

# Cost-Efficient and Confident Sampling for Modern Scientific Discovery

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 greatly 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 that 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 address 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* (LD<sup>1</sup>) (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 tackling a wide array of complex scientific and engineering problems, especially for the PIs’ ongoing collaborations on the study of heavy-ion collisions and real-time engine control of unmanned aircraft vehicles, but for also new collaborations that will be developed over the life of the project.

This project will be led by Fred Hickernell (FH, PI from Illinois Tech), Simon Mak (SM, PI from Duke U), Yuhan Ding (YD, co-PI from Illinois Tech), and Sou-Cheng Terrya Choi (SCTC, SAS & Illinois Tech, Senior Personnel). Our collaborators include Michael McCourt (MM, Intel), Chris Oates (CO, Newcastle U), Art Owen (AO, Stanford U), Jagadeeswaran Rathinavel (JR, Wi-Tronix), Pieterjan Robbe (PR, Sandia National Laboratories & KU Leuven), Illinois Tech PhD students Claude Hall, Jr. (CH) and Aleksei Sorokin (AS), Duke U PhD students Irene Ji (IJ), John Miller (JM), Kevin Li (KL), and Tao Tang (TT), and other students, alumni, and friends.

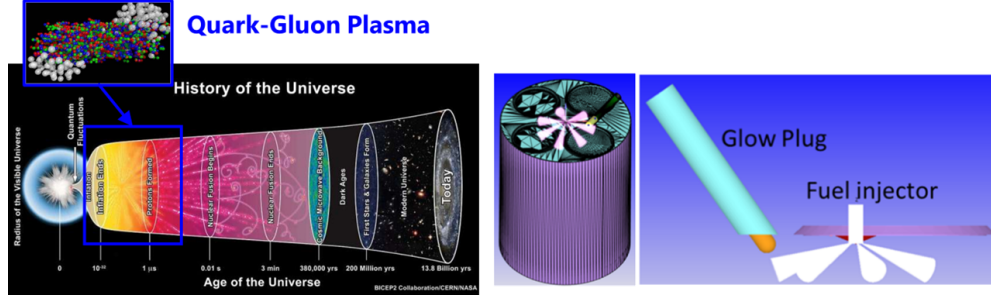
## 1. THE NEED FOR COST-EFFICIENT AND CONFIDENT SAMPLING

**1.1. Motivation.** The nature of scientific discovery has undergone a radical paradigm shift over recent decades. With tremendous advances in experimental methodology and computational technology, scientists now have the capability to generate high-quality data for solving challenging problems which were once thought to be impossible or prohibitively expensive. In the physical sciences, complex phenomena such as universe expansions [77] and rocket propulsion [98, 146] can now be reliably studied via fine-scale virtual (i.e., computer) simulations on high performance computing systems. Similarly, in the biological sciences, fundamental developments in high throughput sequencing have led to key advances in genomics and epidemiology.

There are, however, two critical bottlenecks that greatly hinders the use of this high-quality data for scientific discovery. The first bottleneck is that, for complex problems, *such data can be very costly to generate*, requiring a large investment of computational and/or experimental resources. Take, e.g., the study of the Quark-Gluon Plasma (QGP), a deconfined phase of nuclear matter which filled the Universe shortly after the Big Bang (see recent papers by PI SM [37, 38, 67, 88] with the JETSCAPE collaboration, discussed later in Sect. 4.2). With advances in nuclear physics modeling, this plasma can be simulated by virtually colliding heavy ions together at near-light speeds. The ultimate goal is to perform an inverse problem, which uses such simulations with physical measurements from particle colliders to learn plausible properties of this plasma (call this *t*) and hence shed light on the origins of matter. But such simulations are very costly: a single

---

<sup>1</sup>Acronyms and initials of personnel contain hyperlinks to their full names.



**Figure 1.** (Left) Visualizing the QGP shortly after the Big Bang [89], which became the building blocks of matter in the Universe. (Right) A schematic of the UAV metal engine: fuel is injected at seven nozzles, then ignited via a glow plug.

run at parameter  $t$ , yielding simulation output  $g(t)$ , can take on the order of thousands of CPU hours [37]. This inverse problem may require thousands of simulation runs at different choices of  $t$ , in order to find suitable parameters which best match cosmological measurements. Such a study can thus require *millions* of CPU hours, which places great strain on computational costs. Similar cost bottlenecks are faced in a broad variety of modern science and engineering problems.

The second bottleneck is that the increasing sophistication of scientific experiments results in *massive datasets which take on highly complex forms*. In our nuclear physics application, the simulated output  $g(t)$  consists of fine-scale spatiotemporal flows for the hydrodynamic evolution of the plasma, which can take up to exabytes ( $10^{18}$  bytes) of storage. The efficient use of such massive data is thus critical for timely scientific inference and decision-making. Similar challenges arise in the PI SM’s current collaboration with mechanical engineers at the University of Minnesota, on the engine control of unmanned aircraft vehicles (UAVs). Here, we wish to train a *real-time* control system for UAV engines with low cetane numbers, an important objective for U.S. Army’s single fuel concept. Due to expensive prototyping costs, such systems are typically trained via complex computational fluid dynamics (CFD) simulations, which output a wide range of fine-scale engine characteristics. The key bottleneck is that simulation outputs are massive datasets, consisting of hundreds of response variables and millions of timesteps. The use of such big data for training a statistical model for *real-time* engine control is thus a critical need for practical implementation.

These challenges necessitate two crucial ingredients to facilitate scientific discovery: cost-efficiency and confidence. The first, *cost-efficiency*, refers to statistical methods which aim to maximize model performance given a limited cost (e.g., computational) budget. For the earlier nuclear physics problem, a cost-efficient method strives to provide an accurate solution to the inverse problem given a limited computational budget (and thus limited simulation runs). The second, *confidence*, refers to methods that provide a reliable and estimable measure (or quantification) of model error. Such confidence is essential for verifiable scientific discovery: it provides a quantification of uncertainty for findings, thus protecting against spurious findings. This further allows for theoretically sound stopping rules that guarantee model performance, which is particularly crucial in this cost-constrained setting. For our nuclear physics problem, a confident method yields a measure of uncertainty for the inverse problem, which can be used to guide the amount of simulation runs needed to ensure a desired error tolerance for physics discovery.

Both of the above bottlenecks can be alleviated via a careful integration of novel *sampling* algorithms within the statistical learning framework. Here, “sampling” refers to the drawing of representative samples  $T_1, \dots, T_n$  from a (potentially) complex probability distribution  $F$ , and the use of such samples for scientific inference and decision-making. For the first bottleneck of *costly* scientific data, a careful sampling design of input parameters for the expensive simulator can enable accurate and precise scientific inference in reasonable turnaround times. For the second bottleneck of *massive* data, a judicious sampling of large datasets can allow for timely scientific discovery and useful decision-making. There is, however, much to be done on *cost-efficient* and

*confident* sampling algorithms, given the complexities present in modern scientific problems. We will thus propose a suite of methods, with supporting theory and algorithms, which address this important gap. We first present the prototypical problem of interest, then provide a survey of *low-discrepancy* sampling methods [107], which we extend for our suite of methods.

**1.2. Background.** We first provide a brief background on the prototypical problem we aim to address, then discuss existing literature and its limitations for our motivating applications.

*1.2.1. Prototypical Problem.* Consider the estimation of the *expectation* of a random variable  $Y$  whose distribution is some complicated function,  $g$ , of a random vector  $\mathbf{T}$  with known distribution:

$$(1a) \quad \mu := \mathbb{E}(Y) = \mathbb{E}[g(\mathbf{T})] = ?$$

This arises in many practical problems, e.g., in our **UAV** application,  $\mathbf{T}$  may represent uncertainties in engine operating conditions,  $g(\mathbf{t})$  may denote the engine thrust at operating conditions  $\mathbf{t}$ , and we wish to learn the expected engine thrust  $\mathbb{E}[g(\mathbf{T})]$  under uncertain conditions.

In turn, we define a transformation,  $\Psi$ , of a *standard uniform* random vector,  $\mathbf{X}$ , into  $\mathbf{T}$ , so that we may also write  $Y$  as a function,  $f$ , of  $\mathbf{X}$ :

$$(1b) \quad \mathbf{T} = \Psi(\mathbf{X}), \quad f(\mathbf{x}) = g(\Psi(\mathbf{x})) |\partial \Psi(\mathbf{x}) / \partial \mathbf{x}|, \quad \mathbf{X} \sim \mathcal{U}[0, 1]^d,$$

$$(1c) \quad \mu = \mathbb{E}[f(\mathbf{X})] = \int_{[0, 1]^d} f(\mathbf{x}) \, d\mathbf{x} = ?.$$

This expectation can also be thought of as a *multivariate integration* problem. Here, we write  $Y = f(\mathbf{X})$  because the **LD** sequences described later mimic uniform random vectors to approximate  $\mu$ .

Besides the integral problem in (1), it is often useful to know the *distribution*,  $F$ , *density*,  $\varrho$ , and/or *quantile* function,  $Q$ , of  $Y$ , i.e.,

$$(2a) \quad F(y) := \mathbb{P}(Y \leq y) = \mathbb{P}[f(\mathbf{X}) \leq y] = \int_{[0, 1]^d} \mathbb{1}_{(-\infty, y]}(f(\mathbf{x})) \, d\mathbf{x}, \quad \varrho(y) := F'(y), \quad Q(p) := F^{-1}(p).$$

In the **UAV** problem, we may wish to learn the distribution of engine thrust  $Y$  under uncertain operating conditions; indeed, these are the primary quantities of interest when designing for robust engines. We will address these problems in Sect. 2.4 in the context of our motivating applications.

In practice, the population mean or multivariate integral,  $\mu$ , often cannot be evaluated by analytic means, but it may be estimated by the sample mean,  $\hat{\mu}_n$ , i.e.,

$$(3) \quad \mu \approx \hat{\mu}_n := \frac{1}{n} \sum_{i=1}^n Y_i = \frac{1}{n} \sum_{i=1}^n f(\mathbf{X}_i), \quad Y_i = f(\mathbf{X}_i).$$

Given an error tolerance,  $\varepsilon$ , we want to choose samples,  $\mathbf{X}_1, \dots, \mathbf{X}_n$ , to mimic  $\mathcal{U}[0, 1]^d$  and satisfy

$$(4) \quad |\mu - \hat{\mu}_n| \leq \varepsilon \quad \text{with high confidence.}$$

*1.2.2. Low Discrepancy Sampling.* Choosing  $\mathbf{X}_1, \mathbf{X}_2, \dots$ , for evaluating the sample mean in (3) to be *independent and identically distributed* (**IID**) yields a root mean square error of  $\sqrt{\mathbb{E}[|\mu - \hat{\mu}_n|^2]} = \text{std}(f(\mathbf{X}))n^{-1/2}$ , which is independent of the dimension,  $d$ , but slowly vanishing as  $n \rightarrow \infty$ . The computational cost to satisfy the error tolerance (4) is  $\mathcal{O}(\varepsilon^{-2})$ . Tensor product generalizations of one-dimensional numerical integration rules using grid sampling give errors of  $|\mu - \hat{\mu}_n| = \mathcal{O}(n^{-r/d})$  and a computational cost of  $\mathcal{O}(\varepsilon^{-d/r})$ , where  $r$  is limited by both the sophistication of the rule and the smoothness of the integrand,  $f$ . Such rules may be suited for small dimensions  $d$ , but they are inefficient for the larger  $d$  that occurs in our applications of interest.

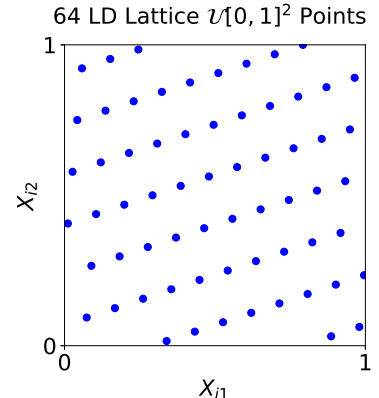
A superior approach that combines the intentional structure of a grid with the essentially dimensionless error bound of **IID** sampling is **LD** sampling,  $\mathbf{X}_1, \mathbf{X}_2, \dots \stackrel{\text{LD}}{\sim} \mathcal{U}[0, 1]^d$ , which has an error bound of [54, 107]

$$(5) \quad |\mu - \hat{\mu}_n| \leq D(\{\mathbf{X}_i\}_{i=1}^n) \|f - \mu\|_{\mathcal{F}}.$$

The discrepancy,  $D(\{\mathbf{X}_i\}_{i=1}^n)$ , corresponds to the norm of the cubature error functional [53] for the function space  $\mathcal{F}$ . The discrepancy is also a measure of how close the empirical distribution (which assigns equal probability to each point) is to the uniform distribution. The discrepancy is typically  $\mathcal{O}(n^{-1+\delta})$ , where  $\delta$  is arbitrarily small and positive. This faster convergence rate translates into a computational cost of  $\mathcal{O}(\varepsilon^{-1-\delta})$  to satisfy error criterion (4), under mild smoothness conditions on  $f$ . The semi-norm  $\|\cdot - \mu\|_{\mathcal{F}}$  is called the *variation* and is a measure of function roughness.

