SIAM J. NUMER. ANAL.
Vol. 61, No. 2, pp. 495–514

PREINTEGRATION VIA ACTIVE SUBSPACE*

SIFAN LIU† AND ART B. OWEN†

Abstract. Preintegration is an extension of conditional Monte Carlo to quasi–Monte Carlo and randomized quasi–Monte Carlo. Conditioning can reduce but not increase the variance in Monte Carlo. For quasi–Monte Carlo it can bring about improved regularity of the integrand with potentially greatly improved accuracy. We show theoretically that, just as in Monte Carlo, preintegration can reduce but not increase the variance when one uses scrambled net integration. Preintegration is ordinarily done by integrating out one of the input variables to a function. In the common case of a Gaussian integral one can also preintegrate over any linear combination of variables. For continuous functions that are differentiable almost everywhere, we propose to choose the linear combination by the first principal component in an active subspace decomposition. We show that the lead eigenvector in an active subspace decomposition is closely related to the vector that maximizes a computationally intractable criterion using a Sobol' index. A numerical example of Asian option pricing finds that this active subspace preintegration strategy is competitive with preintegrating the first principal component of the Brownian motion, which is known to be very effective. The new method outperforms others on some basket and rainbow options where there is no generally accepted counterpart to the principal components construction.

Key words. conditional Monte Carlo, option pricing, quasi–Monte Carlo, randomized quasi-Monte Carlo

MSC codes. 65C05, 65D30

DOI. 10.1137/22M1479129

1. Introduction.

Preintegration [15] is a strategy for high dimensional numerical integration in which one variable is integrated out in a closed form (or by a very accurate quadrature rule) while the others are handled by quasi–Monte Carlo (QMC) sampling. This strategy has been used since the 1950s in Monte Carlo (MC) sampling [16, 45], where it is known as conditional Monte Carlo. It can reduce variance but not increase it. In the Markov chain Monte Carlo (MCMC) literature, such conditioning is called Rao-Blackwellization, although it does not generally bring the optimal variance reduction that results from the Rao–Blackwell theorem. In the MCMC setting it is possible for conditioning to increase variance [10]. See [40] for a survey of conditioning in MCMC.

The advantage of preintegration in QMC goes beyond the variance reduction that arises in MC. After preintegration, a d-dimensional integration problem with a discontinuity or a kink (discontinuity in the gradient) can be converted into a much smoother $(d-1)$-dimensional problem [15]. QMC exploits regularity of the integrand and this additional smoothness brings a benefit on top of the variance reduction that comes from conditioning. It is known that the resulting smoothness depends critically on a monotonicity property of the integrand with respect to the variable being integrated out [11, 14, 15]. Hoyt and Owen [19] give conditions where preintegration

* Received by the editors February 24, 2022; accepted for publication (in revised form) October 25, 2022; published electronically March 10, 2023.
https://doi.org/10.1137/22M1479129
Funding: The work of the authors was supported by National Science Foundation grant IIS-1837931. The work of the first author was also partially supported by the Stanford Data Science Scholars program.
† Department of Statistics, Stanford University, Stanford, CA 94305 USA (sfliu@stanford.edu, owen@stanford.edu).

reduces the mean dimension (that we will define later) of the integrand. It can reduce the mean dimension from proportional to $\sqrt{d}$ to $O(1)$ as $d \to \infty$ in a sequence of ridge functions with a discontinuity that the preintegration smooths out. He [17] studied the error rate of preintegration for scrambled nets applied to functions of the form $f(\boldsymbol{x}) = h(\boldsymbol{x})\mathbf{1}\{\phi(\boldsymbol{x}) \geqslant 0\}$ for a Gaussian variable $\boldsymbol{x}$. That work assumes that h and ϕ are smooth functions and ϕ is monotone in the variable to be preintegrated. Then preintegration has a smoothing effect that when combined with some boundary growth conditions yields a root mean squared error rate of $O(n^{-1+\varepsilon})$, where n is the number of samples. In some very early uses of preintegration in QMC, L'Ecuyer and Lemieux [23] integrate some variables out of a stochastic activity network problem to improve smoothness for applying a lattice rule.

There is presently very little guidance in the literature about which variable to preintegrate over, beyond the monotonicity condition recently studied in [11] and a remark for ridge functions in [19]. Many of the use cases for preintegration involve integration with respect to the multivariate Gaussian distribution, especially for problems arising in finance. In the Gaussian context, we have more choices for the variable to preintegrate over. In addition to preintegration over any one of the d coordinates, preintegration over any linear combination of the variables remains integration with respect to a univariate Gaussian variable. Our proposal is to preintegrate over a linear combination of variables chosen to optimize a measure of variable importance derived from active subspaces [5].

When sampling from a multivariate Gaussian distribution by QMC, even without preintegration, one must choose a square root of the covariance matrix by which to multiply some sampled scalar Gaussian variables. While the choice of square root does not affect MC error, it does affect the QMC error. There are numerous choices for that square root. One can sample via the principal component matrix decomposition as [2] and many others do. For integrands defined with respect to Brownian motions, one can use the Brownian bridge construction studied by [26]. These choices have some potential disadvantages. It is always possible that the integrand is little affected by the first principal component. In a pessimistic scenario, the integrand could depend only on a principal component that is orthogonal to the first one. This is a well-known pitfall in principal components regression [22]. In a related phenomenon, Papageorgiou [36] exhibits an integrand where QMC via the standard construction is more effective than via the Brownian bridge.

Not only might a principal component direction perform poorly, but the first principal component is not necessarily well defined. Although the problem may be initially defined in terms of a specific Gaussian distribution, by a change of variable we can rewrite our integral as an expectation with respect to another Gaussian distribution with a different covariance matrix that has a different first principal component. Or, if the problem is posed with a covariance equal to the d-dimensional identity matrix, then every unit vector (i.e., L^2-norm is 1) is a first principal component direction.

Some proposed methods take account of the specific integrand while formulating a sampling strategy. These include stratifying in a direction chosen from exponential tilting [13], exploiting a linearization of the integrand at $d+1$ special points starting with the center of the domain [20], and a gradient principal component analysis (GPCA) algorithm [47] that we describe in more detail later.

