Task-Agnostic Sample Design for Machine Learning

Bhavya Kailkhura

CASC, Lawrence Livermore National Lab

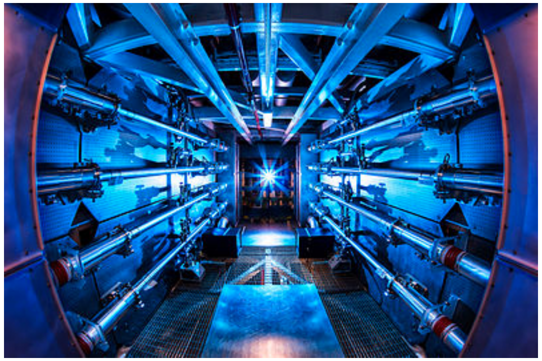
Joint work with: Jay Thiagarajan, Qunwei Li, Jize Zhang, Yi Zhou, Timo Bremer

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

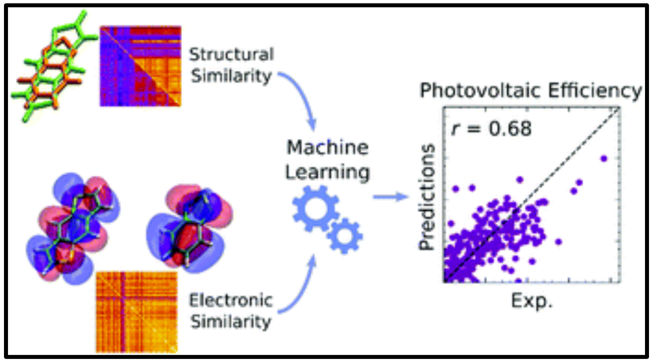
Lawrence Livermore National Laboratory

ML provides incredible opportunities in science

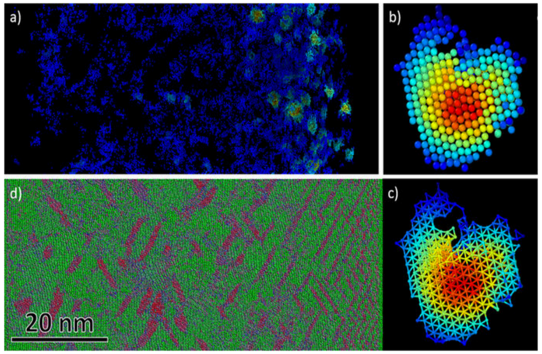
Inertial Confinement Fusion



Material Discovery

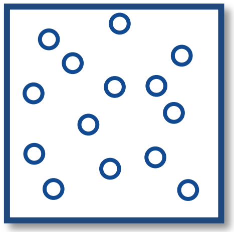


Stockpile Stewardship



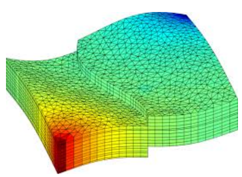
Scientific discoveries fundamentally rely on our understanding of high-fidelity experimental data

A typical scientific data science pipeline



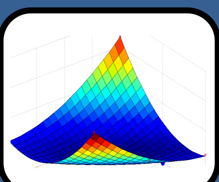
SAMPLE DESIGN

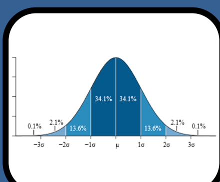
Decide random set of samples to cover the **N**-dimensional parameter space

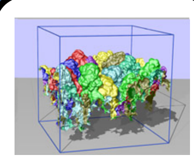


Experiments

Run corresponding experiments to create a baseline of knowledge





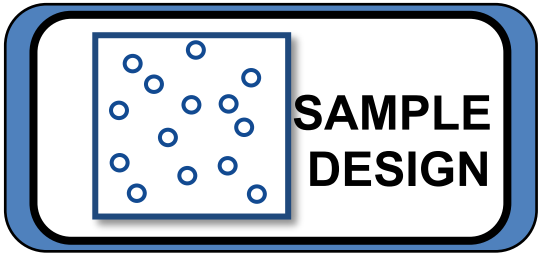


Analyze the resulting ensemble

- Build a reliable predictive model
- Optimization

Scientific experiments are really expensive!

