Z = (Z_1, Z_2, Z_3), \quad Z_i \sim \mathcal{U}(-\pi, \pi)$
- High fidelity model

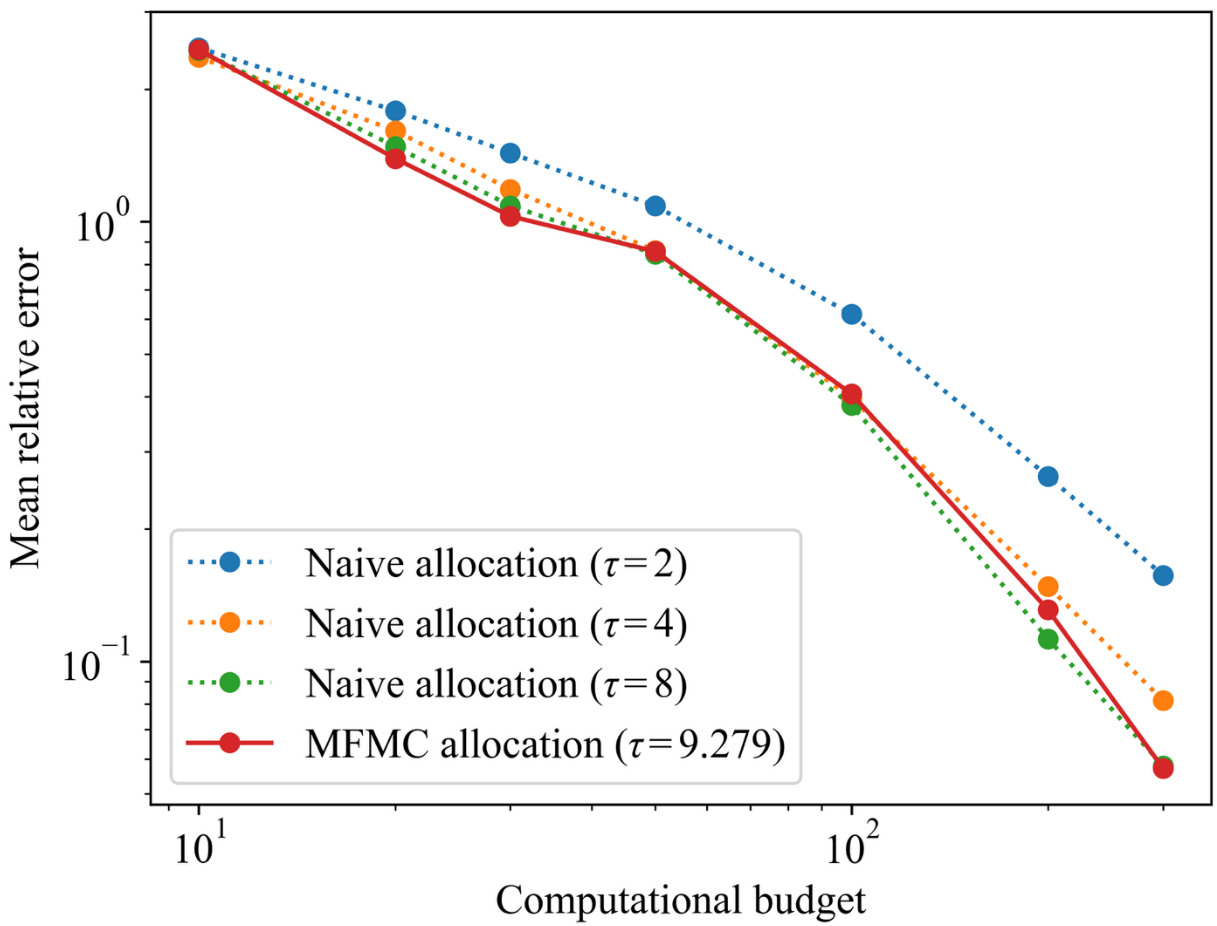
$$f^{(1)}(Z) = \sin Z_1 + 5\sin^2 Z_2 \; + 0.1 Z_3^4 \sin Z_1$$

- Low fidelity model

$$f^{(2)}(Z) = \sin Z_1 + 3\sin^2 Z_2 \; + 0.9 Z_3^2 \sin Z_1$$

- Cost = [1, 0.1]
- Statistics: $\sigma_1 = 3.29$, $\sigma_1 = 3.53$, $\rho_{1,2} = 0.9465$