Popular LD sampling schemes include lattices [32, 107, 125] and digital nets [30, 107]. Fig. 2 displays  $n = 64$  LD lattice points intended to mimic  $\mathcal{U}[0, 1]^2$ . A two-dimensional grid of 64 points would only have 8 different values in each coordinate direction, whereas the LD sample covers 64 different values in each coordinate direction. Such schemes are available in many softwares, including BRODA [80], CUBA [47] MATLAB [135], NAG [136], PyTorch [115], SciPy [139], TensorFlow [134], and uncertainty quantification libraries such as Dakota [1], MUQ [114], UQLab [101], and UQTK [28, 29]. Efforts to identify better LD sequence generators include LatNet Builder [26] and the Magic Point Shop [109].

PI FH, co-PI YD, SCTC, MM, JR, AS and collaborators have developed more comprehensive QMC libraries, GAIL [21] for MATLAB and QMCPy [22] for Python. These libraries include the stopping criteria mentioned below, as well as flexible variable transformations of the form (1b). Our recent effort is focused on QMCPy, and includes an active repository [22], documentation [20], a tutorial [52], and a blog [19].



**Figure 2.** LD lattice points, which have fewer gaps and clusters of points than either the IID or grid points.

**1.2.3. Error Bounds and Stopping Criteria.** Algorithms based on efficient LD sampling are commonly called *quasi-Monte Carlo* (QMC) algorithms. To construct a QMC algorithm, one needs not only an LD sequence but also *reliable* and *practical* bounds on the error  $|\mu - \hat{\mu}_n|$  which are based on the function data,  $f(\mathbf{X}_1), f(\mathbf{X}_2), \dots$ . Such *data-based* error bounds inform the stopping criteria, which guide the determination of sample size  $n$  needed to satisfy the error tolerance (4). One approach is to use  $N$  different randomizations of a single LD sequence to compute  $N$  sample means, and then estimate the error of the grand sample mean,  $\hat{\mu}_{nN} = (\hat{\mu}_n^{(1)} + \dots + \hat{\mu}_n^{(N)})/N$ , via, e.g., bootstrapping. But this would be too costly for our motivating applications since it requires  $nN$  evaluations of the expensive function  $f$  (see first bottleneck in Sect. 1.1).

PI FH and his collaborators have developed two kinds of theoretically justified stopping criteria for LD sampling based on the Fourier coefficients of the data  $f(\mathbf{X}_1), f(\mathbf{X}_2), \dots$ . The first kind determines a bound on  $|\mu - \mu_n|$  by inferring the roughness of  $f$  from the decay of the discrete Fourier coefficients [57, 59, 70]. The second kind uses Bayesian credible intervals for  $|\mu - \mu_n|$  assuming that  $f$  is an instance of a Gaussian process whose hyperparameters are tuned by the function data [56, 65, 66]. The cost of bounding the error  $|\mu - \mu_n|$  is only  $\mathcal{O}(n \log(n))$  for both kinds of stopping criteria. This can be achieved for the Bayesian approach by choosing covariance kernels that match the LD sampling schemes, and thus avoiding the typical  $\mathcal{O}(n^3)$  cost.

**1.3. Limitations of Existing Methodology and Software.** There are, however, fundamental limitations which prevent the effective use of existing methods for tackling the complexities in our applications. These limitations, detailed below, are gaps which we will address in our proposal.

**Limitation 1: Expensive Bayesian Sampling.** Bayesian inference [42] is a popular statistical framework for tackling scientific problems. Here, parameters of interest are sampled from a “posterior” probability distribution  $\varrho$ , which captures both the prior belief of the modeler and evidence from the collected data. In practical problems, the posterior distribution is often highly complex and available only in proportional form. This is further complicated by the *costly* nature of each

posterior evaluation (see first bottleneck in Sect. 1.1). For example, *each* evaluation of the unnormalized posterior density in our heavy-ion physics inverse problem requires thousands of CPU hours [37]. Existing work on LD sampling, however, focuses largely on the sampling from uniform distributions (a literature review is provided later). We develop in Sect. 2.1 a novel cost-efficient LD posterior sampling method which addresses this need, with applications to our heavy-ion physics application and broader scientific problems.

**Limitation 2: Multifidelity Modeling.** For many scientific computing problems, the random variable of interest  $Y_{\boldsymbol{\eta}} = f_{\boldsymbol{\eta}}(\mathbf{X}_{\boldsymbol{\eta}})$  is parameterized by a *fidelity* parameter  $\boldsymbol{\eta}$ , and the desired quantity of interest is the limiting mean  $\mu_{\infty} = \lim_{\boldsymbol{\eta} \rightarrow \boldsymbol{\eta}_{\infty}} \mathbb{E}(Y_{\boldsymbol{\eta}})$ . In our heavy-ion application,  $f_{\boldsymbol{\eta}}(\mathbf{X}_{\boldsymbol{\eta}})$  may represent a observable simulated from a complex partial differential equation (PDE) system modeling the heavy-ion collision, with random coefficients  $\mathbf{X}_{\boldsymbol{\eta}}$  representing uncertainties in simulation inputs, and mesh-size parameters  $\boldsymbol{\eta}$  controlling simulator fidelity. As fidelity increases, the cost of sampling  $Y_{\boldsymbol{\eta}}$  also increases (as each evaluation  $f_{\boldsymbol{\eta}}$  becomes more expensive), and thus the evaluation of many high fidelity  $Y_{\boldsymbol{\eta},i}$  to approximate  $\mu_{\infty}$  can be prohibitively costly.

Multifidelity (also known as multi-level or multi-index) methods [45, 48, 49] choose a sequence of fidelity parameters,  $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \dots$ , and write the desired quantity as a telescoping sum:

$$\mu = \mu_{\infty} = (\mu_{\boldsymbol{\eta}_1} - \mu_{\boldsymbol{\eta}_0}) + (\mu_{\boldsymbol{\eta}_2} - \mu_{\boldsymbol{\eta}_1}) + \dots + (\mu_{\boldsymbol{\eta}_L} - \mu_{\boldsymbol{\eta}_{L-1}}) + (\mu_{\infty} - \mu_{\boldsymbol{\eta}_L}), \quad \mu_{\boldsymbol{\eta}_0} = 0.$$

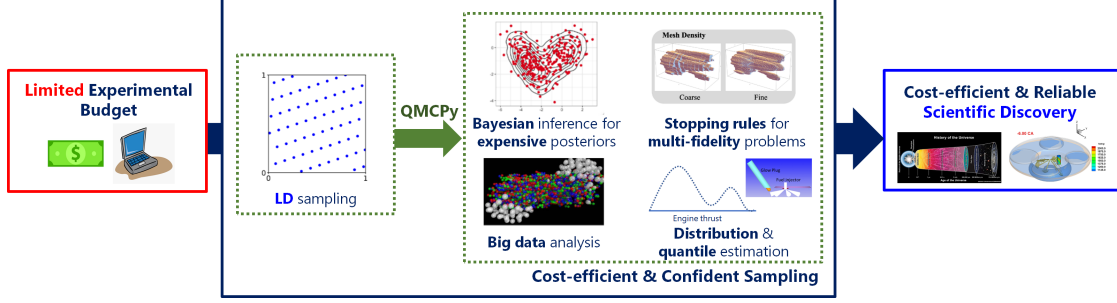
The sequence of fidelity parameters is chosen so that the cost of evaluating  $f_{\boldsymbol{\eta}_l}(\mathbf{X}_{\boldsymbol{\eta}_l})$  increases with  $l$  (greater fidelity), but the sample size required to estimate  $\mu_{\boldsymbol{\eta}_l} - \mu_{\boldsymbol{\eta}_{l-1}}$  accurately decreases with  $l$ . The term  $\mu_{\infty} - \mu_{\boldsymbol{\eta}_L}$  is approximated by zero. Substantial cost-efficiency is gained by using many cheap samples to approximate the low fidelity terms and relatively fewer expensive samples to compute the high fidelity terms. While there is a rich literature on multifidelity methods (including recent work by PIs SM and FH, which we discuss later), there has been little work on developing measures of confidence or stopping criteria (see Sect. 1.2.3) for such multifidelity methods. Such stopping criteria are important for our heavy-ion application, allowing for confident scientific inference with minimal experimental costs. We will address such limitations in Sect. 2.2.2.

**Limitation 3: Big Data Analysis.** Given sophisticated computing technology, scientific simulators typically output massive datasets with complex forms, and the efficient use of such data for scientific discovery is paramount. One solution is to take a small representative *subsample* of the big data, and use this for efficient model training. The careful selection of this subsample is critical for timely decisions. Machine learning algorithms often make use of stochastic gradient descent [6, 131], which takes a new *random* subsample of the big data at each gradient update; such random subsampling, however, may be practically and theoretically inefficient given a computational constraint (more on this later). There has been recent work on extending LD ideas for big data subsampling (discussed later), but such methods typically do not perform well or have sound theoretical guarantees for complex learning models, which are desired with massive training data. To address this, we propose in Sect. 2.3 a new LD subsampling method, which provides provably improved LD big data subsampling for a broad class of learning models.

**Limitation 4: Distribution, Density, and Quantile Estimation.** In many problems, practitioners wish to estimate not only the mean of  $Y = g(\mathbf{T})$ , but also its distribution, density or quantile (as defined in (2)). This is the case in our UAV problem: aerospace engineers wish to estimate not only the expected engine thrust  $E(Y)$  under uncertain operating conditions  $\mathbf{T}$ , but also characterize its full distribution for designing robust engines. The error analysis for such problems, especially for LD sequences, is underdeveloped in the literature, and rigorous, data-based stopping criteria are non-existent. We address this limitation in Sect. 2.4.

**Limitation 5: Software Quality.** QMC software implemented by non-experts may be flawed. FH and MM found that randomized PyTorch Sobol' points fell on the boundaries of  $[0, 1]^d$ , when they never should [116] due to a lack of double precision. Lluís Antoni Jiménez Rugama (LIAJR), a





**Figure 3.** Project workflow: given a limited budget, the proposed tasks (Sect. 2) extend LD sampling to the complex settings motivated by our applications, providing a toolbox for cost-efficient and reliable scientific discovery. QMCpy (Sect. 3.2) serves as an open-source software package that disseminates our suite of methods to the scientific community.

former PhD student of **FH**, alerted that MATLAB’s Sobol’ sequence scrambling was incorrect; MATLAB later corrected this error. After a vigorous discussion on the PyTorch [116] and SciPy [124] issues sites, **AO**, **FH**, and other **QMC** researchers convinced the developers not to omit the first Sobol’ point and to randomize by default. **AO** explained why this is crucial [112]. The need for a vigilant **QMC** software community is addressed in Sect. 3.2. Further, solving complex problems well and efficiently often requires multiple software libraries. For example, in uncertainty quantification, one integrand,  $f(\mathbf{X}_i)$ , may be the output from a PDE library. Not all libraries connect well, nor are there yet standards in the QMC software community on how to pass information from one library to the next. We address this issue in Sect. 3.2.

## 2. A FRAMEWORK FOR COST-EFFICIENT AND CONFIDENT SAMPLING

We now present a suite of novel methods (with supporting theory and algorithms) that extend **LD** sampling to the complex settings from our motivating applications. These methods provide a useful toolbox for cost-efficient and confident sampling to accelerate scientific discovery. Figure 3 shows the workflow for the four proposed tasks, which directly address the limitations in Sect. 1.3.

### 2.1. Bayesian Sampling for Expensive Posteriors [SM lead, FH, IJ, TT, JM]

*2.1.1. Background and Preliminary Results.* Bayesian methods [42] rely on MCMC sampling to explore the posterior distribution  $F$ , which captures information on model parameters (see PI **SM**’s work [63, 97, 100]). However,  $F$  can often be *expensive* to evaluate in many scientific computing problems (see **Limitation 1**). This is compounded by the highly correlated nature of traditional MCMC samplers, which reduces the information provided by each sample [86]. Existing samplers for such problems thus be prohibitively *costly* [75], and an LD posterior sampling method can provide improved Bayesian learning given a computational budget.

