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[1]) (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 [79] and rocket propulsion [101, 150] 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 [38, 39, 69, 91] 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 $\boldsymbol{t}$) and hence shed light on the origins of matter. But such simulations are very costly: a single

[1]Acronyms and initials of personnel contain hyperlinks to their full names.

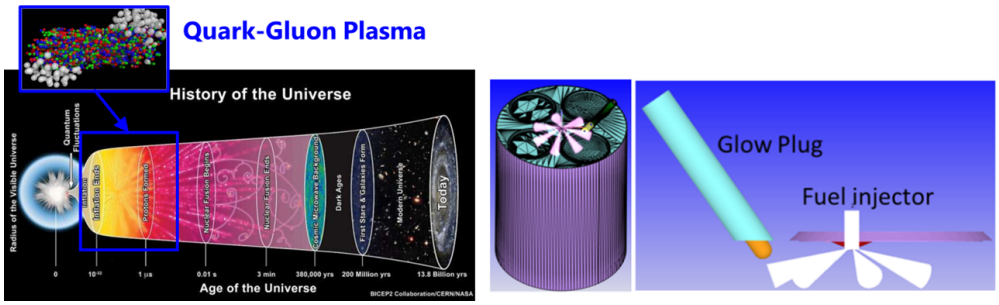

Figure 1. (Left) Visualizing the QGP shortly after the Big Bang [92], which became the building blocks of matter in the Universe. (Right) A schematic of the UAV metal engine: fuel is injected at seven nozzles, then ignited via a glow plug.

run at parameter $\boldsymbol{t}$, yielding simulation output $g(\boldsymbol{t})$, can take on the order of thousands of CPU hours [38]. This inverse problem may require thousands of simulation runs at different choices of $\boldsymbol{t}$, 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 $g(\boldsymbol{t})$ 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 which 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 $\boldsymbol{T}_1, \ldots, \boldsymbol{T}_n$ 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 [110], 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 $\boldsymbol{T}$ with known distribution:

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

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

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

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

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

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

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

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

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, μ, often cannot be evaluated by analytic means, but it may be estimated by the sample mean, $\hat{\mu}_n$, i.e.,

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

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

$$|\mu - \hat{\mu}_n| \le \varepsilon \qquad \text{with high confidence.} \tag{4}$$

1.2.2. Low Discrepancy Sampling. Choosing $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$, for evaluating the sample mean in (3) to be *independent and identically distributed* (IID) yields a root mean square error of $\sqrt{\mathbb{E}[|\mu - \hat{\mu}_n|^2]} = \mathrm{std}(f(\boldsymbol{X}))n^{-1/2}$, which is independent of the dimension, d, but slowly vanishing as $n \to \infty$. 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, $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots \overset{\text{LD}}{\sim} \mathcal{U}[0,1]^d$, which has an error bound of [56, 111]

$$|\mu - \hat{\mu}_n| \le D(\{\boldsymbol{X}_i\}_{i=1}^n)\,\|f - \mu\|_{\mathcal{F}}\,. \tag{5}$$

The discrepancy, $D(\{\boldsymbol{X}_i\}_{i=1}^n)$, corresponds to the norm of the cubature error functional [55] 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 δ 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 [33, 111, 129] and digital nets [31, 111]. 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 [82], CUBA [49] MATLAB [139], NAG [140], PyTorch [119], SciPy [143], TensorFlow [138], and uncertainty quantification libraries such as Dakota [1], MUQ [118], UQLab [104], and UQTk [29, 30]. Efforts to identify better LD sequence generators include LatNet Builder [27] and the Magic Point Shop [113].

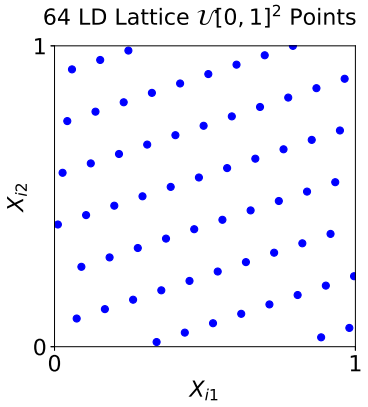

Figure 2. LD lattice points, which have fewer gaps and clusters of points than either the IID or grid points.