Ishigami function example: Results



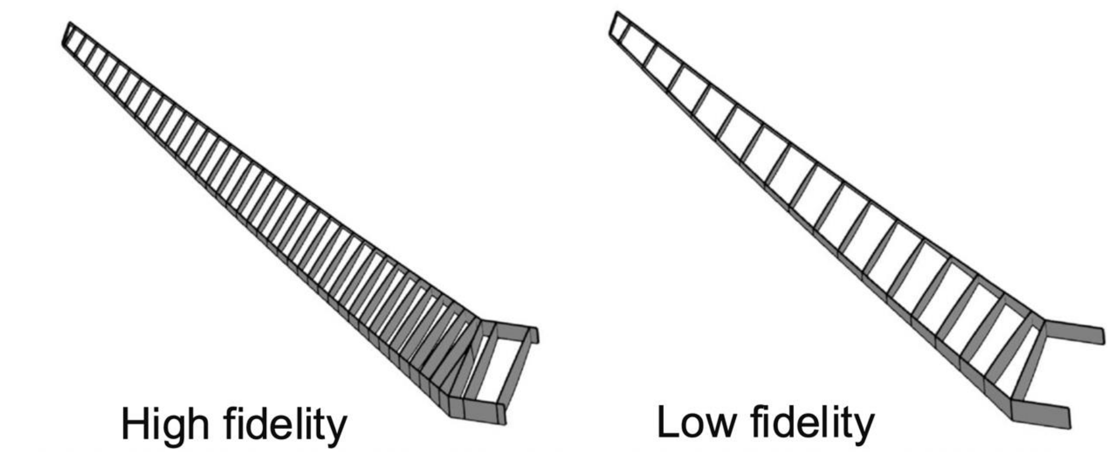
At budget = 100

Allocation	τ	n	m
	2	83	166
Naive	4	71	285
	8	55	444
MFMC	9.279	51	481

$$\tau = \frac{m}{n}$$

Wing structural analysis problem : Set up

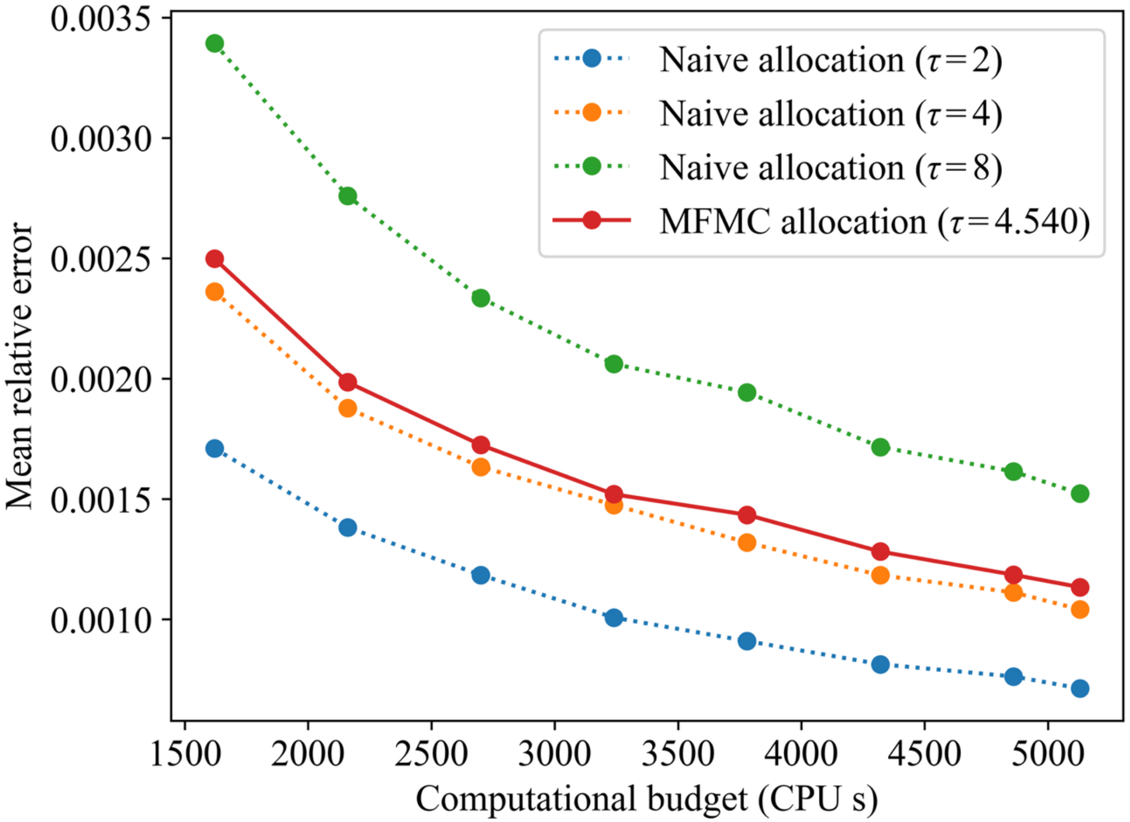
- Input: 4 wing geometry parameters
  - Wing span, dihedral, twist, sweep angles
- Output: maximum von Mises stress



Source: Perron, C., Rajaram, D., & Mavris, D. N. (2021). Multi-fidelity non-intrusive reduced-order modelling based on manifold alignment. Proceedings of the Royal Society A, 477(2253), 20210495

- Cost = [5.4, 4.7] CPU s
- Statistics: $\sigma_1 = 9131.61, \sigma_1 = 8838.04, \rho_{1,2} = 0.9732$

Wing structural analysis problem: Results



At budget = 1,620 CPU s

Allocation	τ	n	m
Naive	2	109	218
	4	66	267
	8	37	301
MFMC	4.54	60	275

$$\tau = \frac{m}{n}$$

Summary and conclusion

- Proposed MFMC budget allocation strategy for hierarchical Kriging
- Hierarchical Kriging with MFMC allocation achieves comparable accuracy
- MFMC allocation functions as a practical guideline for sample allocation