The problem we consider is to compute an approximation to $\mu = \mathbb{E}(f(\boldsymbol{x}))$, where $\boldsymbol{x} \in \mathbb{R}^d$ has the spherical Gaussian distribution denoted by $\mathcal{N}(0, I)$ and f is a continuous function with a gradient almost everywhere that is square integrable. Let

$C = \mathbb{E}(\nabla f(\boldsymbol{x})\nabla f(\boldsymbol{x})^\mathsf{T}) \in \mathbb{R}^{d\times d}$. The r-dimensional active subspace [5] is the space spanned by the r leading (i.e., top) eigenvectors of C. We always rank the eigenvalues and corresponding eigenvectors in descending order. Active subspace methods have been generalized to vector-valued functions [49] and nonlinear dimension reduction [3]. For other uses of the matrix C in numerical computation, see the references in [6]. For $r = 1$, let θ be the leading eigenvector of C normalized to be a unit vector in L^2-norm. We propose to preintegrate f over $\theta^\mathsf{T}\boldsymbol{x} \sim \mathcal{N}(0,1)$. Preintegration can smooth out both kinks and jumps. Because the active subspace direction is derived for continuous integrands it provides a principled preintegration choice for integrands with kinks, but its motivation does not extend to integrands with jumps.

The eigendecomposition of C is an uncentered principal components decomposition of the gradients, also known as the GPCA. The GPCA method [47] also uses the eigendecomposition of C to define a matrix square root for a QMC sampling strategy to reduce effective dimension, but it involves no preintegration. The algorithm in [46] preintegrates the first variable x_1 out of f. Then it applies GPCA to the remaining $d-1$ variables in order to find a suitable $(d-1)\times(d-1)$ matrix square root for the remaining Gaussian variables. They preintegrate over a coordinate variable, while we always preintegrate over the leading eigenvector which is not generally one of the d coordinates. All of the algorithms that involve C typically take a sample to estimate it.

This paper is organized as follows. Section 2 provides some background on randomized QMC (RQMC) and preintegration. Section 3 shows that preintegration never increases the variance of scrambled net integration, regardless of the smoothness of the integrand. Thus this well-known property of conditional MC, which does not always extend to conditional MCMC, does extend to conditional RQMC. In section 4, we describe using the unit vector θ which maximizes the Sobol' index for $\theta^\mathsf{T}\boldsymbol{x}$. This strategy is to find a matrix square root for which the first column maximizes the criterion from section 3. We show that this Sobol' index $\bar{\tau}^2_\theta$ is well defined in that it does not depend on how we parameterize the space orthogonal to θ. We apply active subspace preintegration to option pricing examples in section 5. These include an Asian call option, a basket option, and call on max and call on min options. Section 6 has our conclusions.

2. Background. In this section, we introduce some background of RQMC, preintegration, and active subspaces. First we introduce some notation.

For a positive integer d, we let $1{:}d = \{1, 2, \ldots, d\}$. For a subset $u \subseteq 1{:}d$, we let $-u = 1{:}d\backslash u$. For an integer $j \in 1{:}d$, we use j to represent $\{j\}$ when the context is clear and $-j = 1{:}d\backslash\{j\}$. Let $|u|$ denote the cardinality of u. For $\boldsymbol{x} \in \mathbb{R}^d$, we let $\boldsymbol{x}_u$ be the $|u|$-dimensional vector containing only the x_j with $j \in u$. For $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^d$ and $u \subseteq 1{:}d$, we let $\boldsymbol{x}_u{:}\boldsymbol{z}_{-u}$ be the d-dimensional vector, whose jth entry is x_j if $j \in u$ and z_j if $j \notin u$. We use $\mathbb{N} = \{1, 2, \ldots\}$ for the natural numbers and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. We denote the density and the cumulative distribution function (CDF) of the standard Gaussian distribution $\mathcal{N}(0,1)$ as φ and Φ, respectively. We let Φ^{-1} denote the inverse CDF of $\mathcal{N}(0,1)$. We also use φ to denote the density of the d-dimensional standard Gaussian distribution $\mathcal{N}(0, I_d)$, $\varphi(\boldsymbol{x}) = (2\pi)^{-d/2}\exp(-\|\boldsymbol{x}\|^2/2)$, where $\|\boldsymbol{x}\|$ always denotes the L^2-norm unless otherwise stated. We use $\mathcal{N}(0, I)$ when the dimension of the random variable is clear from context. For a matrix $\Theta \in \mathbb{R}^{d\times d}$, we let $\Theta[u]$ denote the columns of Θ whose indices are in u. For an event A, we define the indicator function $\mathbf{1}_A$ to be 1 if A happens and 0 otherwise. When A contains many levels of subscripts and/or superscripts we use $\mathbf{1}\{A\}$ instead.

2.1. QMC and RQMC. QMC provides a way to estimate $\mu = \int_{[0,1]^d} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ with typically greater accuracy than can be done by MC, while sharing with MC the ability to handle larger dimensions d than can be handled well by classical quadrature methods such as those in [7]. The QMC estimate, like the MC one, takes the form $\hat{\mu}_n = (1/n)\sum_{i=0}^{n-1} f(\boldsymbol{x}_i)$, except that instead of $\boldsymbol{x}_i \overset{\text{IID}}{\sim} \mathbb{U}[0,1]^d$ the sample points are chosen strategically and deterministically to cover the unit cube more evenly as quantified by a discrepancy measure (see [4]) such as the star discrepancy D_n^* (defined in the supplementary material (supplement.pdf [local/web 411KB])).

