DaFT: DerivAtive-Free Thinning

with Stochastic Kernel Embeddings [WiP]

Nikolaos Papadimas, **Mikkel Lykkegaard** and Tim Dodwell

May 5, 2022

Data-Centric Engineering Group, University of Exeter, m.lykkegaard@exeter.ac.uk

Table of contents

1. Motivation: Uncertainty Quantification

2. DaFT Theory

3. Results and a Synthetic Example

4. Future Directions and Discussion

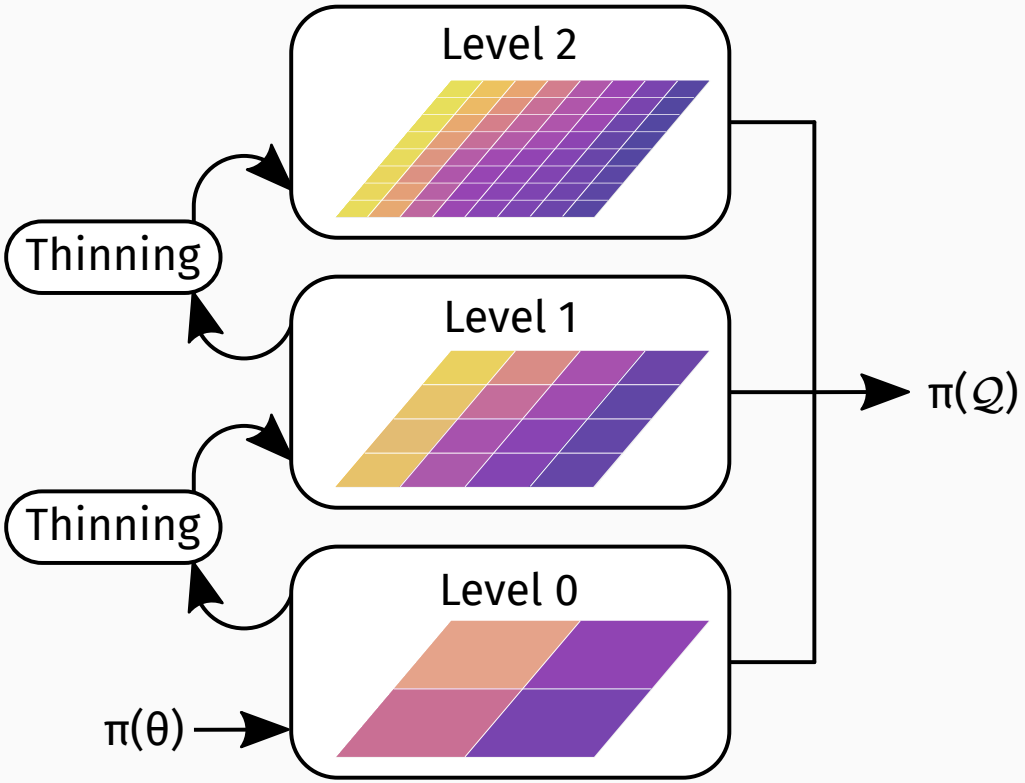
Motivation: Uncertainty Quantification

An Epistomological Hiccup

"All models are wrong but some models are more wrong than others"