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 [54], and a blog [19].

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, $f(\boldsymbol{X}_1), f(\boldsymbol{X}_2), \ldots$. 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, $\hat{\mu}_{nN} = (\hat{\mu}_n^{(1)} + \cdots \hat{\mu}_n^{(N)})/N$, 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 $f(\boldsymbol{X}_1), f(\boldsymbol{X}_2), \ldots$. The first kind determines a bound on $|\mu - \mu_n|$ by inferring the roughness of f from the decay of the discrete Fourier coefficients [59, 61, 72]. 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 [58, 67, 68]. 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 [43] is a popular statistical framework for tackling scientific problems. Here, parameters of interest are sampled from a "posterior" probability distribution ϱ, which captures both the 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 [38]. 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 $Y_{\boldsymbol{\eta}} = f_{\boldsymbol{\eta}}(\boldsymbol{X}_{\boldsymbol{\eta}})$ is parameterized by a *fidelity* parameter $\boldsymbol{\eta}$, and the desired quantity of interest is the limiting mean $\mu_\infty = \lim_{\boldsymbol{\eta}\to\boldsymbol{\eta}_\infty} \mathbb{E}(Y_{\boldsymbol{\eta}})$. In our heavy-ion application, $f_{\boldsymbol{\eta}}(\boldsymbol{X}_{\boldsymbol{\eta}})$ may represent a observable simulated from a complex partial differential equation (PDE) system modeling the heavy-ion collision, with random coefficients $\boldsymbol{X}_{\boldsymbol{\eta}}$ 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 μ_∞ can be prohibitively costly.

Multifidelity (also known as multi-level or multi-index) methods [45, 50, 51] choose a sequence of fidelity parameters, $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots$, and write the desired quantity as a telescoping sum:

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

The sequence of fidelity parameters is chosen so that the cost of evaluating $f_{\boldsymbol{\eta}_l}(\boldsymbol{X}_{\boldsymbol{\eta}_l})$ 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 addressed 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, 135], 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 $Y = g(\boldsymbol{T})$, 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 $\boldsymbol{T}$, 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 [120] due to a lack of double precision. Lluís Antoni Jiménez Rugama (LlAJR), a