The QMC methods we study are digital nets and sequences. To define them, for an integer base $b \geqslant 2$ let $E(\boldsymbol{k},\boldsymbol{c}) = \prod_{j=1}^d [\frac{c_j}{b^{k_j}}, \frac{c_j+1}{b^{k_j}})$ for $\boldsymbol{k} = (k_1,\dots,k_d)$ and $\boldsymbol{c} = (c_1,\dots,c_d)$, where $k_j \in \mathbb{N}_0$ and $0 \leqslant c_j < b^{k_j}$. The sets $E(\boldsymbol{k},\boldsymbol{c})$ are called elementary intervals in base b. They have volume $b^{-|\boldsymbol{k}|}$ where $|\boldsymbol{k}| = \sum_{j=1}^d k_j$. For integers $m \geqslant t \geqslant 0$, the points $\boldsymbol{x}_0,\dots,\boldsymbol{x}_{n-1}$ with $n = b^m$ are a (t,m,d)-net in base b if every elementary interval $E(\boldsymbol{k},\boldsymbol{c})$ with $|\boldsymbol{k}| \leqslant m-t$ contains $b^{m-|\boldsymbol{k}|}$ of those points, which is exactly n times the volume of $E(\boldsymbol{k},\boldsymbol{c})$. Here, t is known as the quality parameter and smaller values are better. It is not always possible to get $t = 0$. The power of digital nets is that the points $\boldsymbol{x}_i$ satisfy $\binom{m-t+d-1}{d-1}$ such stratifications simultaneously. They attain a star discrepancy [27] of $D_n^* = O((\log n)^{d-1}/n)$ after approximating each $[\boldsymbol{0},\boldsymbol{a})$ by sets $E(\boldsymbol{k},\boldsymbol{c})$ [27, Theorem 4.10]. An infinite sequence $\{\boldsymbol{x}_i\}_{i\geqslant 0}$ is called a (t,d)-sequence if for all $k \geqslant 0$ and $m \geqslant t$ the points $\boldsymbol{x}_{kb^m},\dots,\boldsymbol{x}_{(k+1)b^m-1}$ are a (t,m,d)-net in base b.

Randomization techniques can be applied so that one can estimate the error by multiple independent replicates. In RQMC, the points $\boldsymbol{x}_i \sim \mathbb{U}([0,1]^d)$ individually while collectively they form a (t,m,d)-net or sequence with probability 1. One randomization technique that has this property is nested uniform scrambling introduced by [29]. The estimate $\hat{\mu}_n$ taken over a scrambled (t,d)-sequence satisfies a strong law of large numbers if $f \in L^2[0,1]$ [35]. If $f \in L^2[0,1]^d$, then $\mathrm{Var}(\hat{\mu}_n) = o(1/n)$, giving the method asymptotically unbounded efficiency versus MC which has variance σ^2/n for $\sigma^2 = \mathrm{Var}(f(\boldsymbol{x}))$. For smooth enough f, $\mathrm{Var}(\hat{\mu}_n) = O(n^{-3}(\log n)^{d-1})$ [31, 33] with the sharpest sufficient condition in [48]. Also there exists $\Gamma < \infty$ with $\mathrm{Var}(\hat{\mu}_n) \leqslant \Gamma\sigma^2/n$ for all $f \in L^2[0,1]^d$ [32]. This bound involves no powers of $\log(n)$.

For more background about QMC and RQMC, see the supplement (SM1).

2.2. Preintegration. For $j \in 1{:}d$, $\int_{[0,1]^d} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_{[0,1]^{d-1}} \int_0^1 f(\boldsymbol{x})\,\mathrm{d}x_j\,\mathrm{d}\boldsymbol{x}_{-j}$, which we can also write as $\mathbb{E}(f(\boldsymbol{x})) = \mathbb{E}(\mathbb{E}(f(\boldsymbol{x})\,|\,\boldsymbol{x}_{-j}))$ for $\boldsymbol{x} \sim \mathbb{U}[0,1]^d$. For $\boldsymbol{x} \in [0,1]^d$, define $g(\boldsymbol{x}) = g_j(\boldsymbol{x}) = \int_0^1 f(\boldsymbol{x})\,\mathrm{d}x_j$. It simplifies the presentation of some of our results, especially Theorem 3.2 on variance reduction, to keep g defined as above on $[0,1]^d$ even though it only depends on $\boldsymbol{x}_{-j}$. In preintegration,

$$\hat{\mu}_n = \hat{\mu}_{n,j} = \frac{1}{n}\sum_{i=0}^{n-1} g_j(\boldsymbol{x}_i)$$

which, as we noted in the introduction, is conditional MC except that we now use RQMC inputs.

Preintegration can bring some advantages for RQMC. The plain MC variance of $g(\boldsymbol{x})$ is no larger than that of $f(\boldsymbol{x})$ and is generally smaller unless f does not depend at all on x_j. Thus the bound $\Gamma\sigma^2/n$ is reduced. Moreover, the preintegrated integrand g can be much smoother than f and (R)QMC improves on MC by exploiting this additional smoothness. For example, it has been observed that for some option pricing integrands, preintegrating certain variables can remove the discontinuities in the integrand or its gradient [15, 46].

The integrands we consider here are defined with respect to a Gaussian random variable. We are interested in $\mu = \mathbb{E}(f(\boldsymbol{y}))$ for $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{0}, \Sigma)$ with a positive definite covariance $\Sigma \in \mathbb{R}^{d\times d}$. Letting $R_0 \in \mathbb{R}^{d\times d}$ with $R_0 R_0^{\mathsf{T}} = \Sigma$ we can write $\mu = \mathbb{E}(f(R_0\boldsymbol{z}))$ for $\boldsymbol{z} \sim \mathcal{N}(0, I)$. For an orthogonal matrix $Q \in \mathbb{R}^{d\times d}$ we also have $Q\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, I)$. Then taking $\boldsymbol{z} = \Phi^{-1}(\boldsymbol{x})$ componentwise leads us to the estimate $\hat{\mu} = \frac{1}{n}\sum_{i=0}^{n-1} f(R\Phi^{-1}(\boldsymbol{x}_i))$ with $R = R_0 Q$ for RQMC points $\boldsymbol{x}_i$. The choice of Q or equivalently R does not affect the MC variance of $\hat{\mu}$ but it can change the RQMC variance. We will consider some examples later. The mapping Φ^{-1} from $\mathbb{U}[0,1]^d$ to $\mathcal{N}(\boldsymbol{0}, I)$ can be replaced by another one such as the Box–Muller transformation. The choice of transformation does not affect the MC variance but does affect the RQMC variance. Most researchers use Φ^{-1} but [28] advocates for Box–Muller.

When we are using preintegration for a problem defined with respect to an $\mathcal{N}(0, \Sigma)$ random variable we must choose R and then the coordinate j over which to preintegrate. Our approach is to fix $j = 1$ while choosing R so that x_1 is the most important linear combination of $\boldsymbol{x}$ in an active subspace approximation as described in section 4.