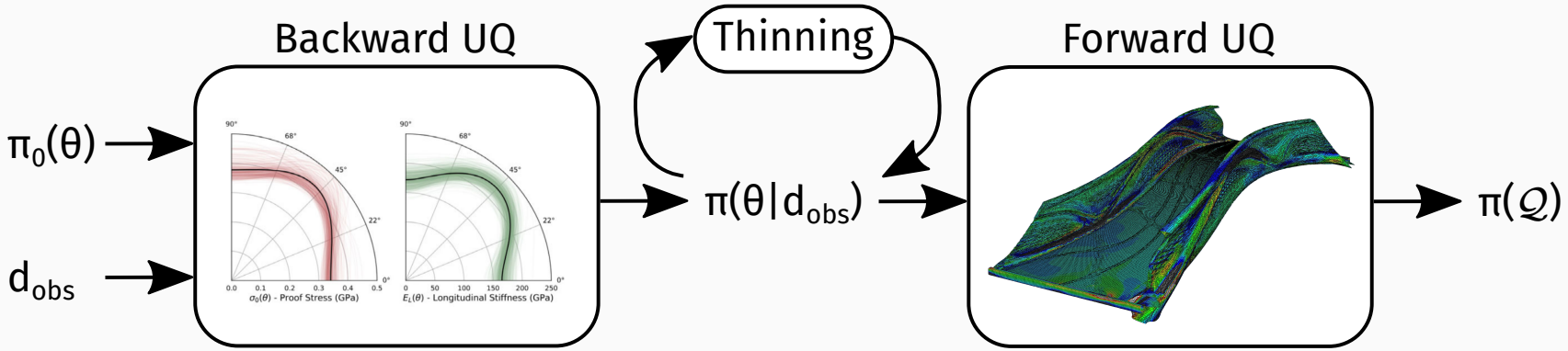
–George Cox/Orwell

Thinning for Multilevel Monte Carlo



- **Which samples** to pick for higher levels?
- Simple random sampling of lower–level samples may lead to **unquantifiable bias**.
- Need **optimal empirical approximation** of the parameter distribution.

Thinning for Multi-Fidelity Models



- Low–fidelity model is **relatively cheap** and allows for running MCMC.
- High–fidelity model is **very expensive** and can only be solved a few times.
- Need **highly compressed** approximation of the posterior.

Image sources: Papadimas and Dodwell (2021), The Alan Turing Institute

Stein Thinning

The Devil is in the Detail

Method proposed by Riabiz et al. (2020) achieves **optimal thinning**, but their kernel requires the **gradient** of the posterior distribution:

$$k_P(x, y) \coloneqq \nabla_x \cdot \nabla_y k(x, y) + \langle \nabla_x k(x, y), \nabla_y \log p(y) \rangle + \langle \nabla_y k(x, y), \nabla_x \log p(x) \rangle + k(x, y) \langle \nabla_x \log p(x), \nabla_y \log p(y) \rangle$$

- The gradient **not readily available** in the context of many PDE–constrained problems.
- Potentially **very large kernel matrix** for MCMC samples, if all $k(x, y)$ are evaluated.

$\Rightarrow$ Develop a **simple**, **fast** and **gradient–free** method for optimal thinning of MCMC output.

DaFT Theory

Maximum Mean Discrepancy

- Need **measure of distance** between empirical distributions.
- **MMD** can be defined by a feature map $\phi : \mathcal{X} \to \mathcal{H}$ where $\mathcal{H}$ is a RKHS:

$$\mathrm{MMD}(P, Q) := \|\mathbb{E}_{X\sim P}[\phi(X)] - \mathbb{E}_{Y\sim Q}[\phi(Y)]\|_{\mathcal{H}} \tag{1}$$

- Using the **"kernel trick"** $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$ we get

$$\begin{aligned}\mathrm{MMD}^2(P, Q) &= \|\mathbb{E}_{X\sim P}[\phi(X)] - \mathbb{E}_{Y\sim Q}[\phi(Y)]\|^2_{\mathcal{H}} \\ &= \mathbb{E}_{X,X'\sim P}[k(X, X')] + \mathbb{E}_{Y,Y'\sim Q}[k(Y, Y')] \\ &\quad - 2\,\mathbb{E}_{X\sim P,Y\sim Q}[k(X, Y)]\end{aligned}$$

- But evaluating the kernel directly is costly and very memory intensive for (typically) large samples sizes.