One approach for **LD** posterior sampling is to minimize the *kernel discrepancy*  $D_K(\{\mathbf{T}_i\}_{i=1}^n, F)$  [54], which measures the difference between the empirical distribution of  $\{\mathbf{T}_i\}_{i=1}^n$  and the posterior  $F$  via a symmetric positive-definite kernel  $K$ . This approach, known as *kernel herding* [16], has a key limitation: the discrepancy requires an analytic form for the integral  $\int K(\mathbf{t}, \cdot) dF(\mathbf{t})$ , which is unattainable for complex posteriors  $F$ . To address this, [15] proposed the “Steinized” kernel:

$$(6) \quad K_{\text{ST}}(\mathbf{t}, \mathbf{x}) = \nabla_{\mathbf{t}} \cdot \nabla_{\mathbf{x}} K(\mathbf{t}, \mathbf{x}) + \nabla_{\mathbf{t}} K(\mathbf{t}, \mathbf{x}) \cdot \nabla_{\mathbf{x}} \log dF(\mathbf{x}) \\ + \nabla_{\mathbf{x}} K(\mathbf{t}, \mathbf{x}) \cdot \nabla_{\mathbf{t}} \log dF(\mathbf{t}) + K(\mathbf{t}, \mathbf{x}) \nabla_{\mathbf{t}} \log dF(\mathbf{t}) \cdot \nabla_{\mathbf{x}} \log dF(\mathbf{x}),$$

where  $\nabla$  and  $\nabla \cdot$  are the gradient and divergence operators, respectively. With this,  $\int K_{\text{ST}}(\mathbf{t}, \cdot) dF(\mathbf{t})$  evaluates to 0, thus yielding a closed form expression for the *kernel Stein discrepancy* (KSD):

$$(7) \quad D_K(\{\mathbf{T}_i\}_{i=1}^n, F) := \sqrt{\frac{1}{n^2} \sum_{i,j=1}^n K_{\text{ST}}(\mathbf{T}_i, \mathbf{T}_j)}.$$

[15] then employs a sequential optimization of the KSD: the first sample  $\mathbf{T}_1^*$  is taken at the mode of  $F$ , then subsequent samples  $\mathbf{T}_2^*, \mathbf{T}_3^*, \dots$  are obtained by sequentially optimizing the KSD, i.e.:

$$(\text{ESP}_n) \quad \mathbf{T}_n^* \leftarrow \arg \min_{\mathbf{t}} \text{KSD}_n(\mathbf{t}) := \arg \min_{\mathbf{t}} D_K(\{\mathbf{T}_i^*\}_{i=1}^{n-1} \cup \mathbf{t}, F), \quad K = K_{\text{ST}}, \quad n = 2, 3, \dots$$

These optimized samples, called *Stein points*, can yield improved representation of the posterior  $F$  over standard MCMC methods for a fixed sample size  $n$ .

However, Stein points have a key drawback: they are *not* cost-efficient. When the posterior  $F$  is expensive, the optimization of the Stein discrepancy for a *single* sample point will require *many* evaluations of  $F$  in the form of its score function  $\nabla_{\mathbf{t}} \log dF(\mathbf{t})$ . For our heavy-ion application, given a fixed budget of  $10^6$  CPU hours, we can afford  $10^3$  evaluations of  $F$ , which translates to only  $n \approx 50$  Stein points - this is inadequate for exploration of complex posterior distributions.

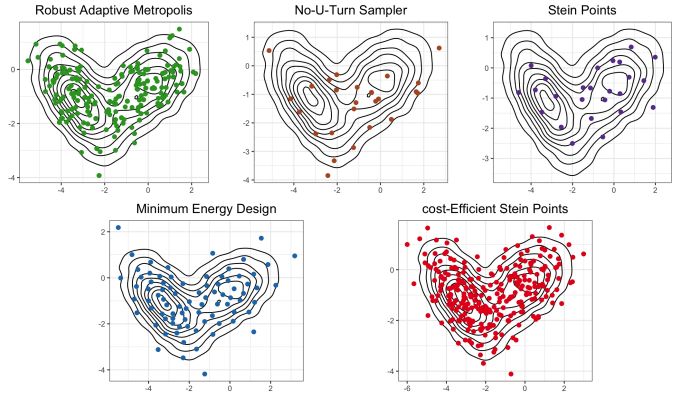
We thus propose the following cost-Efficient Stein Points (ESPs). The key idea is the construction of a sequence of carefully constructed Gaussian process (GP) surrogate models [123] on the expensive objective functions  $\text{KSD}_n$ . Consider first  $(\text{ESP}_n)$  for a given  $n$ . Suppose the posterior in the form of its score function  $\nabla_{\mathbf{t}} \log dF(\mathbf{t})$  has already been evaluated at  $M_n$  candidate points  $\{\mathbf{T}_j\}_{j=1}^{M_n}$ , yielding objective evaluations  $\mathcal{D}_n = \{\text{KSD}_n(\mathbf{T}_j)\}_{j=1}^{M_n}$  via (7). Under a GP prior on  $\text{KSD}_n(\cdot)$ , the posterior distribution of  $\text{KSD}_n(\mathbf{t})$  can be shown to be  $\text{KSD}_n(\mathbf{t})|\mathcal{D}_n \sim \mathcal{N}\{\mu_n(\mathbf{t}), \sigma_n^2(\mathbf{t})\}$ ; specific forms for  $\mu_n(\mathbf{t})$  and  $\sigma_n^2(\mathbf{t})$  can be found in [123]. Using this and following the literature on Bayesian optimization (e.g., [73] and work from PI SM [17]), the *expected improvement* in objective  $\text{KSD}_n$  from evaluating the posterior at a new point  $\mathbf{t}$  takes the closed-form expression:

$$(8) \quad \text{EIKSD}_n(\mathbf{t}) = (\text{KSD}_{n,\min} - \mu_n(\mathbf{t}))\Phi\left(\frac{\text{KSD}_{n,\min} - \mu_n(\mathbf{t})}{\sigma_n(\mathbf{t})}\right) + \sigma_n(\mathbf{t})\phi\left(\frac{\text{KSD}_{n,\min} - \mu_n(\mathbf{t})}{\sigma_n(\mathbf{t})}\right),$$

where  $\text{KSD}_{n,\min}$  is the best observed  $\text{KSD}_n$  value. We thus wish to evaluate the expensive posterior at the point that maximizes  $\text{EIKSD}_n(\mathbf{t})$ . This procedure, of refitting the GP surrogate and evaluating the posterior at the point of greatest expected improvement, is then iterated until a convergence criterion is met. The evaluated point with smallest  $\text{KSD}_n$  is then taken as the next sample  $\mathbf{T}_n^*$  in  $(\text{ESP}_n)$ .

Consider next the *sequence* of optimization problems in  $(\text{ESP}_n)$  for ESP sampling. A key observation is that the posterior evaluations used for solving previous problems  $\text{ESP}_2, \dots, \text{ESP}_{n-1}$  can be directly *reused* for the current problem  $\text{ESP}_n$ , since such evaluations translate directly to evaluations of  $\text{KSD}_n$  via (7). Thus, as sample size  $n$  increase, this allows for increasingly more data on the objective  $\text{KSD}_n$  to generate the  $n$ -th ESP  $\mathbf{T}_n^*$ . This recycling of posterior evaluations enables cost-efficient ESP sampling given a limited computational budget.

To demonstrate the cost-efficiency of ESPs, Fig. 4 compares several state-of-the-art samplers (the robust adaptive Metropolis sampler [138], the No-U-Turn sampler [62], Stein points [15], minimum energy designs [75]) on a 2D two-mixture normal distribution. All samplers are limited to  $B = 500$  posterior evaluations. We see that, as expected, existing samplers that ignore the expensive nature of posterior evaluations provide a poor approximation of  $F$ : they either yield a small sample size ( $n \approx 50$ ), or a highly correlated sample chain. ESPs,



**Figure 4.** Visualizing the sampled points on a 2D two-mixture normal distribution, using four existing posterior samplers and the proposed ESPs. All samplers are limited to 500 posterior evaluations.

on the other hand, provide a noticeably larger sample size  $n = 287$  with low sample correlation. This improved posterior representation is confirmed via a comparison of marginal statistics or distributional metrics.

*2.1.2. Theory.* [Years 1–3] Given these promising results, we will investigate key theoretical properties for ESP sampling. A key property to establish is, given an error tolerance  $\varepsilon$  for posterior approximation, what is the computational cost (i.e., the number of posterior evaluations  $B$ ) required to achieve such a tolerance with ESPs. This provides a theoretical basis for comparison with existing samplers which ignore evaluation costs. Such rates will require a careful integration of cost complexity results for Monte Carlo [45] with Bayesian non-parametrics theory [61]. We will also explore the cost-efficiency of ESPs using a variety of probabilistic surrogate models, particularly models that can learn embedded low-dimensional structure for high-dimensional approximation, and models that can integrate prior scientific information (see papers by PI SM [13, 148, 151] and [67, 68, 88]).

*2.1.3. Randomization and Central Limit Theorems.* [Years 1–2] In QMC, the randomization of LD sequences is an important emerging topic, providing a basis for probabilistic inference (e.g., confidence intervals) on integral estimands [31]. Such randomization is particularly important in this expensive Bayesian setting, where a quantification of estimation uncertainty is desired given limited posterior evaluations. We will develop a randomized implementation of ESPs, where in addition to providing an efficient LD sampling of the posterior, each marginal sample  $\mathbf{T}_n^*$  is random and follows the desired posterior distribution. We will further prove a Central Limit Theorem, which characterizes the asymptotic distribution of integral estimators as  $n \rightarrow \infty$ . With this, we will develop confidence intervals for the proposed randomized ESPs (following [120]), and use this to demonstrate the improved cost-efficiency of ESPs over existing MCMC samplers.

*2.1.4. Implementation and Application.* We will demonstrate the usefulness of ESPs in a wide range of modern scientific problems involving expensive Bayesian inference. This includes our motivating heavy-ion application (see Sect. 1.1), where the Bayesian inference of plasma properties from particle colliders requires costly forward runs (thousands of CPU hours) for each evaluation of the posterior. PI SM has worked extensively in this area (see [36–40, 82, 88]) as a member of the JETSCAPE collaboration (discussed later in Sect. 4.2). We will further explore broader applications of ESPs in Bayesian sensor imaging, rocket engine design and astrophysics, for which the PIs have close ongoing collaborations. The proposed algorithms will be implemented on our open-source package QMCPy (discussed later in Sect. 3.2).

## 2.2. Adaptive Multifidelity Algorithms [FH lead, SCTC, YD, MM, PR, CH, AO, AS]

*2.2.1. Motivation.* A common uncertainty quantification problem in the physical sciences involves the estimation of  $\mu = \mathbb{E}(Y)$ , where  $Y = f(\mathbf{X})$  and  $f$  can only be approximated by  $f_{\boldsymbol{\eta}}$ , with  $\boldsymbol{\eta}$  denoting the fidelity of the approximation. In geophysics,  $f$  may be the solution of a (partial) differential equation modeling fluid flow, whose boundary conditions are given by a random spatial field, and  $\boldsymbol{\eta}$  may denote the mesh size of the numerical solver and the discretization of the random field. In our heavy-ion application,  $f$  may be the exact solution of a particle collision observable from a complex physics model with uncertain plasma properties  $\mathbf{X}$  as inputs, and  $\boldsymbol{\eta}$  may capture the spatial and temporal mesh size of the simulator (see work on this by PI SM [67, 68, 88]).

As mentioned in Sect. 1.3, approximating the true solution by a single high fidelity expectation can be prohibitively costly. Rather, multifidelity methods [45, 48, 49]—an active research area—consider a sequence of problems,  $\{\mu_l = \mathbb{E}[f_{\boldsymbol{\eta}_l}(\mathbf{X}_{\boldsymbol{\eta}_l})]\}_{l=0}^L$ , where the fidelity increases with  $l$  as does the computational cost of evaluating an instance  $f_{\boldsymbol{\eta}_l}(\mathbf{X}_{\boldsymbol{\eta}_l,i})$ . One can efficiently approximate  $\mu$  by using more cheap samples to approximate  $\mu_l - \mu_{l-1}$  for small  $l$  and fewer expensive samples to approximate  $\mu_l - \mu_{l-1}$  for large  $l$ . There has been recent work (see [44, 45], including recent work by PI SM [132]) which show that, under mild conditions on the cost function of the simulator  $f_{\boldsymbol{\eta}}$  and



its convergence rates, multifidelity methods can provide noticeably more accurate and confident estimates over single-fidelity approaches, given a cost budget  $B$ .

Despite this, there has been little work on stopping criteria for multifidelity sampling methods (see [Limitation 2](#)). This is crucial for scientific discovery; in our heavy-ion application, such criteria provide physicists with a confident quantification of uncertainty given a computational budget, or equivalently, a confident estimate of budget required to achieve a desired precision on findings. We propose below two novel directions which address this in a cost-efficient manner.

*2.2.2. Extending Stopping Rules to Multifidelity Problems.* [Years 1-2] Adaptive algorithms for these multifidelity problems use ad hoc stopping criteria. We propose to develop rigorous stopping criteria like those described in [Sect. 1.2.3](#). PI [FH](#), [AS](#), [JR](#), and collaborators have developed stopping criteria for functions of several expectations,  $C(\boldsymbol{\mu})$ , [\[59, 130\]](#). For example, Bayesian posterior means, [\[43\]](#), can be written as the ratio of two expectations,  $C(\boldsymbol{\mu}) = \mu_2/\mu_1$ . Sensitivity indices [\[121, 122, 128\]](#) also involve the computation of more than one expectation. However, in both these cases, the underlying random vector,  $\mathbf{X}$ , is the same for all expectations,  $\mu_1, \mu_2, \dots$ , and only the functions  $f_l$  defining the expectations are different.