2.3. The ANOVA decomposition. For $f \in L^2[0,1]^d$ we can define an analysis of variance (ANOVA) decomposition from [9, 18, 42]. For details see [34, Appendix A.6]. This decomposition reads

$$f(\boldsymbol{x}) = \sum_{u\subseteq 1:d} f_u(\boldsymbol{x})$$

where f_u depends on $\boldsymbol{x}$ only through x_j with $j \in u$ and also $\int_0^1 f_u(\boldsymbol{x})\,\mathrm{d}x_j = 0$ whenever $j \in u$. The term $f_\varnothing$ is the constant function everywhere equal to $\mu = \int_{[0,1]^d} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$. The decomposition is orthogonal in that $\int_{[0,1]^d} f_u(\boldsymbol{x}) f_v(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 0$ if $u \neq v$. To each effect f_u there corresponds a variance component

$$\sigma_u^2 = \mathrm{Var}(f_u(\boldsymbol{x})) = \begin{cases} \int_{[0,1]^d} f_u(\boldsymbol{x})^2\,\mathrm{d}\boldsymbol{x}, & |u| > 0 \\ 0 & \text{else.} \end{cases}$$

The variance components sum to $\sigma^2 = \mathrm{Var}(f(\boldsymbol{x}))$.

The formula for f_u can be written recursively as $f_\varnothing = \mu$ and then for $|u| > 0$,

$$f_u(\boldsymbol{x}) = \int_{[0,1]^{d-|u|}} \biggl(f(\boldsymbol{x}) - \sum_{v\subsetneq u} f_v(\boldsymbol{x}) \biggr)\,\mathrm{d}\boldsymbol{x}_{-u}.$$

The integral above is an expectation with respect to independent random x_j for $j \notin u$. For Gaussian or indeed any other distribution on components x_j one replaces the integral by the analogous expectation. We will use the ANOVA decomposition below when describing how to choose a preintegration variable.

Sobol' indices [43] are derived from the ANOVA decomposition. For $u \subseteq 1{:}d$ these are

$$\underline{\tau}_u^2 = \sum_{v\subseteq u} \sigma_v^2 \quad \text{and} \quad \overline{\tau}_u^2 = \sum_{v\subseteq 1:d} \mathbf{1}_{\{u\cap v\neq\varnothing\}} \sigma_v^2.$$

They provide two ways to judge the importance of the set of variables x_j for $j \in u$. They are usually normalized by σ^2 to get an interpretation as a proportion of variance explained.

The mean dimension of f is $\nu(f) = \sum_{u \subseteq 1:d} |u| \sigma_u^2/\sigma^2$. It satisfies $\nu(f) = \sum_{j=1}^d \overline{\tau}_j^2/\sigma^2$. A ridge function takes the form $f(\boldsymbol{x}) = h(\Theta^{\mathsf{T}}\boldsymbol{x})$ for $\Theta \in \mathbb{R}^{d\times r}$ with $\Theta^{\mathsf{T}}\Theta = I_r$ and a function $h : \mathbb{R}^r \to \mathbb{R}$. For $\boldsymbol{x} \sim \mathcal{N}(0, I_d)$, the distribution of $f(\boldsymbol{x})$ does not depend on d and the mean dimension is $O(1)$ as $d \to \infty$ if h is Lipschitz [19]. If h is an indicator function and $r = 1$, then it is possible to have $\nu(f) = \Omega(\sqrt{d})$ reduced to $O(1)$ by preintegration over a component variable x_j with Θ_{j1} bounded away from zero as $d \to \infty$. See [19] for an example.

3. Preintegration and scrambled net variance. Conditional MC can reduce but not increase the variance of plain MC integration. Here we show that the same thing holds for scrambled nets using the nested uniform scrambling of [29]. The affine linear scrambling of [25] has the same variance and hence the same result. We assume that $f \in L^2[0,1)^d$. For any $f \in L^2[0,1]^d$ we could set $f(\boldsymbol{x}) = 0$ for any $\boldsymbol{x} \in D = [0,1]^d \backslash [0,1)^d$ and get an equivalent function with the same integral and, almost surely, the same RQMC estimate because all $\boldsymbol{x}_i \sim \mathbb{U}[0,1]^d$ avoid D with probability one.

We will preintegrate over one of the d components of $\boldsymbol{x} \in [0,1)^d$. It is also possible to preintegrate over multiple components and reduce the RQMC variance each time, though the utility of that strategy is limited by the availability of suitable closed forms or effective quadratures.

3.1. Walsh function expansions. To get variance formulas for scrambled nets we follow Dick and Pillichshammer [8], who work with Walsh function expansion in $L^2[0,1)^d$, for which they credit Pirsic [38]. Let $\omega_b = e^{2\pi i/b}$ with i being the imaginary unit. For $k \in \mathbb{N}_0$ write $k = \sum_{j\geqslant 0} \kappa_j b^j$ with base b digits $\kappa_j \in \{0, 1, \ldots, b-1\}$. For $x \in [0,1)$ write $x = \sum_{k\geqslant 1} \xi_j b^{-j}$ with base b digits $\xi_j \in \{0, 1, \ldots, b-1\}$. This is unique in the sense that infinitely many ξ_j's must be different from $b-1$.

Using the above notation we can define the kth b-adic Walsh function ${}_b\mathrm{wal}_k{:}[0,1) \to \mathbb{C}$ as ${}_b\mathrm{wal}_k(x) = \omega_b^{\sum_{j\geqslant 1} \xi_j \kappa_{j-1}}$. The summation in the exponent is finite because $k < \infty$. Note that ${}_b\mathrm{wal}_0(x) = 1$ for all $x \in [0,1)$. For $\boldsymbol{x} = (x_1, \ldots, x_d) \in [0,1)^d$ and $\boldsymbol{k} = (k_1, \ldots, k_d) \in \mathbb{N}_0^d$, the d-dimensional Walsh functions are defined as ${}_b\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x}) = \prod_{j=1}^d {}_b\mathrm{wal}_{k_j}(x_j)$. The Walsh series expansion of $f(\boldsymbol{x})$ is

$$f(\boldsymbol{x}) \sim \sum_{\boldsymbol{k}\in\mathbb{N}_0^d} \hat{f}(\boldsymbol{k}) {}_b\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x}), \text{ where } \hat{f}(\boldsymbol{k}) = \int_{[0,1)^d} f(\boldsymbol{x}) \overline{{}_b\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x})}\, \mathrm{d}\boldsymbol{x}.$$

The d-dimensional b-adic Walsh function system is a complete orthonormal basis in $L_2([0,1)^d)$ [8, Theorem A.11] and the series expansion converges to f in L^2.

