Overview

With breakthroughs in experimental methods and computational technology, there are now novel sources of high-quality data for tackling a broad array of pressing problems in science and engineering. However, the generation of such high-fidelity data often requires costly experiments and/or simulations, which can significantly limit the amount of data available for scientific investigation. Given this cost bottleneck, it is of critical importance to develop *cost-efficient sampling methods* for data generation and model training. Furthermore, for scientific inference, such sampling methods need to be performed with *confidence*; they need to be coupled with theoretically sound and data-driven stopping rules, which guarantee the resulting statistical model achieves a desired error tolerance. This is paramount for *reliable* scientific discovery: it provides a quantification of uncertainty for scientific inference, thus protecting against spurious findings.

This project will develop a novel and timely suite of methods that jointly addresses this crucial need for cost-efficient and confident sampling for scientific discovery. Our framework features methodologies (with supporting theory and algorithms) that extend classical low discrepancy (i.e., highly stratified) sampling techniques for a broad range of challenging scenarios encountered in modern scientific problems, including cost-efficient Bayesian inference, efficient subsampling of massive data, multi-fidelity modeling, and density estimation. Major emphasis is placed on demonstrating the effectiveness of these methods for accelerating scientific discoveries, especially for the PIs' ongoing collaborations on the study of heavy-ion collisions and real-time engine control of unmanned aircraft vehicles, but also for new collaborations that will be developed over the project.

Intellectual Merit

Our project investigates four novel directions that extend low discrepancy sampling to complex settings in modern scientific and engineering problems. Each direction is necessitated by a motivating scientific problem from the PIs' multi-disciplinary collaborations and plays an integral role in our proposed suite of methods for accelerating scientific discovery. The first, called cost-efficient Stein points, extends low-discrepancy sampling for expensive Bayesian computation problems, where each posterior evaluation requires a forward run of a costly scientific simulator. The second direction explores adaptive algorithms and stopping rules for confident multifidelity sampling, which is widely used in the physical sciences. The third extends low discrepancy sampling for efficient and confident big data analysis to facilitate real-time decision-making. The last direction investigates low discrepancy sampling for distribution, density, and quantile estimation. Each direction will involve the development of novel methodology, theory, and algorithms, with a keen focus on addressing scientific needs in our aforementioned ongoing collaborations.

Broader Impacts

The proposed suite of methods paves the way for transformative scientific research, equipping practitioners with cost-efficient and confident sampling methods (with supporting theory and algorithms) for accelerating scientific and engineering discoveries. Our project will catalyze closer collaborations between the scientific and data science communities by broadening the application of low-discrepancy sampling for the complex settings featured in modern scientific and engineering problems. Such collaborations will be further strengthened via our open-source Python QMC library QMCPy, which provides an accessible and well-documented software connecting state-of-the-art low-discrepancy methods with the broader scientific community. This project will also train the next generation of science-based computational researchers, who can adeptly work in diverse and multi-disciplinary scientific teams pushing forward the frontiers of scientific knowledge.