Random Kernel Embeddings

- Following Rahimi and Recht (2007), the data is pushed through a **random feature map** $z : \mathbb{R}^d \to \mathbb{R}^D$ to a low–dimensional Euclidean inner product space, where the inner product **approximates the kernel**:

$$k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} \approx z(x)^T z(y). \tag{2}$$

- Some popular shift–invariant kernels can be approximated by

$$z(x) = \sqrt{\frac{2}{D}} \left[\cos(\omega_1^T x + b_1), \ldots, \cos(\omega_D^T x + b_D) \right]^T. \tag{3}$$

  For example, for a Gaussian kernel we use $\omega \sim \mathcal{N}(0, 1) \in \mathbb{R}^d$ and $b \sim \mathcal{U}(0, 2\pi)$.

An approximate MMD

- Putting all this together, we get an **approximate MMD**:

$$\mathrm{MMD}(P, Q) \approx \left\| \frac{1}{n} \sum_{i=1}^{n} z(x_i) - \frac{1}{m} \sum_{j=1}^{m} z(y_j) \right\|_2 \tag{4}$$

  where $X = \{x_i\}_{i=1}^{n} \sim P$ is the entire MCMC sampling history and $Y = \{y_j\}_{j=1}^{m} \sim Q$ is a subset of the MCMC sampling history.

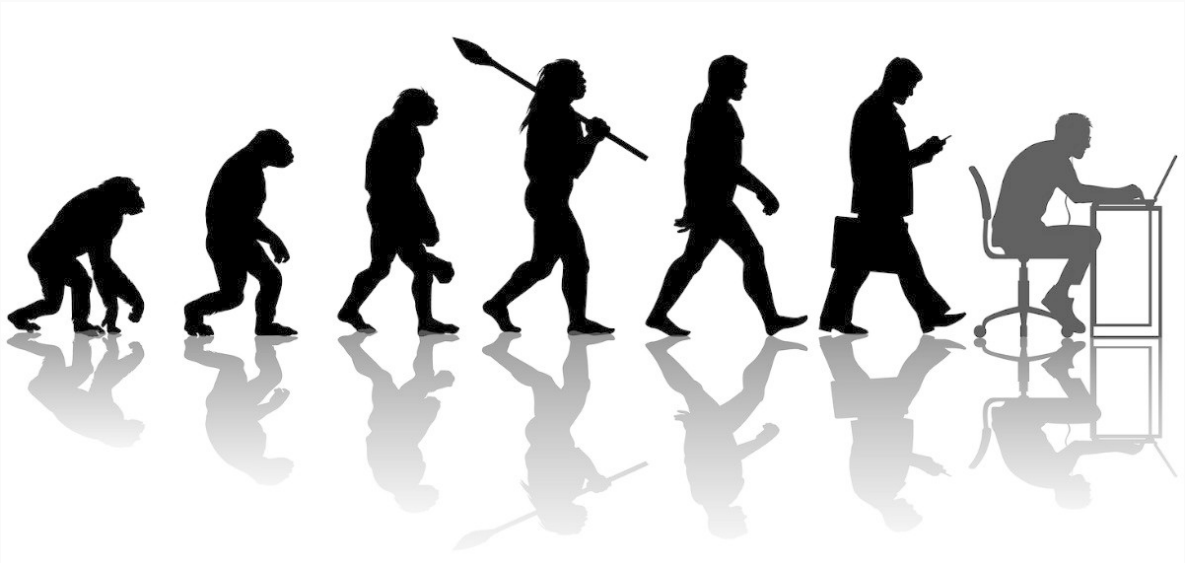
- Given some method for assigning weights $\{w_j\}_{j=1}^{m}$ to the subset $\{y_j\}_{j=1}^{m}$, we can also construct an approximate **weighted** MMD:

$$\mathrm{MMD}_w(P, Q) \approx \left\| \frac{1}{n} \sum_{i=1}^{n} z(x_i) - \frac{\sum_{j=1}^{m} w_j z(y_j)}{\sum_{j=1}^{m} w_j} \right\|_2 \tag{5}$$