For $\boldsymbol{x} \sim \mathbb{U}([0,1]^d)$, the variance of $f(\boldsymbol{x})$ has the decomposition

$$\mathrm{Var}(f(\boldsymbol{x})) = \sum_{\boldsymbol{k}\in\mathbb{N}_0^d\backslash\{\boldsymbol{0}\}} |\hat{f}(\boldsymbol{k})|^2.$$

The variance under scrambled nets is different. To study it we group the Walsh coefficients. For $\boldsymbol{\ell} \in \mathbb{N}_0^d$ let $C_{\boldsymbol{\ell}} = \{\boldsymbol{k} \in \mathbb{N}_0^d \mid \lfloor b^{\ell_j - 1} \rfloor \leqslant k_j < b^{\ell_j}, 1 \leqslant j \leqslant d\}$. Then define $\beta_{\boldsymbol{\ell}}(\boldsymbol{x}) = \sum_{\boldsymbol{k}\in C_{\boldsymbol{\ell}}} \hat{f}(\boldsymbol{k}) {}_b\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x})$. The functions $\beta_{\boldsymbol{\ell}}$ are orthogonal in that $\int_{[0,1)^d} \beta_{\boldsymbol{\ell}}(\boldsymbol{x}) \overline{\beta_{\boldsymbol{\ell}'}(\boldsymbol{x})}\, \mathrm{d}\boldsymbol{x} = 0$ when $\boldsymbol{\ell}' \neq \boldsymbol{\ell}$. For $\boldsymbol{\ell} \neq \boldsymbol{0}$, $\beta_{\boldsymbol{\ell}}(\boldsymbol{x})$ has variance

$$\sigma_{\boldsymbol{\ell}}^2(f) = \int_{[0,1)^d} |\beta_{\boldsymbol{\ell}}(\boldsymbol{x})|^2\, \mathrm{d}\boldsymbol{x} = \sum_{\boldsymbol{k}\in C_{\boldsymbol{\ell}}} |\hat{f}(\boldsymbol{k})|^2.$$

Let $\mathbf{a}_0, \dots, \mathbf{a}_{n-1}$ be a point set in $[0,1]^d$ and let $\boldsymbol{x}_0, \dots, \boldsymbol{x}_{n-1}$ be the scrambled version of $\mathbf{a}_i$'s. Then for the estimator $\hat{\mu} = \hat{\mu}_n = \frac{1}{n}\sum_{i=0}^{n-1} f(\boldsymbol{x}_i)$ we have

$$\mathrm{Var}(\hat{\mu}_n) = \frac{1}{n} \sum_{\boldsymbol{\ell} \in \mathbb{N}_0^d \setminus \{\mathbf{0}\}} \Gamma_{\boldsymbol{\ell}} \sigma_{\boldsymbol{\ell}}^2(f) \tag{3.1}$$

for a collection of gain coefficients $\Gamma_{\boldsymbol{\ell}} \geqslant 0$ that depend on the $\boldsymbol{a}_i$. This expression can also be obtained through a base b Haar wavelet decomposition [30]. Our $\Gamma_{\boldsymbol{\ell}}$ equals $nG_{\boldsymbol{\ell}}$ from [8]. The variance of $\hat{\mu}_n$ under independent and identically distributed MC sampling is $(1/n)\sum_{\boldsymbol{\ell} \in \mathbb{N}_0^d \setminus \{\mathbf{0}\}} \sigma_{\boldsymbol{\ell}}^2(f)$, so $\Gamma_{\boldsymbol{\ell}} < 1$ corresponds to integrating the term $\beta_{\boldsymbol{\ell}}(\boldsymbol{x})$ with less variance than MC does.

If scrambling of [29] or [25] is applied to $\boldsymbol{a}_i$, then

$$\Gamma_{\boldsymbol{\ell}} = \frac{1}{n} \sum_{i,i'=0}^{n-1} \prod_{j=1}^{d} \frac{b\mathbf{1}\left\{\lfloor b^{\ell_j} a_{i,j} \rfloor = \lfloor b^{\ell_j} a_{i',j} \rfloor\right\} - \mathbf{1}\left\{\lfloor b^{\ell_j - 1} a_{i,j} \rfloor = \lfloor b^{\ell_j - 1} a_{i',j} \rfloor\right\}}{b-1}. \tag{3.2}$$

This holds for any $\boldsymbol{a}_i$, not just digital nets. When $\boldsymbol{a}_i$ are the first b^m points of a (t,d)-sequence in base b, then $\Gamma = \sup_{\boldsymbol{\ell}} \Gamma_{\boldsymbol{\ell}} < \infty$ (uniformly in m) so that $\mathrm{Var}(\hat{\mu}) \leqslant \Gamma\sigma^2/n$. Similarly, for any $\boldsymbol{\ell} \in \mathbb{N}_0^d$ we have $\Gamma_{\boldsymbol{\ell}} \to 0$ as $n = b^m \to \infty$ in a (t,d)-sequence in base b from which we get $\mathrm{Var}(\hat{\mu}_n) = o(1/n)$. For a (t,m,d)-net in base b, one can show that the gain coefficients $\Gamma_{\boldsymbol{\ell}} = 0$ for all $\boldsymbol{\ell}$ with $|\boldsymbol{\ell}| \leqslant m - t$.