Sample design is crucial for the success of scientific ML



- Excellent generalization
- Low sampling rates
- Controlled variance

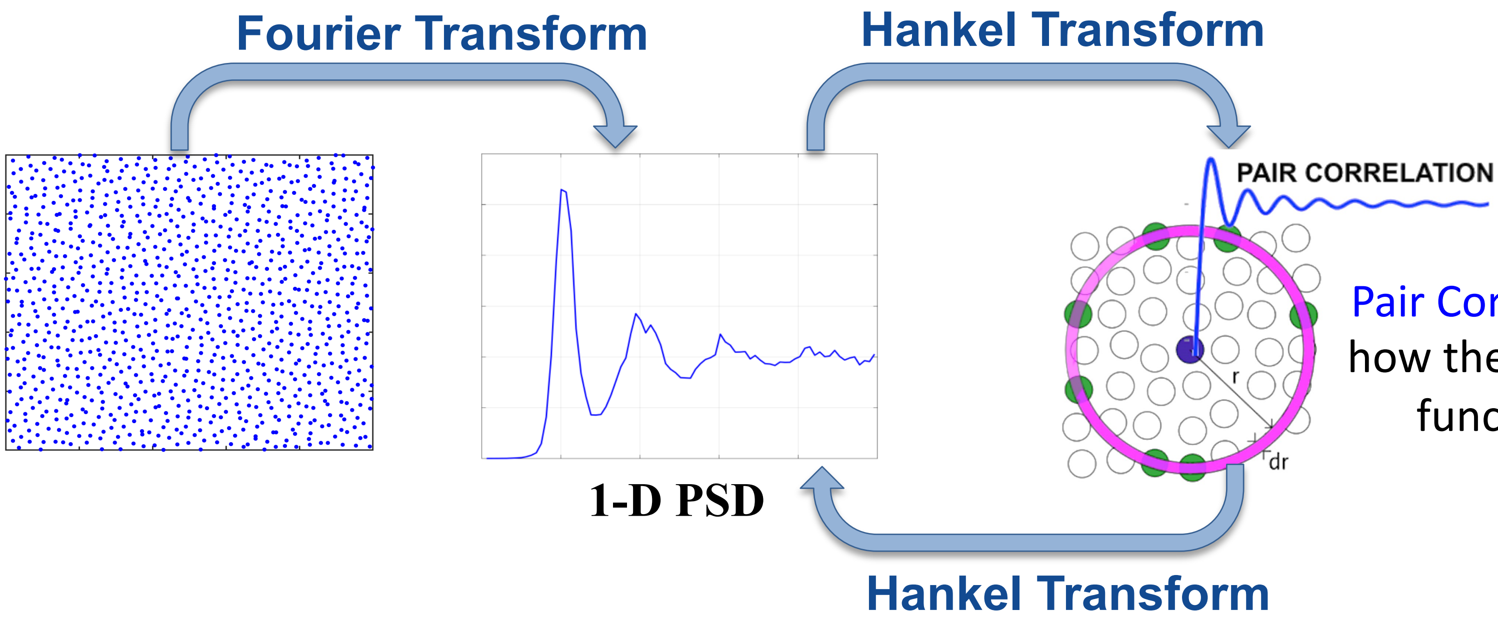
Plethora of methods

- Uniform random
- Latin Hypercubes
- Voronoi Tessellation
- Orthogonal arrays
- Quasi Monte Carlo
- …

Given a fixed sampling budget, which experiments to run to acquire the most amount of information?

A new spectral sampling theory for sample design

Characterize spatial properties using the Pair Correlation Function (PCF) and develop a mathematical connection to Power Spectral Density (PSD)



Pair Correlation: Measures how the density varies as a function of distance

A neat theoretical connection: $P(k) = 1 + \rho(2\pi)^{\frac{d}{2}} k^{1-\frac{d}{2}} H_{\frac{d}{2}-1}\left(r^{\frac{d}{2}-1}(G(r)-1)\right)$

*B. Kailkhura, et. al., “A spectral approach for the design of experiments: Design, analysis and algorithms.” The Journal of Machine Learning Research 19.1 (2018): 1214-1259.

Risk minimization using Monte Carlo estimates

Consider the following general setup to learn the function $h : X \to Y$ by minimizing the ***population risk***:

$$\boxed{R_P(h) \triangleq \mathbb{E}_{P(x,y)}[l(h(x), y)] = \int l(h(x), y) dP(x, y)}$$

In general, the joint distribution *P(x, y)* is unknown, we minimize the ***empirical risk***

$$\boxed{R_S(h) \triangleq \frac{1}{N} \sum_{i=1}^{N} l(h(x_i), y_i)}$$

The generalization error is defined as

$$\boxed{\mathrm{gen}(h) \triangleq \mathbb{E}_S[(R_P(h) - R_S(h))^2]} \quad = \quad bias^2 + var(R_S(h))$$

Connecting generalization error with spectral sampling

We restrict our analysis to homogeneous sampling patterns, which are unbiased

Lemma 1. *The generalization error in terms of the power spectra of both the sampling pattern and the loss function in the toroidal domain can be obtained as:*

$$gen(h) \triangleq \frac{1}{N} \int_{\Theta} \mathbb{E}(\mathcal{P}_S(\omega))\mathcal{P}_l(\omega) d\omega \tag{8}$$

Theorem 2. *The generalization error for isotropic homogeneous sampling patterns (in polar coordinates) is given by*