Minimising the MMD

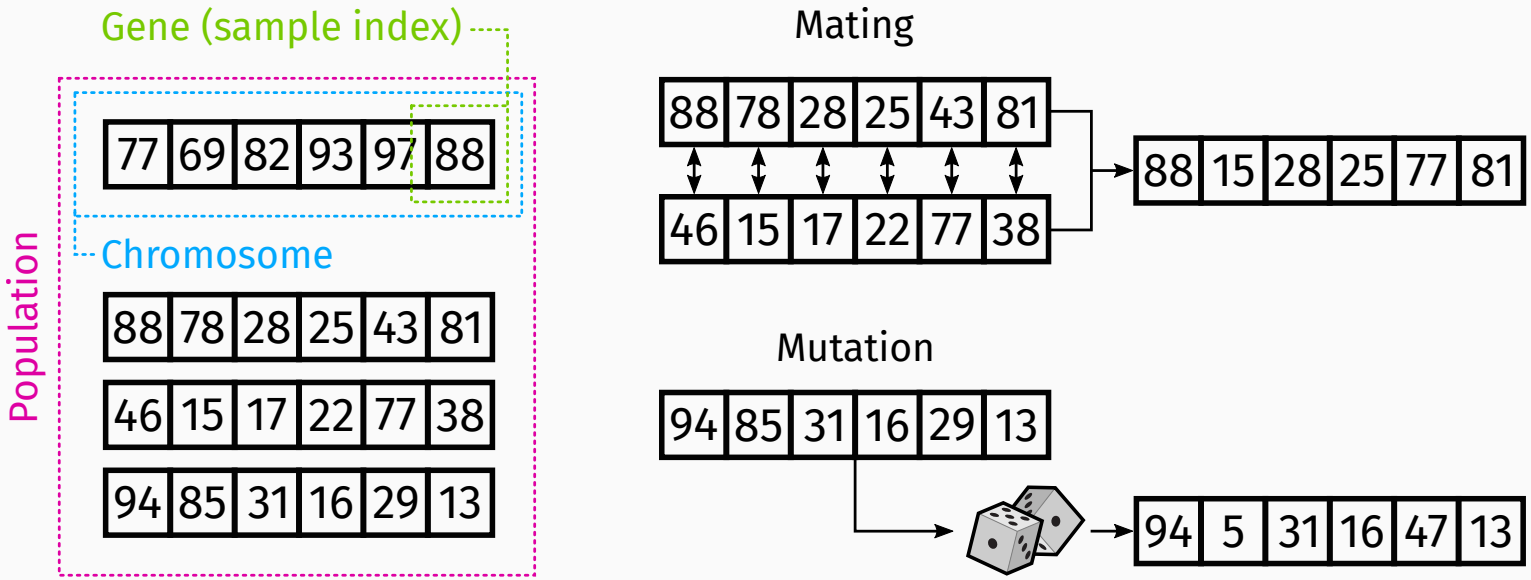
$$Y^\star = \arg\min_{Y \in X} \mathrm{MMD}_X(Y) \tag{6}$$

- **Discrete optimisation problem** with $n!/(n-m)!$ possible states!
- Possibly many **near–optimal solutions**.
- **Genetic Algorithm** heuristically yields a population of solutions.
- Not rigorous but it **works**.



Genetic Algorithm

- Population of candidates $\mathcal{Y} = \{Y_i\}_{i=1}^{N}$ with size N.
- Fitness function $f(Y_i) = \mathrm{MMD}_X(Y_i)^{-1}$.
- Highest fitness chromosomes are preserved across generations.
- Mating/mutation selection probabilities $p_s(Y_i) = \frac{f(Y_i)}{\sum_{i=1}^{N} f(Y_i)}$.



Results and a Synthetic Example

Convergence of the approximate MMD

Two samples from a banana distribution.