3.2. Walsh decomposition after preintegration.

PROPOSITION 3.1. *For $f \in L^2[0,1)^d$ and $j \in 1{:}d$, let g be f preintegrated over x_j. Then for $\boldsymbol{k} \in \mathbb{N}_0^d$,*

$$\hat{g}(\boldsymbol{k}) = \begin{cases} \hat{f}(\boldsymbol{k}), & k_j = 0 \\ 0, & k_j > 0. \end{cases} \tag{3.3}$$

Proof. We write

$$\hat{g}(\boldsymbol{k}) = \int_{[0,1)^{d-1}} \int_0^1 g(\boldsymbol{x}) \overline{{}_b\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x})}\, \mathrm{d}x_j\, \mathrm{d}\boldsymbol{x}_{-j} = \int_{[0,1)^{d-1}} g(\boldsymbol{x}) \int_0^1 \overline{{}_b\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x})}\, \mathrm{d}x_j\, \mathrm{d}\boldsymbol{x}_{-j}$$

because $g(\boldsymbol{x})$ does not depend on x_j. If $k_j > 0$, then the inner integral vanishes, establishing the second clause in (3.3). If $k_j = 0$, then ${}_b\mathrm{wal}_{k_j}(x_j) = 1$ for all x_j and the inner integral equals $\prod_{\ell \neq j} {}_b\mathrm{wal}_{k_\ell}(x_j) = {}_b\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x})$, establishing the first clause. ∎

THEOREM 3.2. *For $\boldsymbol{a}_0, \dots, \boldsymbol{a}_{n-1} \in [0,1)^d$ let $\boldsymbol{x}_0, \dots, \boldsymbol{x}_{n-1}$ be a scrambled version of them using the algorithm from* [29] *or* [25]. *Let $f \in L^2[0,1)^d$, and for $j \in 1{:}d$, let g be f preintegrated over x_j. Then*

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=0}^{n-1} g(\boldsymbol{x}_i)\right) \leqslant \mathrm{Var}\left(\frac{1}{n}\sum_{i=0}^{n-1} f(\boldsymbol{x}_i)\right).$$

Proof. With either f or g we have the same gain coefficients $\Gamma_{\boldsymbol{\ell}}$ for $\boldsymbol{\ell} \in \mathbb{N}_0^d$. However

$$\sigma_{\boldsymbol{\ell}}^2(g) = \sum_{\boldsymbol{k} \in C_{\boldsymbol{\ell}}} |\hat{g}(\boldsymbol{k})|^2 = \sum_{\boldsymbol{k} \in C_{\boldsymbol{\ell}}, k_j = 0} |\hat{f}(\boldsymbol{k})|^2 \leqslant \sum_{\boldsymbol{k} \in C_{\boldsymbol{\ell}}} |\hat{f}(\boldsymbol{k})|^2 = \sigma_{\boldsymbol{\ell}}^2(f).$$

The result now follows from (3.1). ∎

Theorem 3.2 shows that preintegration does not increase the variance under scrambling. This holds whether or not the underlying points are a digital net, though of course the main case of interest is for scrambling of digital nets and sequences. It holds because the RQMC error is a sum of uncorrelated errors (one per Walsh coefficient) and preintegration can decrease but not increase their variances. A similar phenomenon happens for RQMC by random shifting of lattice rules or by digital shifting of digital nets. For those RQMC methods also, preintegration cannot increase the variance. The proofs for those cases are in supplementary material, section SM3.

A variance reduction for conditional RQMC was previously found by [1]. Their Theorem 4.3 shows that after conditioning on a certain linear combination of variables, the variance of a randomly shifted lattice rule approach for some barrier options cannot be greater than the unconditional sampling variance. Their proof is specific to barrier options. They note that their argument would also apply to digital shifts.

Preintegration has another benefit that is not captured by Theorem 3.2. By reducing the input dimension from d to $d-1$ we will be using a $(t', m, d-1)$-net in base b with $t' \leqslant t$ and possibly $t' < t$, since the lowest possible quality parameter commonly increases with the dimension. For scrambled net sampling, reducing the dimension reduces an upper bound on the variance. For any function $f \in L^2[0,1)^d$, the variance using a scrambled (t,m,d)-net is at most b^{t+d} times the MC variance. Reducing the dimension reduces the bound to $b^{t'+d-1}$ times the MC variance.

As remarked above, preintegration of f over a variable that f uses will reduce the variance under scrambled net sampling. This reduction does not require f to be monotone in x_j, though such cases have the potential to bring a greater improvement [14, 15].

3.3. Choice of preintegration variable. In order to choose x_j to preintegrate over, we can look at the variance reduction we get. Preintegrating over x_j reduces the scrambled net variance by

$$\frac{1}{n}\sum_{\boldsymbol{\ell}\in\mathbb{N}_0^d\setminus\{\mathbf{0}\}}\Gamma_{\boldsymbol{\ell}}\sigma^2_{\boldsymbol{\ell}}(f)\mathbf{1}_{\{\ell_j>0\}} = \frac{1}{n}\sum_{\boldsymbol{\ell}\in\mathbb{N}_0^d\setminus\{\mathbf{0}\}}\Gamma_{\boldsymbol{\ell}}\sum_{\boldsymbol{k}\in C_{\boldsymbol{\ell}}}|\hat{f}(\boldsymbol{k})|^2\mathbf{1}_{\{k_j>0\}}. \tag{3.4}$$

Evaluating this quantity for each $j \in 1{:}d$ might be more expensive than getting a good estimate of μ. However, we don't need to find the best j. Any j where f depends on x_j will bring some improvement. Below we develop a principled and computationally convenient strategy of choosing the j which is most important as measured by a Sobol' index [43] from global sensitivity analysis [39].

A convenient proxy replacement for (3.4) is

$$\frac{1}{n}\sum_{\boldsymbol{k}\in\mathbb{N}_0^d\setminus\{\mathbf{0}\}}|\hat{f}(\boldsymbol{k})|^2\mathbf{1}_{\{k_j>0\}} = \frac{1}{n}\sum_{u\subseteq 1:d}\sigma_u^2\mathbf{1}_{\{j\in u\}}, \tag{3.5}$$

where σ_u^2 is the ANOVA variance component of f for the set u. The equality above follows because the ANOVA can be defined by collecting up the terms involving x_j for $j \in u$ from the orthogonal decomposition as Sobol' did similarly in [42] using Haar functions. The right-hand side of (3.5) equals $\overline{\tau}_j^2/n$. It counts all the variance components in which variable j participates. Conditioning on x_j removes $\overline{\tau}_j^2/n$ from the plain MC variance [46]. From the orthogonality properties of ANOVA effects it follows that

$$\overline{\tau}_j^2 = \frac{1}{2}\int_{[0,1]^{d+1}}\bigl(f(\boldsymbol{z}_j{:}\boldsymbol{x}_{-j}) - f(\boldsymbol{x})\bigr)^2\,\mathrm{d}\boldsymbol{z}_j\,\mathrm{d}\boldsymbol{x}.$$

The Jansen estimator [21] is an estimate of the above integral that can be done by a $d+1$ dimensional MC or QMC or RQMC sampling algorithm. Our main interest in this Sobol' index estimator is that we use it as a point of comparison to the use of active subspaces in choosing a projection of a Gaussian vector along which to preintegrate.