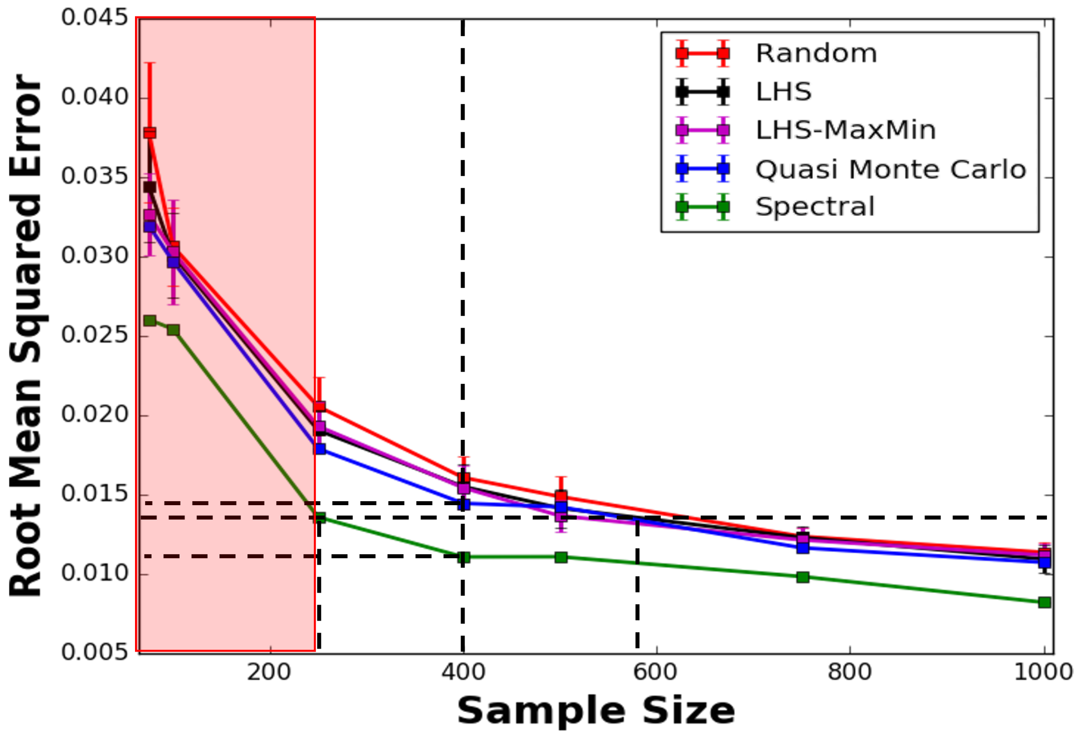
$$gen(h) \triangleq \frac{\mu(\mathcal{S}^{d-1})}{N} \int_0^{\infty} \rho^{d-1} \mathbb{E}(\hat{\mathcal{P}}_S(\rho))\hat{\mathcal{P}}_l(\rho) d\rho, \tag{9}$$

An ideal sampling power spectrum must attain zero values in the low frequency regime

B. Kailkhura, et. al., “A Look at the Effect of Sample Design on Generalization through the Lens of Spectral Analysis”.
Pilleboue, Adrien, et al. "Variance analysis for Monte Carlo integration." ACM Transactions on Graphics (TOG) 34.4 (2015): 1-14.

Predicting peak pressure in NIF 1-d hotspot simulator

We use random forest regressor to learn peak pressure by varying 2 input parameters and performance is evaluated on 10K unseen test samples



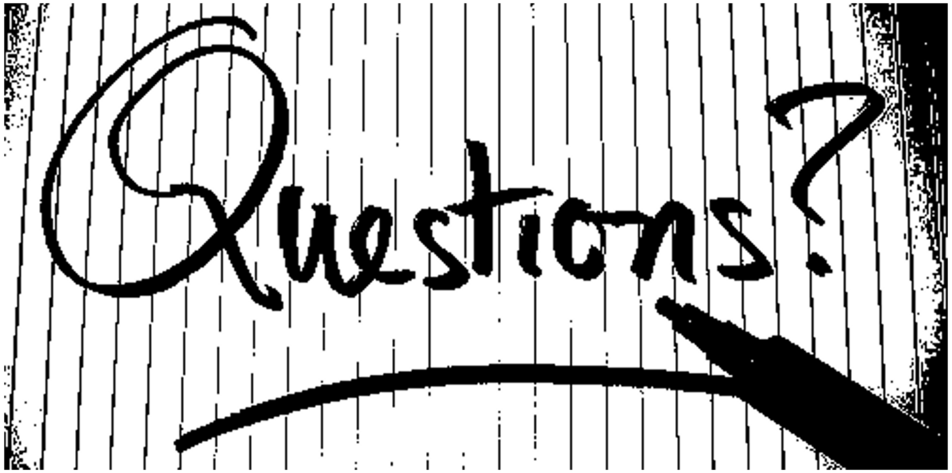
Spectral sampling

- ~ 30% less test error
- ~ 50% less samples
- Low Variance

Summary

- A general theoretical framework for studying the generalization performance of task-agnostic sampling patterns
- Spectral sampling is an effective alternative to creating baseline of knowledge in small data scientific ML applications
- Exploiting the connection between Fourier and Spatial statistics enables the design of sampling patterns that outperform existing methods at low sampling rates

Improved sample designs can enable unprecedented capabilities in computational sciences



Contact

Bhavya Kailkhura
Center for Applied Scientific Computing
Lawrence Livermore National Laboratory
Email: kailkhura1@llnl.gov