For multifidelity problems each  $\mu_l$  depends on a different  $\mathbf{X}_{\eta_l} \sim \mathcal{U}[0, 1]^{d_l}$  with a different  $d_l$ . Thus, adaptive algorithms need to manage [LD](#) sequences of different dimensions as well as different sample sizes. Adaptive decisions must be made on whether to devote more effort to sampling the low or high fidelity terms. The wealth of literature on multifidelity methods and our experience on developing rigorous stopping criteria for single fidelity problems gives us confidence in success.

*2.2.3. Implementation and Application.* We will demonstrate the effectiveness of the above developments for confident sampling in a broad spectrum of scientific problems involving multifidelity modeling. This includes our motivating heavy-ion application ([Sect. 1.1](#)), where the simulators for particle collisions have multiple fidelity parameters involving spatial meshes and time-steps at different stages (see PI [SM](#)'s papers [\[67, 68\]](#)). We will further explore broader applications in geophysical applications, which the PIs have close ongoing collaborations (see [\[147\]](#)). The proposed algorithms will be implemented and fully documented in [QMCPy](#) (see [Sect. 3.2](#))

### 2.3. Big Data Subsampling [[SM](#) lead, [AO](#), [IJ](#), [KL](#)]

*2.3.1. Motivation and Preliminary Results.* Big data is ubiquitous with advances in technology and computing. In our [UAV](#) application, the output of the numerical simulator can require terabytes of storage (see [Sect. 1.2](#)). A key challenge is that learning algorithms need to be *scalable* to extract useful information from such data for *real-time* decisions, e.g., real-time engine control for [UAV](#) flight. One strategy is to iteratively train the model on small batches of the data, typically sampled uniformly at random. This *subsampling* scales up state-of-the-art machine learning algorithms, such as stochastic gradient descent (SGD, [\[6\]](#)) and stochastic gradient boosting [\[41\]](#).

Consider SGD, which minimizes the loss  $L(\theta; \mathcal{T}) = N^{-1} \sum_{m=1}^N l(\theta; \mathbf{T}_m)$  over model parameters  $\theta \in \mathbb{R}^q$ , where  $\mathcal{T} = \{\mathbf{T}_m\}_{m=1}^N \subset \mathbb{R}^d$  is the large training data. Standard gradient descent [\[108\]](#) is impractical here, since they require evaluation of the full gradient  $N^{-1} \sum_{m=1}^N \nabla_{\theta} l(\theta; \mathbf{T}_m)$ , which is very expensive with  $N$  large. Mini-batch SGD [\[6\]](#) approximates this via a subsample  $\mathcal{T}_s^{[l]} \subset \mathcal{T}$  of size  $n \ll N$ , taken [IID](#) and uniformly from  $\mathcal{T}$ . The descent steps are iterated until convergence:

$$(9) \quad \theta^{[l+1]} \leftarrow \theta^{[l]} - \zeta \left( \frac{1}{n} \sum_{\mathbf{T} \in \mathcal{T}_s^{[l]}} l(\theta; \mathbf{T}) \right), \quad l = 1, 2, \dots,$$

where  $\zeta$  is the gradient descent step size. Mini-batch SGD is widely used for scalable training of neural networks and deep learning models with big data [\[131\]](#).

Mini-batch SGD, however, has a key limitation. Since gradients are estimated by *random* subsampling, the solution sequence  $(\theta^{[l]})_{l=1}^{\infty}$  converges to a *noise ball* of radius  $\mathcal{O}(n^{-1})$  around the global optimum  $\theta^*$ . For small subsample sizes  $n$  (as necessitated from our cost-constrained

setting), SGD can thus return estimates *very far* from  $\theta^*$ . Our solution is to choose an **LD** dataset that well-represents the big data  $\mathcal{X}$ . This is known as “data squashing” (termed by **AO** in [111]), and encompasses work on leverage-score subsampling [90], coresets [3, 9, 64], experimental design [142, 143], and work by **PI SM** [79, 92–94]. **LD** data squashing for SGD is a timely problem, but one largely unaddressed in the literature for complex non-linear models (see **Limitation 3**).

We propose a new data squashing method which makes use of **LD** subsampling of big data  $\mathcal{T}$  for accelerating SGD. The preliminary result below guarantees the *existence* of such a subsample:

**Theorem 1.** *Let  $\mathcal{T} = \{\mathbf{T}_m\}_{m=1}^N$  (the “big data”) be any set of points on  $[0, 1]^d$ , and suppose the feasible space  $\Theta$  is convex. Further suppose  $n \leq \sqrt{N}$  and the loss function  $l$  is convex with mild regularity conditions. Then there exists a subsample  $\mathcal{T}_s \subseteq \mathcal{T}$  of size  $n$  which, when used within the SGD iterative updates (9), yields a solution sequence  $(\theta^{[l]})_{l=1}^\infty$  converging to a noise ball of radius  $\mathcal{O}\{(\log n)^{3d+1}/n^2\}$  around the global optimum  $\theta^*$ .*

This theorem guarantees that, under mild assumptions on the loss function  $l$  (satisfied by a broad range of learning models), there exists an **LD** subsample of the big data which, when used within SGD, converges a noise ball of radius  $\mathcal{O}\{(\log n)^{3d+1}/n^2\}$  around the desired solution  $\theta^*$ . Thus, with a carefully chosen **LD** subsample, the proposed “**LD**-batch SGD” can yield *improved* optimization over standard mini-batch SGD, which converges to a *larger* noise ball of radius  $\mathcal{O}(n^{-1})$ . Put another way, this suggests that **LD**-batch SGD can yield comparable performance to mini-batch SGD with far fewer optimization iterations, thus allowing for *large computational savings for big data analysis*. We will tackle the following tasks to flesh out a comprehensive methodological framework for **LD**-batch SGD.

**2.3.2. Optimization of LD Subsample.** [Years 1–2] While the existence of an **LD** subsample is promising, one challenge is in finding such a subset efficiently. We will find this via the following optimization approach. Define the so-called “data kernel” using the big data  $\mathcal{T} = \{\mathbf{T}_m\}_{m=1}^N$ :

$$(10) \quad K_{\text{data}}(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{k} \in \mathbb{Z}^d \setminus \mathbf{0}} \lambda_{\mathbf{k}} \phi_{\mathbf{k}}(\mathbf{x}) \overline{\phi_{\mathbf{k}}(\mathbf{y})}, \quad \phi_{\mathbf{k}}(\mathbf{x}) = e^{2\pi i \mathbf{k}^T \mathbf{x}} - b_{\mathbf{k}}, \quad b_{\mathbf{k}} = \frac{1}{N} \sum_{m=1}^N e^{2\pi i \mathbf{k}^T \mathbf{T}_m}.$$

where  $i$  is the imaginary number, and  $\lambda_{\mathbf{k}} = \prod_{j=1}^d \max(1, 2\pi|k_j|)^{-2}$ . Here, the coefficients  $b_{\mathbf{k}}$  can be efficiently computed via non-linear fast Fourier transform [140]. The data kernel  $K_{\text{data}}$  has two nice properties. We can show that the subsample  $\mathcal{X}_s$  minimizing the kernel discrepancy with  $K_{\text{data}}$  yields the improved rate in Theorem 1. We can also show that, with all coefficients computed, this discrepancy can be evaluated in  $\mathcal{O}(n^2)$  work, *independent* of  $N$  (the big data size). We will develop *scalable algorithms to optimize this data discrepancy for LD subsampling*, leveraging recent developments in accelerated gradient descent [71] and randomized algorithms [91].

**2.3.3. Implementation and Application.** [Years 2–3] We will show the usefulness of the proposed **LD** subsampling on a broad range of scientific applications. In particular, we will showcase its effectiveness on our **UAV** application. The goal here is to train an *efficient* predictive model that can be used for *real-time UAV* flight, facilitating engine control decisions within a timeframe of several milliseconds. The challenge is that such a model has to be trained using massive simulation data from computational fluid dynamics models. We will show our **LD** subsampling approach can indeed facilitate this real-time control for **UAV** flight. We will also provide specific *implementations* of **LD**-batch SGD for a broad range of popular learning models (e.g., regression, neural networks, kernel methods), with full documentation on **QMCPy** (see Sect. 3.2).

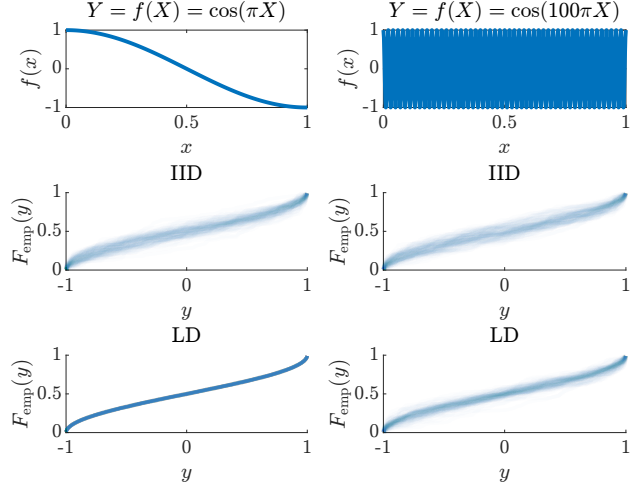
## 2.4. Distribution, Density and Quantile Estimation [FH lead, **AO**, **AS**]

**2.4.1. Motivation and Preliminary Results.** Besides expectations, we may wish to know the distribution, density, and/or quantile function of  $Y = f(\mathbf{X})$ . As opposed to estimating a scalar,  $\mu$ , we now need to approximate functions,  $F$ ,  $\varrho$ , and/or  $Q$ , which require function approximation

algorithms and error criteria involving function norms. This is critical in our **UAV** application: engineers are interested in estimating and quantifying uncertainty on the *distribution* of engine thrust under uncertain operating conditions, which enable robust engine design and control.

Although the distribution,  $F$ , is defined as an integral in (2), the integrand as it stands,  $\mathbb{1}_{(-\infty, y]}(f(\cdot))$ , is insufficiently smooth for **QMC** theory to apply directly, and thus there has been little developments on **LD** sampling for such a setting (see **Limitation 4**). Approximating the density (derivative of  $F$ ) and quantile function (inverse of  $F$ ) are harder problems. Despite such theoretical challenges, **LD** sequences show promise in approximating distributions, densities, and quantiles. Consider two different definitions of  $Y = f(X)$ ,  $X \sim \mathcal{U}[0, 1]$  as shown in Fig. 5. For both choices the distribution function of  $Y$  is  $F : y \mapsto \sin^{-1}(y)/\pi + 1/2$ . Fig. 5 displays  $N = 100$  replications of the empirical distribution,  $F_{\text{emp}}$  based on  $n = 64$  **IID** and **LD** samples. The **LD** points provide better approximations of the distribution, and they provide better approximations for smoother  $f$ .

A recent study of **LD** sequences for kernel density estimation [4] has demonstrated that in practice they outperform **IID** sequences. The authors also derive theoretical error bounds on this estimator for **LD** sequences, but such bounds appear to be too loose. A tight quantification of uncertainty for density estimation is critical in our **UAV** application, enabling sound design and control decisions. There has also been little work on investigating (and mitigating) the curse-of-dimensionality for density estimation with **LD** sequences, a fundamental topic in **QMC** [31]. In **QMC**, effective dimension is defined roughly as the number of coordinates of the input vector  $\mathbf{X}$  that make substantial contributions to  $f(\mathbf{X})$ . One expects that small effective dimension plays an important role in approximating  $\varrho$  with **LD** sequences, just as it is in approximating  $\mu = \mathbb{E}[f(\mathbf{X})]$  with **LD** sequences [31]. This is critical in our **UAV** problem, where there are many design and control parameters for engine operation.



**Figure 5.** Comparison of the empirical distribution functions of  $Y = f(X)$  using  $n = 64$  **IID** and **LD**  $X_i$  and for two definitions of  $f$ . The empirical distributions generated by **LD** points are closer to the true distribution than **IID** sampling.

*2.4.2. Exploring the Effectiveness of LD Sequences.* [Year 1–3] The theory for using **LD** sequences to estimate distributions, densities, and quantiles is in its infancy, which tempers our ambitions for these problems. Our first step will be to empirically compare the efficiency of **LD** sequences to **IID** sequences by inserting **LD** sequences into standard algorithms that use **IID** sequences, most likely some form of kernel smoothing [141]. We will experimentally explore how to choose the band-width of kernel smoothing methods. We will explore how the nominal dimension, the effective dimension, and the smoothness of  $f$  influence the effectiveness of **LD** sequences. We will investigate its effectiveness in extensive experiments on the test function library [5].

Based on our computational experiments, we will derive data-driven algorithms that compute optimal bandwidth for smoothing and determine the sample size needed to satisfy the error criterion. We will pursue the two approaches used for stopping criteria for computing  $\mu$ , namely i) analyzing the Fourier coefficients of  $f$ , and ii) assuming  $f$  comes from a Gaussian process. We will demonstrate the effectiveness of such methods on our **UAV** flight application.