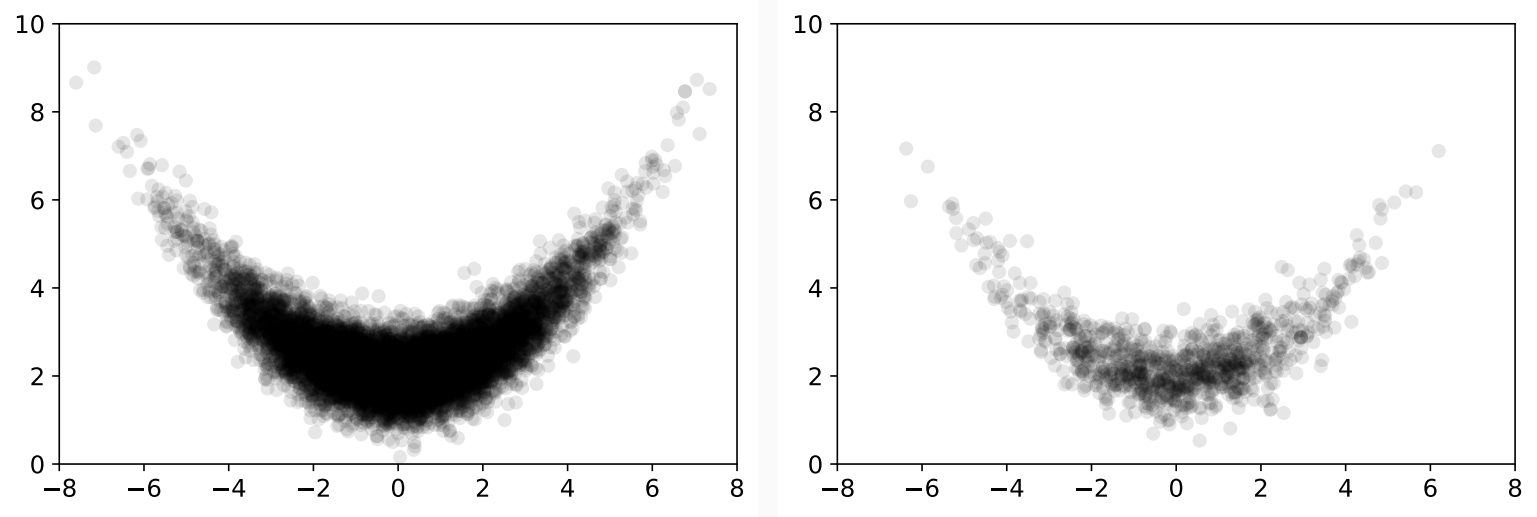


Figure 1: Two samples from a banana–shaped distribution, with (left) $n = 10000$ and (right) $m = 1000$ datapoints.

Convergence of the approximate MMD

Some numerical evidence of convergence.