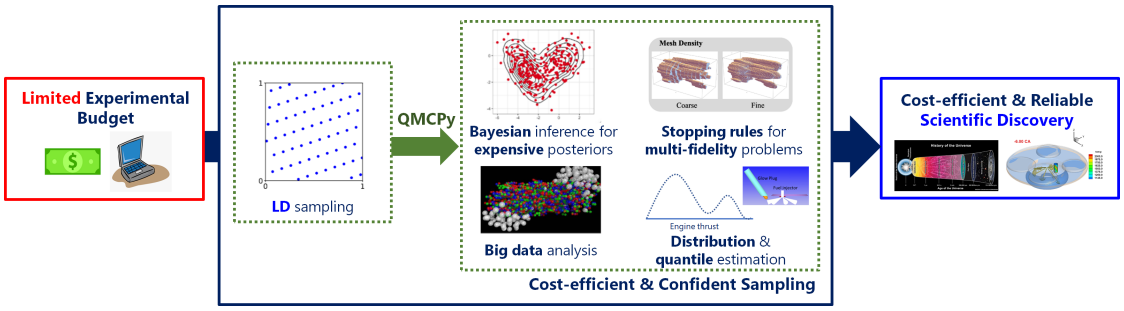

Figure 3. Project workflow: given a limited budget, the proposed tasks (Sect. 2) extend LD sampling to the complex settings motivated by our applications, providing a toolbox for cost-efficient and reliable scientific discovery. QMCPy (Sect. 3.2) serves as an open-source software package that disseminates 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 [120] and SciPy [128] 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 [116]. 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, $f(\boldsymbol{X}_i)$, 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 [43] rely on MCMC sampling to explore the posterior distribution F, which captures information on model parameters (see PI SM's work [65, 100, 103]). 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 [89]. Existing samplers for such problems thus be prohibitively *costly* [77], 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* $D_K(\{\boldsymbol{T}_i\}_{i=1}^n, F)$ [56], which measures the difference between the empirical distribution of $\{\boldsymbol{T}_i\}_{i=1}^n$ 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 $\int K(\boldsymbol{t},\cdot)\,\mathrm{d}F(\boldsymbol{t})$, which is unattainable for complex posteriors F. To address this, [15] proposed the "Steinized" kernel:

$$K_{\mathrm{ST}}(\boldsymbol{t},\boldsymbol{x}) = \nabla_{\boldsymbol{t}}\cdot\nabla_{\boldsymbol{x}}K(\boldsymbol{t},\boldsymbol{x}) + \nabla_{\boldsymbol{t}}K(\boldsymbol{t},\boldsymbol{x})\cdot\nabla_{\boldsymbol{x}}\log\mathrm{d}F(\boldsymbol{x}) + \nabla_{\boldsymbol{x}}K(\boldsymbol{t},\boldsymbol{x})\cdot\nabla_{\boldsymbol{t}}\log\mathrm{d}F(\boldsymbol{t}) + K(\boldsymbol{t},\boldsymbol{x})\nabla_{\boldsymbol{t}}\log\mathrm{d}F(\boldsymbol{t})\cdot\nabla_{\boldsymbol{x}}\log\mathrm{d}F(\boldsymbol{x}), \tag{6}$$

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

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

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

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

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 $\nabla_{\boldsymbol{t}} \log \mathrm{d}F(\boldsymbol{t})$. 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 [127] on the expensive objective functions KSD_n. Consider first (ESP_n) for a given n. Suppose the posterior in the form of its score function $\nabla_{\boldsymbol{t}} \log \mathrm{d}F(\boldsymbol{t})$ has already been at evaluated at M_n candidate points $\{\boldsymbol{T}_j\}_{j=1}^{M_n}$, yielding objective evaluations $\mathcal{D}_n = \{\mathrm{KSD}_n(\boldsymbol{T}_j)\}_{j=1}^{M_n}$ via (7). Under a GP prior on $\mathrm{KSD}_n(\cdot)$, the posterior distribution of $\mathrm{KSD}_n(\boldsymbol{t})$ can be shown to be $\mathrm{KSD}_n(\boldsymbol{t})|\mathcal{D}_n \sim \mathcal{N}\{\mu_n(\boldsymbol{t}), \sigma_n^2(\boldsymbol{t})\}$; specific forms for $\mu_n(\boldsymbol{t})$ and $\sigma_n^2(\boldsymbol{t})$ can be found in [127]. Using this and following the literature on Bayesian optimization (e.g., [75] and work from PI SM [17]), the *expected improvement* in objective KSD_n from evaluating the posterior at a new point $\boldsymbol{t}$ takes the closed-form expression:

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

where $\mathrm{KSD}_{n,\min}$ is the best observed KSD_n value. We thus wish to evaluate the expensive posterior at the point that maximizes $\mathrm{EIKSD}_n(\boldsymbol{t})$. 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 KSD_n is then taken as the next sample $\boldsymbol{T}_n^*$ in (ESP_n).

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

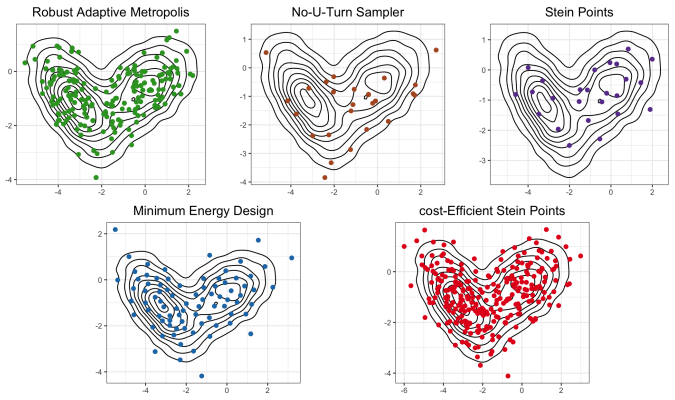

Figure 4. Visualizing the sampled points on a 2D two-mixture normal distribution, using four existing posterior samplers and the proposed ESPs. All samplers are limited to 500 posterior evaluations.