### 3. BROADER IMPACTS

**3.1. Dissemination to the Broad Scientific Community.** For the proposed suite of methods to become an integral tool for modern scientific discovery, scientists and engineers need to be convinced that our sampling framework can achieve confident and accurate estimates in a cost-efficient manner. This necessitates a *strategic outreach* and *targeted dissemination* of our methodologies to scientific communities who may not be familiar with the potential of **LD** sampling for accelerating scientific discovery. We will achieve this via a *multi-disciplinary* publication plan that targets a broad multi-disciplinary audience. We will publish in not only statistics, mathematics, and computer science journals, but also top subject-matter journals to make the proposed tools accessible to the scientific community. We will post all work as freely accessible papers on arXiv, with links to open source and well-documented code on GitHub to allow for easier dissemination of our methods to end-users. We will also give talks, tutorials and workshops at prominent statistics, machine learning, mathematics and scientific conferences, educating different communities on promising developments on novel **LD** methods for furthering scientific progress.

By broadening **QMC** methods for tackling complex scientific and engineering problems, our project will naturally *introduce new application areas* to the benefits of **LD** sampling. The accessibility of state-of-the-art **QMC** methods in our outreach will spur on novel advancements in many scientific disciplines. We will focus specifically on four areas:

*3.1.1. Uncertainty Quantification (UQ).* UQ is the science of quantifying and managing uncertainty in computational and physical systems [126, 127]. Since data are typically expensive for UQ, a key focus is the design of sampling points for experimentation. **QMC** can thus yield *great computational savings* for UQ (see [50] for a convincing application in fluid flows). PI **SM** has ongoing multi-disciplinary collaborations on UQ in aerospace engineering [10, 11, 84, 85, 98, 146], nuclear physics [8, 36–38, 67, 68, 82, 88], astrophysics [96, 99, 151] and neuroscience [144, 145], and we will introduce the benefits of **LD** sampling for UQ in such disciplines.

*3.1.2. Machine Learning (ML).* Machine learning is a rapidly growing area with broad applications in science and engineering. Given the prevalent use of Monte Carlo in ML algorithms [6, 41, 117], **QMC** will undoubtedly have a significant impact in this area (see [78] for a stunning application of **QMC** in image rendering, and [12] for an application of **QMC** for learning PDEs). In addition to **LD** big data subsampling (Sect. 2.3), we will *show the advantages of QMC over IID sampling in cutting-edge ML problems*, an area where both PIs have prior publication record [74, 94, 99].

*3.1.3. High Energy Physics.* Physicists have begun to recognize the advantages of **LD** sampling for generating parton distributions [25]. We will continue our recent discussions with this group to explore how to best employ **LD** sampling for a cost-efficient simulation of such problems. PI **SM** has an extensive publication record in high energy physics (see Section 4.2 and [8, 36–38, 67, 68, 82, 88]).

*3.1.4. Bayesian Methodology.* In addition to expensive posterior sampling (Sect. 2.1), we will explore two Bayesian areas which can greatly benefit from **QMC**. The first is probabilistic numerics (PN), which presumes the solution of a mathematical problem follows a prior random process. PI **FH** with his former PhD student **JR** developed a fast Bayesian cubature method [65] using lattice **LD** sampling. Bayesian optimization (**BO**) aims to minimize a black-box objective function, where evaluations are optimized via an acquisition function in the form of a high-dimensional integral. **MM** illustrates the advantages of **LD** sampling for **BO** [19, qEI with **QMCPy**], and PI **SM** has also worked in this area [17, 95]. We will demonstrate the *efficacy of LD sampling* in **BO** and **PN**.

**3.2. QMCPy as a Proving Ground.** As mentioned in Sects. 1.2.2 and 1.2.3, **FH**, **YD**, **SCTC**, **AS**, **MM**, **JR**, **PR**, and collaborators have created the open source Python **QMC** library **QMCPy** [22]. **QMCPy** has a clear architecture and consistent user interface. It includes **LD** sequences, variable transformations, and stopping rules. **QMCPy** is hosted on a GitHub repository with documentation, a suite of doctests and unit tests, Jupyter notebook demos, and an issues board.



With this architecture, **QMCPy** will serve as an ideal vessel for bridging our suite of methods (and state-of-the-art **QMC** algorithms) to the scientific community, thus addressing **Limitation 5**.

**QMCPy** will continue to grow and *stand in the breach* between research code from individual groups and large-scale software packages. Research groups need to compare their new ideas with the best available. Developers for **LD** generators need to test them on a variety of use cases and as key components of **QMC** algorithms. Those with new **QMC** algorithms need to test them with the best generators. We will connect **QMCPy** to other libraries in Sect. 1.2.2, taking advantage of what they offer in **QMCPy** and pushing our more established developments into them. We recently collaborated with Uncertainty Quantification and Model Bridge (UM-Bridge) [27] to make **QMCPy** compatible with UM-Bridge. A recent gathering at an international **QMC** conference initiated by **FH** raised the need to standardize formats for **LD** generators to allow them to be shared across software libraries. We will pursue initiatives that promote same look-and-feel across **QMC** software.

**3.3. Promoting Proper QMC Practice and Code.** **QMCPy**—software, documentation, academic articles, and conference presentations—will *showcase the right way* to do **LD** sampling. As an example, the adoption of PyTorch into **QMCPy** and the tutorial given by **FH** at MCQMC 2020 [23, 51] prompted a vigorous discussion on the PyTorch issues site [116] that migrated to the SciPy issues site [124]. **AO**, **FH**, and other **QMC** researchers convinced the developers to not omit the first Sobol’ point, but to randomize by default. Keeping the first point preserves the net property of the first  $2^m$  Sobol’ points and randomization can speed up convergence [112]. In these discussions, it was pointed out that UQLab [101], OpenTurns [110], and other packages routinely drop the first Sobol’ point, a bad but understandable practice. The arguments we provided to PyTorch and SciPy developers addressed their concerns. We expect this project to produce fruitful discussions between **QMC** practitioners, which will promote better practice.

Having the eyes of the **QMC** community on **QMCPy** will more *quickly uncover and eradicate bugs*. **FH** found that randomized PyTorch Sobol’ points fell on the boundaries of  $[0, 1]^d$ , when they never should [116]. This was due to a lack of double precision, as discovered by **MM**. **LIAJR** found that the Sobol’ scrambling in MATLAB was incorrect. This was rectified in R2017a. This highlights how having a larger community using a software library leads to higher quality code.

### 3.4. Training the Next Generation of *Science-Based Computational Researchers*.

There is a critical need for individuals who have the computational and mathematical tools to adeptly work in modern *scientific* teams, geared towards pushing forward the frontier of scientific knowledge and engineering through improved science-based data science tools. Students involved in this project will be given a *multidisciplinary* background in science-based data science, and will learn relevant methods in statistics, mathematics and computer science. This will serve them well in transitioning to not just academic research teams but also government and industry positions.

Computational mathematics and statistics require the use of others’ code, hopefully in the form of well-developed software packages. New scholars also need to be trained not only on how to use such packages, but how to contribute to them as well. Students supported by this project will learn to write clean, efficient code that fits package architecture, is documented, and passes doctests. Students will learn about repositories and software engineering tools, which will prepare them well for academic and/or industry opportunities.

**3.5. Incorporating Diversity, Equity and Inclusion** As with past projects, in seeking students, we will give preference to underrepresented minorities, women, and students from colleges where research experiences are rare. As noted in Sects. 4.1 and 4.2, we have had a strong track record of mentoring students from *diverse* backgrounds and experiences. Five of **FH**’s fifteen students who earned PhDs are women, three of whom are academics. **FH** supervised an African-American student, **CH**, in the Summer Undergraduate Research Experience (SURE) program at Illinois Tech, which provides research opportunities particularly for underrepresented groups. **CH** is now a PhD student at Illinois Tech. Two of **SM**’s seven current PhD students are women,



and five of **SM**'s ten undergraduate thesis advisees are women. Many of our undergraduates have enrolled in graduate programs. **SM** is on the leadership board of the IMS New Researchers Group, which aims to provide career development opportunities for young statisticians, particularly those from underrepresented backgrounds and institutions.

We will leverage these connections to further promote diversity and inclusion, providing students from underrepresented backgrounds and institutions with ample research training and opportunities. The senior personnel on this project include two women (**SCTC** and **YD**) and one early-career scholar (**SM**). Our senior personnel and collaborators include folks with diverse technical expertise, institutions and backgrounds. The students that we mentor in this project will also learn and benefit from thinking from these diverse perspectives.

#### 4. RESULTS FROM PRIOR NSF SUPPORT

**4.1. NSF-DMS-1522687, *Stable, Efficient, Adaptive Algorithms for Approximation and Integration*, \$270,000, August 2015 – July 2018.** Fred Hickernell (**FH**, PI) and Gregory E. Fasshauer (**GEF**, co-PI) led this project, and **SCTC** contributed as senior personnel. Other contributors were **FH**'s research students **YD** (PhD 2015), **LJ** (PhD 2016), **LIAJR** (PhD 2016), Da Li (**DL**, MS 2016), Jiazhen Liu (**JL**, MS 2018), JR (PhD 2019), Xin Tong (**XT**, MS 2014, PhD 2020 at the University of Illinois at Chicago), Kan Zhang (**KZ**, PhD student), Yizhi Zhang (**YZ**, PhD 2018), and Xuan Zhou (**KZ**, PhD 2015). Articles, theses, software, and preprints supported in part by this grant include [2, 18, 21, 33, 34, 46, 55–60, 65, 69, 70, 72, 83, 87, 102–106, 118, 119, 149, 150, 152, 153].

*4.1.1. Intellectual Merit from Prior NSF Support.* **FH**, **SCTC**, **YD**, **LIAJR**, **DL**, **JR**, **XT**, **YZ**, and developed several adaptive algorithms for univariate integration, function approximation, and optimization [18, 24, 33, 137, 149], multivariate integration [57, 59, 70], and multivariate function approximation for Banach spaces,  $\mathcal{F}$ , defined by series representations [34, 35].

*4.1.2. Broader Impacts from Prior NSF Support.* Publications by **GEF**, **FH**, **SCTC**, students, and collaborators are listed above. We have spoken at many applied mathematics, statistics, and computational science conferences and given colloquium/seminar talks to mathematics and statistics departments. **FH** co-organized the 2016 Spring Research Conference, gave an invited tutorial at MCQMC 2016, was a program leader for the SAMSI 2017–18 Quasi-Monte Carlo (QMC) Program, and received the 2016 Joseph F. Traub Prize for Achievement in Information-Based Complexity. Our adaptive algorithms have been implemented in GAIL [21], which has been used in the graduate Monte Carlo at Illinois Tech. The PIs mentored a number of research students; female students include **YD**, **LJ**, **JL**, **XT**, and Xiaoyang Zhao (MS 2017).

**4.2. NSF CSSI Frameworks 2004571 (Subaward WSU20076). *X-Ion Collisions with a Statistically and Computationally Advanced Program Envelope (X-SCAPE)*, \$696,442, July 2020 – June 2024.** High-energy colliders study the interaction between subatomic particles and environments produced in the collision of protons with protons and with nuclei. This requires an elaborate theoretical, statistical and computational framework. The X-SCAPE (JETSCAPE) collaboration is a multi-disciplinary team of physicists, computer scientists, and statisticians, who are engaged in the construction of such a framework. **SM** is a Duke co-PI in this ongoing project, and is responsible for leading statistical and ML developments.

*4.2.1. Intellectual Merit from Prior NSF Support.* The JETSCAPE collaboration has developed the first open-source simulation framework for the high energy sector of heavy-ion collisions and a Bayesian framework to rigorously compare event generators with experimental data. This has resulted in numerous publications in top physics journals and conferences [7, 8, 36–39, 76, 81, 82, 113, 129, 133] and top statistical / ML journals and conferences [13, 14, 63, 100, 145, 147, 148, 151].

*4.2.2. Broader Impacts from Prior NSF Support.* The primary broader impacts of the X-SCAPE collaboration have been in the training of its graduate students and postdocs, through regular

meetings, collaboration gatherings, and joint projects. The collaboration also influences the training of the wider US nuclear physics workforce through its annual winter school and workshops. **SM** is supporting several students on this project, including two female PhD students and two female undergraduates. **SM** is co-organizing the 2023 Spring Research Conference, and received the 2022 Rosenbluth-Blackwell Award for exceptional career achievements by a junior Bayesian researcher.

## 5. STRENGTHS OF THIS TEAM AND COLLABORATION PLAN

Our team combines senior personnel with diverse backgrounds, career stages, and institutions. We will have regular meetings to share progress and collaborate on papers and software. These will be held both at our own institutions and via video conference between institutions.

**5.1. Senior Personnel.** **FH** has been the lead PI on the GAIL [21] MATLAB software project that contains many of the stopping criteria in **QMCPy**. His expertise is in the numerical analysis of QMC and other multivariate problems as well as theoretically justified adaptive numerical algorithms. As a former editorial board member for major computational mathematics journals, a Fellow of the Institute of Mathematical Statistics, and a co-leader of SAMSI’s program on QMC in 2017-18, **FH** understands the interface between computational mathematics and statistics.