4. Active subspace method. An expectation defined with respect to $\boldsymbol{x} \sim \mathcal{N}(0,\Sigma)$ for positive definite $\Sigma \in \mathbb{R}^{d\times d}$ can always be written as an expectation with respect to $\boldsymbol{x} \sim \mathcal{N}(0,I)$. For a unit vector $\theta \in \mathbb{R}^d$, we will preintegrate over $\boldsymbol{x}^\mathsf{T}\theta \sim \mathcal{N}(0,1)$ and then the problem is to make a principled choice of θ. It would not be practical to seek an optimal choice.

Our proposal is to use active subspaces [5]. As mentioned in the introduction we let

$$C = \mathbb{E}(\nabla f(\boldsymbol{x})\nabla f(\boldsymbol{x})^\mathsf{T})$$

and then let $\Theta[1{:}r]$ comprise the r leading eigenvectors of C. The original purpose for using active subspaces is to approximate $f(\boldsymbol{x}) \approx \tilde{f}(\Theta[1{:}r]^\mathsf{T}\boldsymbol{x})$ for some function $\tilde{f}$ on $\mathbb{R}^r$. It is well known that one can construct functions where the active subspace will be a bad choice over which to approximate. For instance, with $f(\boldsymbol{x}) = \sin(10^6 x_1) + 100x_2$ the $r=1$ active subspace provides a function of x_1 alone, while a function of x_2 alone can provide a better approximation. Active subspaces remain useful for approximation because the motivating problems are usually not so pathological and there is a human in the loop to catch such things. They also have an enormous practical advantage in that one set of evaluations of ∇f can be used in the search for Θ instead of having every candidate Θ require its own evaluations of ∇f. Using active subspaces for integration retains that advantage.

In our setting, we take $r=1$ and preintegrate over $\theta^\mathsf{T}\boldsymbol{x}$, where θ is the leading eigenvector of C. That is, θ maximizes $\theta^\mathsf{T}\mathbb{E}(\nabla f(\boldsymbol{x})\nabla f(\boldsymbol{x})^\mathsf{T})\theta$ over all d-dimensional unit vectors. Now suppose that instead of using $f(\boldsymbol{x})$ we use $f_Q(\boldsymbol{x}) = f(Q\boldsymbol{x})$ for an orthogonal matrix $Q \in \mathbb{R}^{d\times d}$. Then $\mathbb{E}(\nabla f_Q(\boldsymbol{x})\nabla f_Q(\boldsymbol{x})^\mathsf{T}) = Q^\mathsf{T}CQ$, which is similar to C. It has the same eigenvalues and the leading eigenvector is $\tilde{\theta} = Q^\mathsf{T}\theta$.

4.1. Connection to a Sobol' index. The discussion in section 3 motivates preintegration of $f(\boldsymbol{x})$ for $\boldsymbol{x} \sim \mathcal{N}(0,I)$ over a linear combination $\theta^\mathsf{T}\boldsymbol{x}$ having the largest Sobol' index over unit vectors θ. For $\theta_1 = \theta$, let $\Theta = (\theta_1, \theta_2, \ldots, \theta_d) \in \mathbb{R}^{d\times d}$ be an orthogonal matrix and write

$$f_\Theta(\boldsymbol{x}) = f(\Theta\boldsymbol{x}) = f(x_1\theta_1 + x_2\theta_2 + \cdots + x_d\theta_d).$$

Then we define $\overline{\tau}^2_\theta(f)$ to be $\overline{\tau}^2_1$ in the ANOVA decomposition of $f_\Theta(\boldsymbol{x})$. First we show that $\overline{\tau}^2_\theta$ does not depend on $\Theta_{-1} \equiv \Theta[2{:}d]$, the last $d-1$ columns of Θ.

Let z, $\tilde{z}$, and $\boldsymbol{y}$ be independent with distributions $\mathcal{N}(0,1)$, $\mathcal{N}(0,1)$, and $\mathcal{N}(0,I_{d-1})$, respectively. Let $\boldsymbol{x} = z\theta_1 + \Theta_{-1}\boldsymbol{y}$ and $\tilde{\boldsymbol{x}} = \tilde{z}\theta_1 + \Theta_{-1}\boldsymbol{y}$. Using the Jansen formula for $\theta = \theta_1$,

$$\overline{\tau}^2_\theta = \frac{1}{2}\int_{\mathbb{R}}\int_{\mathbb{R}}\int_{\mathbb{R}^{d-1}} \bigl(f(z\theta_1 + \Theta_{-1}\boldsymbol{y}) - f(\tilde{z}\theta_1 + \Theta_{-1}\boldsymbol{y})\bigr)^2 \varphi(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}\varphi(\tilde{z})\,\mathrm{d}\tilde{z}\varphi(z)\,\mathrm{d}z. \tag{4.1}$$

Now for an orthogonal matrix $Q \in \mathbb{R}^{(d-1)\times(d-1)}$, let

$$\widetilde{\Theta} = \Theta\begin{pmatrix} 1 & \mathbf{0}_{d-1}^\mathsf{T} \\ \mathbf{0}_{d-1} & Q \end{pmatrix} = (\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_d)$$

where $\tilde{\theta}_1 = \theta_1$. In this parameterization we get

$$\overline{\tau}_\theta^2 = \frac{1}{2}\int_{\mathbb{R}}\int_{\mathbb{R}}\int_{\mathbb{R}^{d-1}} \big(f(z\theta_1 + \Theta_{-1}Q\boldsymbol{y}) - f(\tilde{z}\theta_1 + \Theta_{-1}Q\boldsymbol{y})\big)^2 \varphi(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}\varphi(\tilde{z})\,\mathrm{d}\tilde{z}\varphi(z)\,\mathrm{d}z$$

which matches (4.1) after a change of variable. There is an even stronger invariance property in this setup. The random variable $\mathbb{E}(f_\Theta(\boldsymbol{x})\,|\,\boldsymbol{x}_{-1})$ has a distribution that does not depend on $\theta_2, \ldots, \theta_d$.