To demonstrate the cost-efficiency of ESPs, Fig. 4 compares several state-of-the-art samplers (the robust adaptive Metropolis sampler [142], the No-U-Turn sampler [64], Stein points [15], minimum energy designs [77]) 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,

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 ε 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 [47] with Bayesian non-parametrics theory [63]. 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, 152, 155] and [69, 70, 91]).

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 [32]. 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 $\boldsymbol{T}_n^*$ 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 $n \to \infty$. With this, we will develop confidence intervals for the proposed randomized ESPs (following [124]), 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 [37–41, 84, 91]) 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 $Y = f(\boldsymbol{X})$ 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 $\boldsymbol{X}$ as inputs, and $\boldsymbol{\eta}$ may capture the spatial and temporal mesh size of the simulator (see work on this by PI SM [69, 70, 91]).

As mentioned in Sect. 1.3, approximating the true solution by a single high fidelity expectation can be prohibitively costly. Rather, multifidelity methods [45, 50, 51]—an active research area—consider a sequence of problems, $\{\mu_l = \mathbb{E}[f_{\boldsymbol{\eta}_l}(\boldsymbol{X}_{\boldsymbol{\eta}_l})]\}_{l=0}^L$, where the fidelity increases with l as does the computational cost of evaluating an instance $f_{\boldsymbol{\eta}_l}(\boldsymbol{X}_{\boldsymbol{\eta}_l,i})$. One can efficiently approximate μ 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 [46, 47], including recent work by PI SM [136]) 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, $C(\boldsymbol{\mu})$, [61, 134]. For example, Bayesian posterior means, [44], can be written as the ratio of two expectations, $C(\boldsymbol{\mu}) = \mu_2/\mu_1$. Sensitivity indices [125, 126, 132] also involve the computation of more than one expectation. However, in both these cases, the underlying random vector, $\boldsymbol{X}$, is the same for all expectations, $\mu_1, \mu_2, \ldots$, and only the functions f_l defining the expectations are different.

For multifidelity problems each μ_l depends on a different $\boldsymbol{X}_{\boldsymbol{\eta}_l} \sim \mathcal{U}[0,1]^{d_l}$ 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 [69, 70]). We will further explore broader applications in geophysical applications, which the PIs have close ongoing collaborations (see [151]). 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 [42].

Consider SGD, which minimizes the loss $L(\theta; \mathcal{T}) = N^{-1} \sum_{m=1}^{N} l(\theta; \boldsymbol{T}_m)$ over model parameters $\theta \in \mathbb{R}^q$, where $\mathcal{T} = \{\boldsymbol{T}_m\}_{m=1}^N \subset \mathbb{R}^d$ is the large training data. Standard gradient descent [112] is impractical here, since they require evaluation of the full gradient $N^{-1} \sum_{m=1}^{N} \nabla_\theta l(\theta; \boldsymbol{T}_m)$, 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:

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

where ζ 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 [135].

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 θ^*. For small subsample sizes n (as necessitated from our cost-constrained

setting), SGD can thus return estimates *very far* from θ^*. 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 [115]), and encompasses work on leverage-score subsampling [93], coresets [3, 9, 66], experimental design [146, 147], and work by PI SM [81, 95–97]. 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 $\mathcal{T} = \{\boldsymbol{T}_m\}_{m=1}^N$ (the "big data") be any set of points on $[0,1]^d$, and suppose the feasible space Θ 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 θ^*.*

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 θ^*. 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 $\mathcal{T} = \{\boldsymbol{T}_m\}_{m=1}^N$:

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

where i is the imaginary number, and $\lambda_{\boldsymbol{k}} = \prod_{j=1}^d \max(1, 2\pi |k_j|)^{-2}$. Here, the coefficients $b_{\boldsymbol{k}}$ can be efficiently computed via non-linear fast Fourier transform [144]. The data kernel K_{data} has two nice properties. We can show that the subsample $\mathcal{X}_s$ minimizing the kernel discrepancy with K_{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 [73] and randomized algorithms [94].