**SM** is an Assistant Professor at Duke, specializing in Bayesian computation, big data analytics, and computer experiments. As an Associate Editor for *Technometrics* (a top engineering statistics journal) and *Data Science in Science*, and the recipient of major awards from the American Statistical Association and the International Society of Bayesian Analysis, **SM** provides expertise on statistical / ML methods and applications. **SM** will lead the activities at Duke and oversee the proposed efforts on Bayesian sampling and big data subsampling.

**YD** is an associate teaching professor at Illinois Tech, who teaches Monte Carlo methods in finance and programming for data analytics. She is experienced in the theory and implementation of QMC and adaptive algorithms, and is one of the main developers of the GAIL project.

**SCTC** serves as a Principal Data Scientist at SAS in the financial risk group. She is a research associate professor at Illinois Tech. **SCTC** has co-led the GAIL and **QMCPy** projects and is an expert in best engineering practices for numerical software. She is a co-winner of the 2011 Society for Industrial and Applied Mathematics Activity Group on Linear Algebra best paper prize.

**5.2. Students.** **AS** began his PhD in applied mathematics at Illinois Tech in Fall 2021 after completing a dual BS/MS degree at Illinois Tech and doing a large majority of the coding of **QMCPy** to date. **CH** is an African American Illinois Tech PhD student who began in Fall 2022. **IJ**, **TT**, **JM** and **KL** are Statistical Science PhD students at Duke University, working on theory and methods for computer experiments, uncertainty quantification, Bayesian computation and big data analytics, with applications to physics and engineering. PhD students supported by this project will work with the senior personnel to tackle major theoretical and methodological challenges. Undergraduate students will focus on new features that can be implemented mostly over the course of a summer. They will be mentored by senior personnel and PhD students.

**5.3. Collaborators.** The responsibilities of these unpaid collaborators will be to discuss research ideas of common interest, involve their research groups as appropriate, and publish significant results. We have identified above some of the projects that they may be involved in. **MM** convinced his company to fund the early development of **QMCPy**. He wanted to spread the advantages of **LD** sampling to the tech industry. **MM** will advise us on the continued development of **QMCPy**, and continue to help us spread the word among his network in the machine learning community. **AO** has engaged with the PIs in conversations about **QMC** for many years and is particularly an expert in randomized **QMC**. **AO** has taken a keen interest in **QMCPy** and used it in his own research. **PR** has wide experience in multi-level methods and uncertainty quantification applications of **QMC**. **CO** is an expert in **PN**. He will provide expertise on the use of Stein discrepancy points and also Bayesian numerics for stopping criteria for **QMC** methods.

## References Cited

- [1] Adams BM, Bohnhoff WJ, Dalbey KR, Ebeida MS, Eddy JP, Eldred MS, Hooper RW, Hough PD, Hu KT, Jakeman JD, Khalil M, Maupin KA, Monschke JA, Ridgway EM, Rushdi AA, Seidl DT, Stephens JA, Swiler LP, Tran A, Winokur JG (2022) Dakota, A Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis: Version 6.16 User's Manual. Sandia National Laboratories
- [2] Ala G, Fasshauer GE, Francomano E, McCourt M, Ganci S (2017) An augmented MFS approach for brain activity reconstruction. *Math Comput Simul* 141:3–15, DOI 10.1016/j.matcom.2016.11.009
- [3] Bachem O, Lucic M, Krause A (2017) Practical coresets constructions for machine learning. arXiv preprint arXiv:170306476
- [4] Ben Abdellah A, L'Ecuyer P, Owen AB, Puchhammer F (2021) Density estimation by randomized quasi-Monte Carlo. *SIAM/ASA J Uncertain Quantif* 9(280-301), DOI 10.1137/19M1259213
- [5] Bingham D (2017) The virtual library of simulation experiments: Test functions and data sets. URL <https://www.sfu.ca/~ssurjano/index.html>
- [6] Bottou L (2010) Large-scale machine learning with stochastic gradient descent. In: *Proceedings of COMPSTAT'2010*, Springer, pp 177–186
- [7] Cao S, Park C, Barbieri RA, Bass SA, Bazow D, Bernhard J, Coleman J, Fries R, Gale C, He Y, Heinz U, Jacak BV, Jacobs PM, Jeon S, Kordell M, Kumar A, Luo T, Majumder A, Nejahi Y, Pablos D, Pang LG, Putschke JH, Roland G, Rose S, Schenke B, Schwiebert L, Shen C, Sirimanna C, Soltz RA, Velicanu D, Vujanovic G, Wang XN, Wolpert RL (2017) Multistage Monte Carlo simulation of jet modification in a static medium. *Phys Rev C* 96:024,909, DOI 10.1103/PhysRevC.96.024909, URL <https://link.aps.org/doi/10.1103/PhysRevC.96.024909>
- [8] Cao S, Chen Y, Coleman J, Mulligan J, Jacobs PM, Soltz RA, Angerami A, Arora R, Bass SA, Cunqueiro L, Dai T, Du L, Ehlers R, Elfner H, Everett D, Fan W, Fries RJ, Gale C, Garza F, He Y, Heffernan M, Heinz U, Jacak BV, Jeon S, Ke W, Kim B, Kordell M, Kumar A, Majumder A, Mak S, McNelis M, Nattrass C, Oliinychenko D, Park C, Paquet JF, Putschke JH, Roland G, Silva A, Schenke B, Schwiebert L, Shen C, Sirimanna C, Tachibana Y, Vujanovic G, Wang XN, Wolpert RL, Xu Y (2021) Determining the jet transport coefficient  $\hat{q}$  from inclusive hadron suppression measurements using Bayesian parameter estimation. *Phys Rev C* 104:024,905, DOI 10.1103/PhysRevC.104.024905, URL <https://link.aps.org/doi/10.1103/PhysRevC.104.024905>
- [9] Chan TM (2006) Faster core-set constructions and data-stream algorithms in fixed dimensions. *Comput Geom* 35(1-2):20–35
- [10] Chang YH, Zhang L, Wang X, Yeh ST, Mak S, Sung CL, Wu CFJ, Yang V (2019) Kernel-smoothed proper orthogonal decomposition-based emulation for spatiotemporally evolving flow dynamics prediction. *AIAA Journal* 57(12):5269–5280
- [11] Chang YH, Wang X, Zhang L, Li Y, Mak S, Wu CFJ, Yang V (2021) Reduced-order modeling for complex flow emulation by common kernel-smoothed proper orthogonal decomposition. *AIAA Journal* 59(9):3291–3303
- [12] Chen J, Du R, Li P, Lyu L (2019) Quasi-Monte Carlo sampling for machine-learning partial differential equations. arXiv preprint arXiv:191101612
- [13] Chen J, Mak S, Joseph VR, Zhang C (2020) Function-on-function kriging, with applications to three-dimensional printing of aortic tissues. *Technometrics* 63:384–395
- [14] Chen J, Mak S, Joseph VR, Zhang C (2022) Adaptive design for gaussian process regression under censoring. *The Annals of Applied Statistics* 16(2):744–764

- [15] Chen WY, Mackey L, Gorham J, Briol FX, Oates C (2018) Stein points. In: Dy J, Krause A (eds) Proceedings of the 35th International Conference on Machine Learning, PMLR, Proceedings of Machine Learning Research, vol 80, pp 844–853, URL <https://proceedings.mlr.press/v80/chen18f.html>
- [16] Chen Y, Welling M, Smola A (2012) Super-samples from kernel herding. arXiv preprint arXiv:12033472
- [17] Chen Z, Mak S, Wu CFJ (2019) A hierarchical expected improvement method for Bayesian optimization. arXiv preprint arXiv:191107285
- [18] Choi SCT, Ding Y, Hickernell FJ, Tong X (2017) Local adaption for approximation and minimization of univariate functions. J Complexity 40:17–33, DOI 10.1016/j.jco.2016.11.005
- [19] Choi SCT, Hickernell FJ, McCourt MJ, Sorokin A (2020) QMCPy blogs. URL <https://qmcpy.wpcomstaging.com>
- [20] Choi SCT, Hickernell FJ, Jagadeeswaran R, McCourt MJ, Sorokin A (2020) QMCPy documentation. URL <https://qmcpy.readthedocs.io/en/latest/>
- [21] Choi SCT, Ding Y, Hickernell FJ, Jiang L, Jiménez Rugama LIA, Li D, Jagadeeswaran R, Tong X, Zhang K, Zhang Y, Zhou X (2021) GAIL: Guaranteed Automatic Integration Library (versions 1.0–2.3.2). MATLAB software, [http://gailgithub.github.io/GAIL\\_Dev/](http://gailgithub.github.io/GAIL_Dev/), DOI 10.5281/zenodo.4018189
- [22] Choi SCT, Hickernell FJ, Jagadeeswaran R, McCourt M, Sorokin A (2022) QMCPy: A quasi-Monte Carlo Python library. DOI 10.5281/zenodo.3964489, URL <https://qmcsoftware.github.io/QMCSoftware/>
- [23] Choi SCT, Hickernell FJ, Jagadeeswaran R, McCourt MJ, Sorokin AG (2022) Quasi-Monte Carlo software. In: Keller A (ed) Monte Carlo and Quasi-Monte Carlo Methods: MCQMC, Oxford, England, August 2020, Springer, Cham, Springer Proceedings in Mathematics and Statistics, pp 23–50, <https://arxiv.org/abs/2102.07833>
- [24] Clancy N, Ding Y, Hamilton C, Hickernell FJ, Zhang Y (2014) The cost of deterministic, adaptive, automatic algorithms: Cones, not balls. J Complexity 30:21–45, DOI 10.1016/j.jco.2013.09.002
- [25] Courtoy A, Huston J, Nadolsky P, Xie K, Yan M, Yuan CP (2022) Parton distributions need representative sampling. preprint, URL <https://arxiv.org/pdf/2205.10444.pdf>
- [26] Darmon Y, Godin M, L’Ecuyer P, Jemel A, Marion P, Munger D (2018) LatNet builder. URL <https://github.com/umontreal-simul/latnetbuilder>
- [27] Davis A, Parno M, Reinartz A, Seelinger L (2022) UQ and model bridge (UM-bridge). URL <https://um-bridge-benchmarks.readthedocs.io/en/docs/>
- [28] Debusschere B (2022) UQ toolkit. URL <https://www.sandia.gov/uqtoolkit/>
- [29] Debusschere B, Najm HN, Pébay PP, Knio OM, Ghanem RG, Maître OPL (2004) Numerical challenges in the use of polynomial chaos representations for stochastic processes. SIAM J Sci Comput pp 698–719, URL [10.1137/S1064827503427741](https://doi.org/10.1137/S1064827503427741)
- [30] Dick J, Pillichshammer F (2010) Digital Nets and Sequences: Discrepancy Theory and Quasi-Monte Carlo Integration. Cambridge University Press, Cambridge
- [31] Dick J, Kuo FY, Sloan IH (2013) High-dimensional integration: the quasi-Monte Carlo way. Acta Numerica 22:133–288
- [32] Dick J, Kritzer P, Pillichshammer F (2022) Lattice Rules: Numerical Integration, Approximation, and Discrepancy. Springer Series in Computational Mathematics, Springer Cham, DOI <https://doi.org/10.1007/978-3-031-09951-9>
- [33] Ding Y (2015) Guaranteed adaptive univariate function approximation. PhD thesis, Illinois Institute of Technology
- [34] Ding Y, Hickernell FJ, Jiménez Rugama LIA (2020) An adaptive algorithm employing continuous linear functionals. In: L’Ecuyer P, Tuffin B (eds) Monte Carlo and Quasi-Monte Carlo Methods: MCQMC, Rennes, France, July 2018, Springer, Cham, Springer Proceedings in