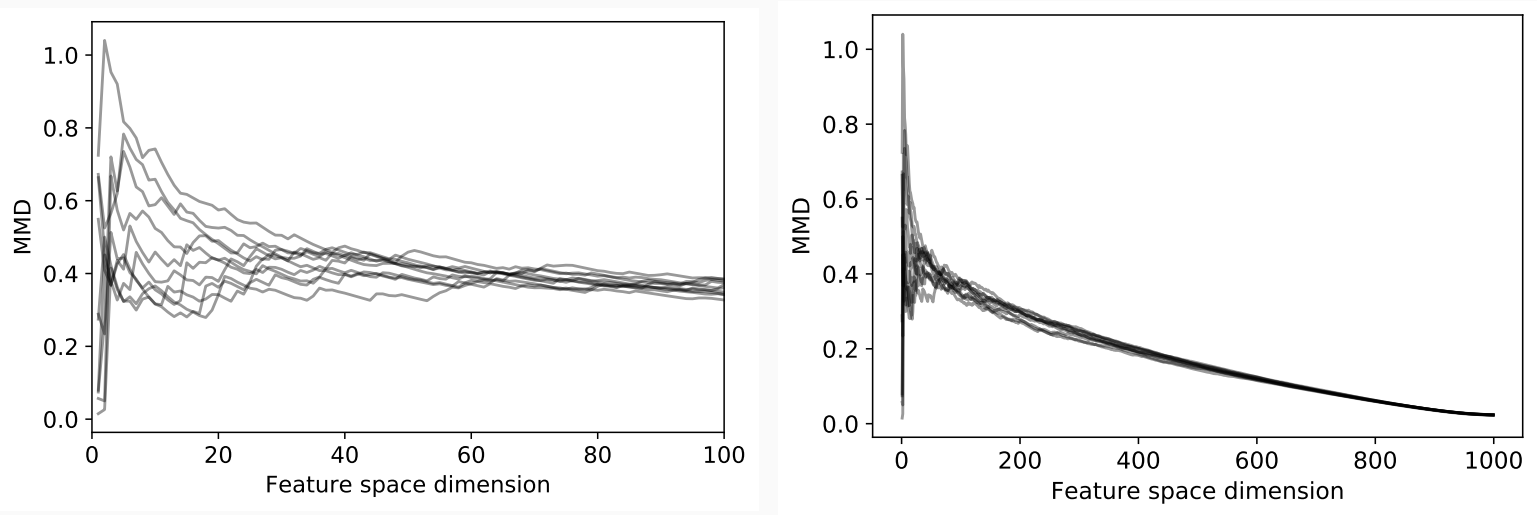


Figure 2: Convergence of the approximate MMD with increasing feature space dimension (D) using the samples shown in Figure 1. Each line represents a random feature space, each constructed as described above.