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 $Y = f(\boldsymbol{X})$. As opposed to estimating a scalar, μ, we now need to approximate functions, F, ϱ, 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_{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.

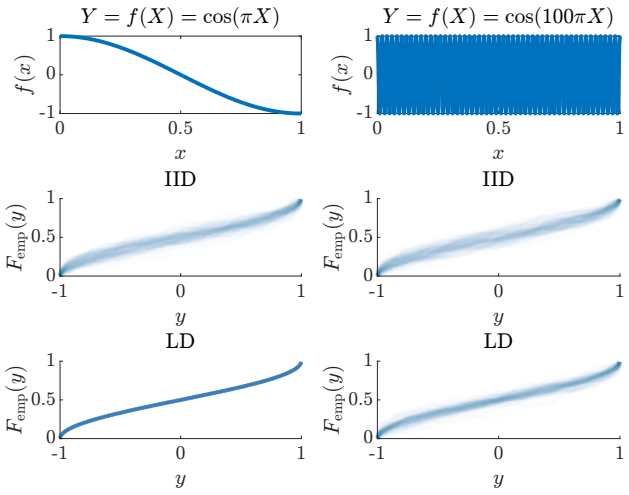

Figure 5. Comparison of the empirical distribution functions of $Y = f(X)$ using $n = 64$ IID and LD X_i and for two definitions of f. The empirical distributions generated by LD points are closer to the true distribution than IID sampling.

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 [32]. In QMC, effective dimension is defined roughly as the number of coordinates of the input vector $\boldsymbol{X}$ that make substantial contributions to $f(\boldsymbol{X})$. One expects that small effective dimension plays an important role in approximating ϱ with LD sequences, just as it is in approximating $\mu = \mathbb{E}[f(\boldsymbol{X})]$ with LD sequences [32]. This is critical in our UAV problem, where there are many design and control parameters for engine operation.

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 [145]. 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 μ, 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 [130, 131]. 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 [52] for a convincing application in fluid flows). PI SM has ongoing multi-disciplinary collaborations on UQ in aerospace engineering [10, 11, 87, 88, 101, 150], nuclear physics [8, 37–39, 69, 70, 84, 91], astrophysics [99, 102, 155] and neuroscience [148, 149], 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, 42, 121], QMC will undoubtedly have a significant impact in this area (see [80] 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 [76, 97, 102].

3.1.3. High Energy Physics. Physicists have begun to recognize the advantages of LD sampling for generating parton distributions [26]. 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, 37–39, 69, 70, 84, 91]).

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 [67] 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, 98]. 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) [28] 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, 53] prompted a vigorous discussion on the PyTorch issues site [120] that migrated to the SciPy issues site [128]. 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 [116]. In these discussions, it was pointed out that UQLab [104], OpenTurns [114], 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 [120] This was due to a lack of double precision, as discovered by MM. LlAJR 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), LlAJR (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, 34, 35, 48, 57–62, 67, 71, 72, 74, 86, 90, 105–109, 122, 123, 153, 154, 156, 157].

4.1.1. Intellectual Merit from Prior NSF Support. FH, SCTC, YD, LlAJR, DL, JR, XT, YZ, and developed several adaptive algorithms for univariate integration, function approximation, and optimization [18, 24, 34, 141, 153], multivariate integration [59, 61, 72], and multivariate function approximation for Banach spaces, $\mathcal{F}$, defined by series representations [35, 36].

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 algorithmshave been 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, 37–40, 78, 83, 84, 117, 133, 137] and top statistical / ML journals and conferences [13, 14, 65, 103, 149, 151, 152, 155].

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 identifed 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 quantifciation 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.