- Mathematics and Statistics, vol 324, pp 161–181, DOI 10.1007/978-3-030-43465-6\\_8
- [35] Ding Y, Hickernell FJ, Kritzer P, Mak S (2020) Adaptive approximation for multivariate linear problems with inputs lying in a cone. In: Hickernell FJ, Kritzer P (eds) *Multivariate Algorithms and Information-Based Complexity*, DeGruyter, Berlin/Boston, pp 109–145, DOI 10.1515/9783110635461-007
  - [36] Ehlers R, Angerami A, Arora R, Bass S, Cao S, Chen Y, Du L, Dai T, Elfner H, Fan W, et al (2022) Bayesian analysis of qgp jet transport using multi-scale modeling applied to inclusive hadron and reconstructed jet data. arXiv preprint arXiv:220807950
  - [37] Everett D, Ke W, Paquet JF, Vujanovic G, Bass SA, Du L, Gale C, Heffernan M, Heinz U, Liyanage D, Luzum M, Majumder A, McNelis M, Shen C, Xu Y, Angerami A, Cao S, Chen Y, Coleman J, Cunqueiro L, Dai T, Ehlers R, Elfner H, Fan W, Fries RJ, Garza F, He Y, Jacak BV, Jacobs PM, Jeon S, Kim B, Kordell M, Kumar A, Mak S, Mulligan J, Nattrass C, Oliinychenko D, Park C, Putschke JH, Roland G, Schenke B, Schwiebert L, Silva A, Sirimanna C, Soltz RA, Tachibana Y, Wang XN, Wolpert RL (2021) Multisystem Bayesian constraints on the transport coefficients of QCD matter. *Phys Rev C* 103:054,904, DOI 10.1103/PhysRevC.103.054904, URL <https://link.aps.org/doi/10.1103/PhysRevC.103.054904>
  - [38] Everett D, Ke W, Paquet JF, Vujanovic G, Bass SA, Du L, Gale C, Heffernan M, Heinz U, Liyanage D, Luzum M, Majumder A, McNelis M, Shen C, Xu Y, Angerami A, Cao S, Chen Y, Coleman J, Cunqueiro L, Dai T, Ehlers R, Elfner H, Fan W, Fries RJ, Garza F, He Y, Jacak BV, Jacobs PM, Jeon S, Kim B, Kordell M, Kumar A, Mak S, Mulligan J, Nattrass C, Oliinychenko D, Park C, Putschke JH, Roland G, Schenke B, Schwiebert L, Silva A, Sirimanna C, Soltz RA, Tachibana Y, Wang XN, Wolpert RL (2021) Phenomenological constraints on the transport properties of QCD matter with data-driven model averaging. *Phys Rev Lett* 126:242,301, DOI 10.1103/PhysRevLett.126.242301, URL <https://link.aps.org/doi/10.1103/PhysRevLett.126.242301>
  - [39] Everett D, Oliinychenko D, Luzum M, Paquet JF, Vujanovic G, Bass S, Du L, Gale C, Heffernan M, Heinz U, et al (2022) Role of bulk viscosity in deuteron production in ultrarelativistic nuclear collisions. *Physical Review C* 106(6):064,901
  - [40] Fan W, Vujanovic G, Bass S, Majumder A, Angerami A, Arora R, Cao S, Chen Y, Dai T, Du L, Ehlers R, Elfner H, Fries RJ, Gale C, He Y, Heffernan M, Heinz U, Jacak BV, Jacobs PM, Jeon S, Ji Y, Kauder K, Kasper L, Ke W, Kelsey M, Kordell M, Kumar A, Latessa J, Lee YJ, Liyanage D, Lopez A, Luzum M, Mak S, Mankolli A, Martin C, Mehryar H, Mengel T, Mulligan J, Nattrass C, Oliinychenko D, Paquet JF, Putschke JH, Roland G, Schenke B, Schwiebert L, Sengupta A, Shen C, Silva A, Sirimanna C, Soeder D, Soltz RA, Soudi I, Staudenmaier J, Strickland M, Tachibana Y, Velkovska J, Wang XN, Wolpert RL, Zhao W (2022) Multi-scale evolution of charmed particles in a nuclear medium. arXiv preprint arXiv:220800983
  - [41] Friedman JH (2002) Stochastic gradient boosting. *Comput Statist & Data Anal* 38(4):367–378
  - [42] Gelman A, Carlin JB, Stern HS, Rubin DB (1995) *Bayesian Data Analysis*. Chapman and Hall/CRC
  - [43] Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2013) *Bayesian Data Analysis*. CRC Texts in Statistical Science, Chapman & Hall
  - [44] Giles MB (2008) Multilevel Monte Carlo path simulation. *Operations Research* 56(3):607–617
  - [45] Giles MB (2015) Multilevel Monte Carlo methods. *Acta Numer* 24:259–328, DOI 10.1017/S096249291500001X
  - [46] Gilquin L, Jiménez Rugama LIA, Arnaud E, Hickernell FJ, Monod H, Prieur C (2017) Iterative construction of replicated designs based on Sobol’ sequences. *C R Math Acad Sci Paris* 355:10–14, DOI 10.1016/j.crma.2016.11.013



- [47] Hahn T (2005) CUBA—a library for multidimensional numerical integration. *Comput Phys Commun* 168:78–95, DOI 10.1016/j.cpc.2005.01.010, URL <https://feynarts.de/cuba/>
- [48] Haji-Latif AL, Nobile F, Tempone R (2016) Multi-index Monte Carlo: when sparsity meets sampling. *Numer Math* 132:767–806, DOI 10.1007/s00211-015-0734-5
- [49] Heinrich S (2001) Multilevel Monte Carlo methods. In: Margenov S, Wasniewski J, Yalamov PY (eds) *Large-Scale Scientific Computing, Third International Conference, LSSC 2001*, Springer-Verlag, Lecture Notes in Computer Science, vol 2179, pp 58–67
- [50] Herrmann L, Schwab C (2020) Multilevel quasi-Monte Carlo uncertainty quantification for advection-diffusion-reaction. In: L’Ecuyer P, Tuffin B (eds) *Monte Carlo and Quasi-Monte Carlo Methods: MCQMC*, Rennes, France, July 2018, Springer, Cham, Springer Proceedings in Mathematics and Statistics, vol 324, pp 31–67, DOI 10.1007/978-3-030-43465-6\\_2
- [51] Hickernell FJ (2020) Quasi-Monte Carlo software. URL <https://mcqmc20.web.ox.ac.uk/programme>
- [52] Hickernell FJ, Sorokin A (2020) Quasi-Monte Carlo software tutorial. URL [https://media.ed.ac.uk/playlist/dedicated/51612401/1\\_0z0wec2z/1\\_2k12mwiw](https://media.ed.ac.uk/playlist/dedicated/51612401/1_0z0wec2z/1_2k12mwiw)
- [53] Hickernell FJ (1998) A generalized discrepancy and quadrature error bound. *Math Comp* 67:299–322, DOI 10.1090/S0025-5718-98-00894-1
- [54] Hickernell FJ (1999) Goodness-of-fit statistics, discrepancies and robust designs. *Statist Probab Lett* 44:73–78, DOI 10.1016/S0167-7152(98)00293-4
- [55] Hickernell FJ (2018) The trio identity for quasi-Monte Carlo error analysis. In: Glynn P, Owen A (eds) *Monte Carlo and Quasi-Monte Carlo Methods: MCQMC*, Stanford, USA, August 2016, Springer-Verlag, Berlin, Springer Proceedings in Mathematics and Statistics, pp 3–27, DOI 10.1007/978-3-319-91436-7
- [56] Hickernell FJ, Jagadeeswaran R (2019) Comment on “Probabilistic integration: A role in statistical computation?”. *Statist Sci* 34:23–28, DOI 10.1214/18-STS660
- [57] Hickernell FJ, Jiménez Rugama LIA (2016) Reliable adaptive cubature using digital sequences. In: Cools R, Nuyens D (eds) *Monte Carlo and Quasi-Monte Carlo Methods: MCQMC*, Leuven, Belgium, April 2014, Springer-Verlag, Berlin, Springer Proceedings in Mathematics and Statistics, vol 163, pp 367–383, arXiv:1410.8615 [math.NA]
- [58] Hickernell FJ, Choi SCT, Jiang L, Jiménez Rugama LIA (2018) Monte Carlo simulation, automatic stopping criteria for. In: Davidian M, Everitt B, Kenett RS, Molenberghs G, Piegorsch W, Ruggeri F (eds) *Wiley StatsRef-Statistics Reference Online*, John Wiley & Sons Ltd., DOI 10.1002/9781118445112.stat08035
- [59] Hickernell FJ, Jiménez Rugama LIA, Li D (2018) Adaptive quasi-Monte Carlo methods for cubature. In: Dick J, Kuo FY, Woźniakowski H (eds) *Contemporary Computational Mathematics — a celebration of the 80th birthday of Ian Sloan*, Springer-Verlag, pp 597–619, DOI 10.1007/978-3-319-72456-0
- [60] Hickernell FJ, Kritzer P, Woźniakowski H (2020) Exponential tractability of linear tensor product problems. In: *2018 MATRIX Annals*, MATRIX Book Series, Springer Nature Switzerland AG, pp 61–78
- [61] Hjort NL, Holmes C, Müller P, Walker SG (2010) *Bayesian Nonparametrics*, vol 28. Cambridge University Press
- [62] Hoffman MD, Gelman A, et al (2014) The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J Mach Learn Res* 15(1):1593–1623
- [63] Huang C, Joseph VR, Mak S (2022) Population quasi-monte carlo. *Journal of Computational and Graphical Statistics* pp 1–14
- [64] Huggins J, Campbell T, Broderick T (2016) Coresets for scalable Bayesian logistic regression. In: *Advances in Neural Information Processing Systems*, pp 4080–4088
- [65] Jagadeeswaran R, Hickernell FJ (2019) Fast automatic Bayesian cubature using lattice sampling. *Stat Comput* 29:1215–1229, DOI 10.1007/s11222-019-09895-9

- [66] Jagadeeswaran R, Hickernell FJ (2022) Fast automatic Bayesian cubature using Sobol' sampling. In: Botev Z, Keller A, Lemieux C, Tuffin B (eds) *Advances in Modeling and Simulation: Festschrift in Honour of Pierre L'Ecuyer*, Springer, Cham, pp 301–318, DOI 10.1007/978-3-031-10193-9\_15
- [67] Ji Y, Mak S, Soeder D, Paquet J (2021) A graphical Gaussian process model for multi-fidelity emulation of expensive computer codes. arXiv preprint arXiv:210800306
- [68] Ji Y, Yuchi HS, Soeder D, Paquet JF, Bass SA, Joseph VR, Wu C, Mak S (2022) Multi-stage multi-fidelity gaussian process modeling, with application to heavy-ion collisions. arXiv preprint arXiv:220913748
- [69] Jiménez Rugama LIA, Gilquin L (2018) Reliable error estimation for Sobol' indices. *Statistics and Computing* 28:725–738, DOI 10.1007/s11222-017-9759-1
- [70] Jiménez Rugama LIA, Hickernell FJ (2016) Adaptive multidimensional integration based on rank-1 lattices. In: Cools R, Nuyens D (eds) *Monte Carlo and Quasi-Monte Carlo Methods: MCQMC*, Leuven, Belgium, April 2014, Springer-Verlag, Berlin, Springer Proceedings in Mathematics and Statistics, vol 163, pp 407–422, arXiv:1411.1966
- [71] Jin C, Netrapalli P, Jordan MI (2018) Accelerated gradient descent escapes saddle points faster than gradient descent. In: *Conference On Learning Theory*, PMLR, pp 1042–1085
- [72] Johnson T, Fasshauer GE, Hickernell FJ (2018+) Characterizing reproducing kernel Hilbert spaces. In preparation
- [73] Jones DR, Schonlau M, Welch WJ (1998) Efficient global optimization of expensive black-box functions. *Journal of Global Optimization* 13(4):455–492
- [74] Joseph VR, Mak S (2021) Supervised compression of big data. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 14(3):217–229
- [75] Joseph VR, Dasgupta T, Tuo R, Wu CJ (2015) Sequential exploration of complex surfaces using minimum energy designs. *Technometrics* 57(1):64–74
- [76] Kauder K (2019) JETSCAPE v1.0 quickstart guide. *Nuclear Phys A* 982:615–618
- [77] Kaufman CG, Bingham D, Habib S, Heitmann K, Frieman JA (2011) Efficient emulators of computer experiments using compactly supported correlation functions, with an application to cosmology. *The Annals of Applied Statistics* 5(4):2470–2492
- [78] Keller A (2013) Quasi-Monte Carlo image synthesis in a nutshell. In: Dick J, Kuo FY, Peters GW, Sloan IH (eds) *Monte Carlo and Quasi-Monte Carlo Methods 2012*, Springer-Verlag, Berlin, Springer Proceedings in Mathematics and Statistics, vol 65, pp 213–249, DOI 10.1007/978-3-642-41095-6
- [79] Krishna A, Mak S, Joseph R (2019) Distributional clustering: A distribution-preserving clustering method. arXiv preprint arXiv:191105940
- [80] Kucherenko S (2022) BRODA. URL <https://www.broda.co.uk/index.html>
- [81] Kumar A, Tachibana Y, Pablos D, Sirimanna C, Fries R, Angerami A, Bass S, Cao S, Coleman J, Cunqueiro L (2019) The JETSCAPE framework: p+p results. arXiv preprint arXiv:191005481
- [82] Kumar A, Tachibana Y, Sirimanna C, Vujanovic G, Cao S, Majumder A, Chen Y, Du L, Ehlers R, Everett D, et al (2022) Inclusive jet and hadron suppression in a multi-stage approach. arXiv preprint arXiv:220401163
- [83] Li D (2016) Reliable quasi-Monte Carlo with control variates. Master's thesis, Illinois Institute of Technology
- [84] Li Y, Wang X, Mak S, Yeh ST, Lin LH, Wu CFJ, Yang V (2017) A two-stage transfer function identification methodology and its applications to bi-swirl injectors. In: 53rd AIAA/SAE/ASEE Joint Propulsion Conference, p 4933
- [85] Li Y, Wang X, Mak S, Sung CL, Wu J, Yang V (2018) Uncertainty quantification of flame transfer function under a Bayesian framework. In: 2018 AIAA Aerospace Sciences Meeting, p 1187