Thimomenos Distribution

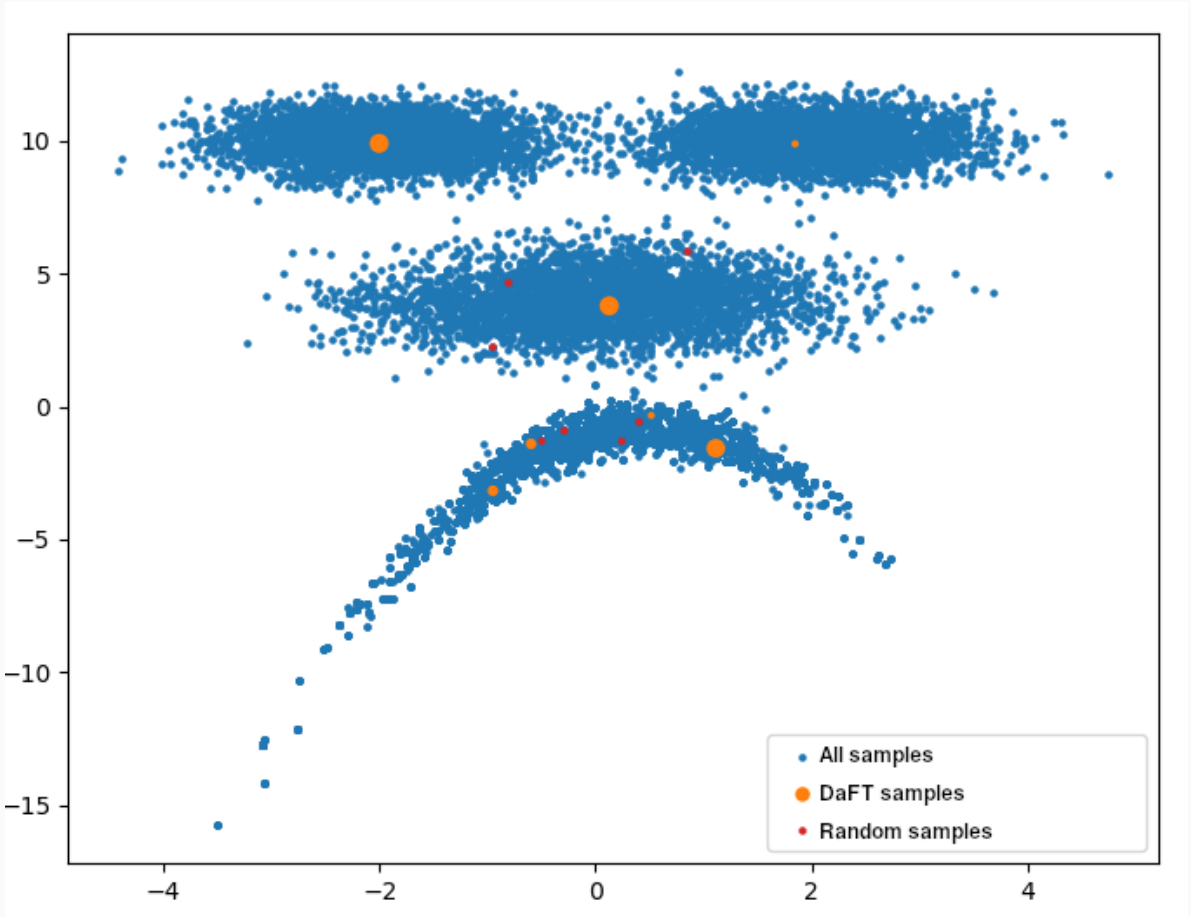


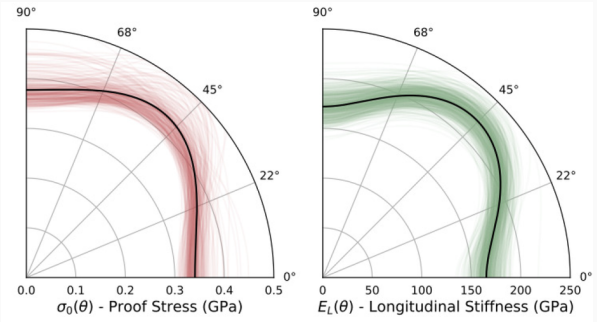
Figure 3: Random samples from the famous Thimomenos [greek: "angryman"] distribution. All samples (blue), weighted DaFT samples (orange) and simple random samples (red).

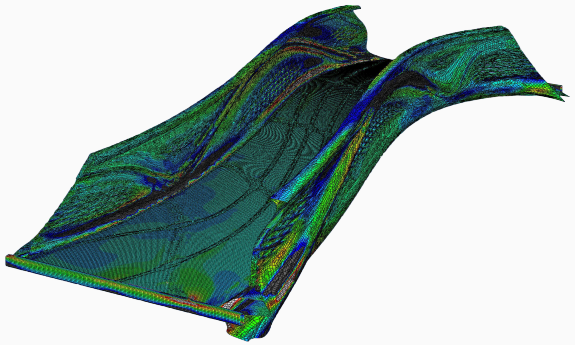
Future Directions and Discussion

3D Printed Bridge

- Posterior distribution of **material parameters** from **laboratory experiments**.
- Very **expensive bridge model** and a limited computational budget.
- **Quantify the uncertainy** in $\mathcal{Q}$ (e.g. maximum stress/strain or displacement) with **few model evaluations**.

$\Rightarrow$ Choose samples from posterior using DaFT.





Discussion

- A **simple** and **flexible** approach to thinning of samples.
- Applications to MLMC, MCMC, multi-fidelity and surrogate models.
- **Very fast** approximate MMD with a choice of different kernels.
- Genetic algorithm yields a **population** of near–optimal solutions.

- **Required dimension** of the feature space is problem–dependent and currently **unclear**.
- The **genetic algorithm** approach **works well** but it is not the most rigorous approach. $\Rightarrow$ Alternatives?
- Samples are no longer random. $\Rightarrow$ Consequences?

Questions?

m.lykkegaard@exeter.ac.uk

References

Papadimas, N. and Dodwell, T. (2021). A hierarchical bayesian approach for calibration of stochastic material models. *Data-Centric Engineering*, 2:e20.

Rahimi, A. and Recht, B. (2007). Random Features for Large-Scale Kernel Machines. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

Riabiz, M., Chen, W., Cockayne, J., Swietach, P., Niederer, S. A., Mackey, L., and Oates, C. J. (2020). Optimal Thinning of MCMC Output. *arXiv:2005.03952 [math, stat]*. arXiv: 2005.03952.