- [86] Link WA, Eaton MJ (2012) On thinning of chains in MCMC. *Methods in Ecology and Evolution* 3(1):112–115
- [87] Liu J (2018) Adaptive quadrature with a general error criterion. Master’s thesis, Illinois Institute of Technology
- [88] Liyanage D, Ji Y, Everett D, Heffernan M, Heinz U, Mak S, Paquet JF (2022) Efficient emulation of relativistic heavy ion collisions with transfer learning. *Physical Review C* 105(3):034,910
- [89] Lopes A (2020) Recreating Big Bang matter on Earth. <https://phys.org/news/2020-12-recreating-big-earth.html>, accessed: 2022-12-27
- [90] Ma P, Mahoney MW, Yu B (2015) A statistical perspective on algorithmic leveraging. *J Mach Learn Res* 16(1):861–911
- [91] Mahoney MW (2011) Randomized algorithms for matrices and data. arXiv preprint arXiv:11045557
- [92] Mak S, Joseph VR (2017) Projected support points, with application to optimal MCMC reduction. arXiv preprint arXiv:170806897
- [93] Mak S, Joseph VR (2018) Minimax and minimax projection designs using clustering. *J Comput and Graph Statist* 27(1):166–178
- [94] Mak S, Joseph VR (2018) Support points. *Ann Statist* 46(6A):2562–2592
- [95] Mak S, Wu CFJ (2019) Analysis-of-marginal-tail-means (ATM): A robust method for discrete black-box optimization. *Technometrics* 61(4):545–559
- [96] Mak S, Xie Y (2018) Maximum entropy low-rank matrix recovery. *IEEE J Sel Topics in Signal Process* 12(5):886–901
- [97] Mak S, Bingham D, Lu Y (2016) A regional compound Poisson process for hurricane and tropical storm damage. *J R Stat Soc Ser C Appl Stat* 65(5):677–703
- [98] Mak S, Sung CL, Wang X, Yeh ST, Chang YH, Joseph VR, Yang V, Wu CFJ (2018) An efficient surrogate model for emulation and physics extraction of large eddy simulations. *J Amer Statist Assoc* 113(524):1443–1456
- [99] Mak S, Yuchi HS, Xie Y (2021) Information-guided sampling for low-rank matrix completion. In: *ICML Workshop on Information Theoretic Methods for Rigorous, Responsible, and Reliable Machine Learning*
- [100] Mak S, Zhou Y, Hoang L, Wu C (2022) Tsec: a framework for online experimentation under experimental constraints. *Technometrics*, to appear
- [101] Marelli S, Sudret B (2014) UQLab: A framework for uncertainty quantification in MATLAB. In: *The 2nd International Conference on Vulnerability and Risk Analysis and Management (ICVRAM 2014)*, ASCE Library, pp 2554–2563, URL <https://www.uqlab.com>
- [102] Martin B, Elsherbeni A, Fasshauer GE, Hadi M (2018) Improved FDTD method around dielectric and pec interfaces using RBF-FD techniques. In: *2018 International Applied Computational Electromagnetics Society Symposium (ACES)*, DOI 10.23919/ROPACES.2018.8364209
- [103] McCourt M, Fasshauer GE (2017) Stable likelihood computation for Gaussian random fields. In: *Pesenson I, Le Gia Q, Mayeli A, Mhaskar H, Zhou DX (eds) Recent Applications of Harmonic Analysis to Function Spaces, Differential Equations, and Data Science: Novel Methods in Harmonic Analysis*, vol 2, Basel, Birkäuser, pp 917–942
- [104] McCourt MJ, Fasshauer GE, Kozak D (2018) A nonstationary designer space-time kernel. In: *32nd Conference on Neural Information Processing Systems (NIPS 2018)*
- [105] Mishra PK, Nath SK, Sen MK, Fasshauer GE (2018) Hybrid Gaussian-cubic radial basis function for scattered data interpolation. *Comput Geosciences* 22:1203–1218, DOI 10.1007/s10596-018-9747-3
- [106] Mishra PK, Fasshauer GE, Sen MK, Ling L (2019) A stabilized radial basis-finite difference (RBF-FD) method with hybrid kernels. *Comput Math Appl* 77:2354–2368, DOI 10.1016/j.

- camwa.2018.12.027
- [107] Niederreiter H (1992) Random Number Generation and Quasi-Monte Carlo Methods. CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM, Philadelphia
  - [108] Nocedal J, Wright S (2006) Numerical Optimization. Springer Science & Business Media
  - [109] Nuyens D (2017) Magic point shop. URL <https://people.cs.kuleuven.be/~dirk.nuyens/qmc-generators/>
  - [110] OpenTURNS Developers (2020) An open source initiative for the Treatment of Uncertainties, Risks 'N Statistics. URL <http://www.openturns.org>
  - [111] Owen A (2003) Data squashing by empirical likelihood. Data Min Knowl Disc 7(1):101–113
  - [112] Owen AB (2020) On dropping the first Sobol' point. arXiv preprint arXiv:200808051
  - [113] Park C, Angerami A, Bass S, Cao S, Coleman J, Cunqueiro L, Dai T, Du L, Elfner H, Everett D (2019) Multi-stage jet evolution through QGP using the JETSCAPE framework: inclusive jets, correlations and leading hadrons. arXiv preprint arXiv:190205934
  - [114] Parno M, Davis A, Seelinger L, Marzouk Y (2014) MIT uncertainty quantification (muq) library. URL [https://mituq.bitbucket.io/source/\\_site/index.html](https://mituq.bitbucket.io/source/_site/index.html)
  - [115] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S (2019) PyTorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems 32:8026–8037
  - [116] PyTorch Developers (2020) PyTorch discussion on Sobol' sequence implementation. URL <https://github.com/pytorch/pytorch/issues/32047>
  - [117] Quiroz M, Kohn R, Villani M, Tran MN (2018) Speeding up MCMC by efficient data sub-sampling. Journal of the American Statistical Association 114:831–843
  - [118] Rashidinia J, Fasshauer GE, Khasi M (2016) A stable method for the evaluation of Gaussian radial basis function solutions of interpolation and collocation problems. Comput Math Appl 72(1):178–193, DOI 10.1016/j.camwa.2016.04.048
  - [119] Rashidinia J, Khasi M, Fasshauer GE (2018) A stable Gaussian radial basis function method for solving nonlinear unsteady convection-diffusion-reaction equations. Comput Math Appl 75:1831–1850, DOI 10.1016/j.camwa.2017.12.007
  - [120] Rosenthal JS (2017) Simple confidence intervals for MCMC without CLTs. Electronic Journal of Statistics 11(1):211–214
  - [121] Saltelli A (2002) Making best use of model evaluations to compute sensitivity indices. Comput Phys Commun 145:280–297, DOI 10.1016/S0010-4655(02)00280-1
  - [122] Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, Saisana M, Tarantola S (2008) Global Sensitivity Analysis: The Primer. John Wiley & Sons Ltd., Chichester, England
  - [123] Santner TJ, Williams BJ, Notz WI, Williams BJ (2003) The Design and Analysis of Computer Experiments. Springer
  - [124] SciPy Developers (2020) scipy discussion of Sobol' sequence implementation. URL <https://github.com/scipy/scipy/pull/10844>
  - [125] Sloan IH, Joe S (1994) Lattice Methods for Multiple Integration. Oxford University Press, Oxford
  - [126] SmartUQ Developers (2020) SmartUQ. URL <https://www.smartuq.com/resources/uncertainty-quantification/>
  - [127] Smith RC (2014) Uncertainty Quantification: theory, implementation, and applications. SIAM, Philadelphia
  - [128] Sobol' IM (2001) Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Math Comput Simul 55(1-3):271–280
  - [129] Soltz R (2018) Bayesian extraction of  $q$  with multi-stage jet evolution approach. In: Proceedings, 9th International Conference on Hard and Electromagnetic Probes of High-Energy

- Nuclear Collisions: Hard Probes, pp 1–5
- [130] Sorokin AG, Jagadeeswaran R (2022+) Monte Carlo for vector functions of integrals. In: Hinrichs A, Kritzer P, Pillichshammer F (eds) Monte Carlo and Quasi-Monte Carlo Methods: MCQMC, Linz, Austria, July 2022, Springer, Cham, Springer Proceedings in Mathematics and Statistics, in preparation for submission for publication
  - [131] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 15(1):1929–1958
  - [132] Sung CL, Ji Y, Tang T, Mak S (2022) Stacking designs: designing multi-fidelity computer experiments with confidence. arXiv preprint arXiv:221100268
  - [133] Tachibana Y, Angerami A, Bass S, Cao S, Coleman J, Cunqueiro L, Dai T, Du L, Elfner H, Everett D (2018) Jet substructure modification in a QGP from a multi-scale description of jet evolution with JETSCAPE. arXiv preprint arXiv:181206366
  - [134] TF Quant Finance Contributors (2021) TF Quant Finance: TensorFlow based Quant Finance library. URL <https://github.com/google/tf-quant-finance>
  - [135] The MathWorks, Inc (2020) MATLAB R2020b. Natick, MA
  - [136] The Numerical Algorithms Group (2021) The NAG Library. Oxford, Mark 27 edn
  - [137] Tong X (2014) A guaranteed, adaptive, automatic algorithm for univariate function minimization. Master’s thesis, Illinois Institute of Technology
  - [138] Vihola M (2012) Robust adaptive metropolis algorithm with coerced acceptance rate. *Statistics and Computing* 22(5):997–1008
  - [139] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat I, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, Vijaykumar A, Bardelli AP, Rothberg A, Hilboll A, Kloeckner A, Scopatz A, Lee A, Rokem A, Woods CN, Fulton C, Masson C, Häggström C, Fitzgerald C, Nicholson DA, Hagen DR, Pasechnik DV, Olivetti E, Martin E, Wieser E, Silva F, Lenders F, Wilhelm F, Young G, Price GA, Ingold GL, Allen GE, Lee GR, Audren H, Probst I, Dietrich JP, Silterra J, Webber JT, Slavič J, Nothman J, Buchner J, Kulick J, Schönberger JL, de Miranda Cardoso J, Reimer J, Harrington J, Rodríguez JLC, Nunez-Iglesias J, Kuczynski J, Tritz K, Thoma M, Newville M, Kümmerer M, Bolingbroke M, Tartre M, Pak M, Smith NJ, Nowaczyk N, Shebanov N, Pavlyk O, Brodtkorb PA, Lee P, McGibbon RT, Feldbauer R, Lewis S, Tygier S, Sievert S, Vigna S, Peterson S, More S, Pudlik T, Oshima T, Pingel TJ, Robitaille TP, Spura T, Jones TR, Cera T, Leslie T, Zito T, Krauss T, Upadhyay U, Halchenko YO, Vázquez-Baeza Y, Contributors (2020) Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods* 17(3):261–272
  - [140] Wahls S, Poor HV (2015) Fast numerical nonlinear Fourier transforms. *IEEE Transactions on Information Theory* 61(12):6957–6974
  - [141] Wand MP, Jones MC (1995) Kernel Smoothing. Chapman & Hall, New York
  - [142] Wang H, Ma Y (2021) Optimal subsampling for quantile regression in big data. *Biometrika* 108(1):99–112
  - [143] Wang H, Zhu R, Ma P (2018) Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association* 113(522):829–844
  - [144] Wang H, Xie L, Cuozzo A, Mak S, Xie Y (2020) Uncertainty quantification for inferring Hawkes networks. In: *Advances in Neural Information Processing Systems*
  - [145] Wang H, Xie L, Xie Y, Cuozzo A, Mak S (2022) Sequential change-point detection for mutually exciting point processes over networks. *Technometrics*, to appear
  - [146] Yeh ST, Wang X, Sung CL, Mak S, Chang YH, Zhang L, Wu CJ, Yang V (2018) Common proper orthogonal decomposition-based spatiotemporal emulator for design exploration.



- AIAA Journal 56(6):2429–2442
- [147] Yuchi HS, Mak S, Xie Y (2022) Bayesian uncertainty quantification for low-rank matrix completion. *Bayesian Analysis* , to appear
  - [148] Zhang R, Mak S, Dunson D (2022) Gaussian process subspace prediction for model reduction. *SIAM Journal on Scientific Computing* 44(3):A1428–A1449
  - [149] Zhang Y (2018) Guaranteed, adaptive, automatic algorithms for univariate integration: Methods, costs and implementation. PhD thesis, Illinois Institute of Technology
  - [150] Zhao X (2017) Simulating the Heston model via the QE method with a specified error tolerance. Master’s thesis, Illinois Institute of Technology
  - [151] Zheng X, Mak S, Xie Y (2021) Online high-dimensional change-point detection using topological data analysis. *Technometrics* , to appear
  - [152] Zhou X (2015) Function approximation with kernel methods. PhD thesis, Illinois Institute of Technology
  - [153] Zhou X, Hickernell FJ (2016) Tractability of the radial function approximation problem with kernels of a product form. In: Cools R, Nuyens D (eds) *Monte Carlo and Quasi-Monte Carlo Methods: MCQMC*, Leuven, Belgium, April 2014, Springer-Verlag, Berlin, Springer Proceedings in Mathematics and Statistics, vol 163, pp 583–598, arXiv:1411.0790 [math.NA]