THEOREM 4.1. *Let f satisfy $\mathbb{E}(f(\boldsymbol{x})^2) < \infty$ for $\boldsymbol{x} \sim \mathcal{N}(0, I_d)$ and let $\Theta \in \mathbb{R}^{d\times d}$ be an orthogonal matrix with columns θ_j for $j = 1, \ldots, d$. Then the distribution of $\mathbb{E}(f_\Theta(\boldsymbol{x})|\boldsymbol{x}_{-1})$ does not depend on the last $d-1$ columns of Θ.*

The proof is in the supplementary material (SM2). Another consequence of Theorem 4.1 is that $\underline{\tau}_\theta^2(f) = \underline{\tau}_1^2(f_\Theta)$ is unaffected by $\theta_2, \ldots, \theta_d$. Because the variance of f is unchanged by making an orthogonal matrix transformation of its inputs, the normalized Sobol' indices $\underline{\tau}_\theta^2/\sigma^2$ and $\overline{\tau}_\theta^2/\sigma^2$ are also invariant.

Finding the optimal θ would ordinarily require an expensive search because every estimate of $\overline{\tau}_\theta^2$ for a given θ would require its own collection of evaluations of f. Using a Poincaré inequality from [44] we can bound that Sobol' index by

$$\overline{\tau}_\theta^2(f) \leqslant \mathbb{E}((\theta^{\mathsf{T}}\nabla f(\boldsymbol{x}))^2) = \theta^{\mathsf{T}} C \theta.$$

The active subspace direction thus maximizes an upper bound on the Sobol' index for a projection. Next we develop a deeper correspondence between these two measures.

For a unit vector $\theta \in \mathbb{R}^d$, we can write $f(\boldsymbol{x}) = f(\theta\theta^{\mathsf{T}}\boldsymbol{x} + (I - \theta\theta^{\mathsf{T}})\boldsymbol{x})$. If $\boldsymbol{x}, \boldsymbol{z}$ are independent $\mathcal{N}(0, I)$ vectors, then we can change the component of $\boldsymbol{x}$ parallel to θ by changing the argument of f to be $\theta\theta^{\mathsf{T}}\boldsymbol{z} + (I - \theta\theta^{\mathsf{T}})\boldsymbol{x}$. This leaves the resulting point unchanged in the $d-1$ dimensional space orthogonal to θ. Let $\tilde{x} = \theta^{\mathsf{T}}\boldsymbol{x}$ and $\tilde{z} = \theta^{\mathsf{T}}\boldsymbol{z}$. Then $\tilde{x}, \tilde{z} \sim \mathcal{N}(0,1)$ and $(I - \theta\theta^{\mathsf{T}})\boldsymbol{x} \sim \mathcal{N}(0, I - \theta\theta^{\mathsf{T}})$ are all independent. If f is differentiable, then by the mean value theorem

$$f(\theta\theta^{\mathsf{T}}\boldsymbol{z} + (I - \theta\theta^{\mathsf{T}})\boldsymbol{x}) - f(\theta\theta^{\mathsf{T}}\boldsymbol{x} + (I - \theta\theta^{\mathsf{T}})\boldsymbol{x}) = \theta^{\mathsf{T}}\nabla f(\theta\tilde{y} + (I - \theta\theta^{\mathsf{T}})\boldsymbol{x})(\tilde{z} - \tilde{x})$$

for a real number $\tilde{y}$ between $\tilde{x}$ and $\tilde{z}$. Using the Jansen formula, the Sobol' index for this projection is

$$\frac{1}{2}\theta^{\mathsf{T}}\mathbb{E}\Big((\tilde{z} - \tilde{x})^2 \nabla f(\theta\tilde{y} + (I - \theta\theta^{\mathsf{T}})\boldsymbol{x})\nabla f(\theta\tilde{y} + (I - \theta\theta^{\mathsf{T}})\boldsymbol{x})^{\mathsf{T}}\Big)\theta \tag{4.2}$$

which matches $\theta^{\mathsf{T}}\mathbb{E}(\nabla f(\boldsymbol{x})\nabla f(\boldsymbol{x})^{\mathsf{T}})\theta$ over a $d-1$ dimensional subspace but differs from it as follows. First, it includes a weight factor $(\tilde{z} - \tilde{x})^2$ that puts more emphasis on pairs of inputs where $\theta^{\mathsf{T}}\boldsymbol{x}$ and $\theta^{\mathsf{T}}\boldsymbol{z}$ are far from each other. Second, the evaluation point projected onto θ equals $\tilde{y}$, which lies between two independent $\mathcal{N}(0,1)$ variables instead of having the $\mathcal{N}(0,1)$ distribution, and its exact location depends on details of f and there could be more than one such $\tilde{y}$ for some f. The formula simplifies in an illustrative way for quadratic functions f.

PROPOSITION 4.2. *If $f : \mathbb{R}^d \to \mathbb{R}$ is a quadratic function and $\theta \in \mathbb{R}^d$ is a unit vector, then the Sobol' index $\overline{\tau}_\theta^2$ is*

$$\theta^{\mathsf{T}}\mathbb{E}\left(\nabla f\left(\frac{\theta\theta^{\mathsf{T}}\boldsymbol{x}}{\sqrt{2}} + (I - \theta\theta^{\mathsf{T}})\boldsymbol{x}\right)\nabla f\left(\frac{\theta\theta^{\mathsf{T}}\boldsymbol{x}}{\sqrt{2}} + (I - \theta\theta^{\mathsf{T}})\boldsymbol{x}\right)^{\mathsf{T}}\right)\theta. \tag{4.3}$$

Proof. If f is quadratic, then $\tilde{y} = (\tilde{z} + \tilde{x})/2 \sim \mathcal{N}(0, 1/2)$ and $\tilde{z} - \tilde{x} \sim \mathcal{N}(0, 2)$ and $(I - \theta\theta^{\mathsf{T}})\boldsymbol{x}$ are all independent. Then $\mathbb{E}((\tilde{z} - \tilde{x})^2) = 2$ and $\theta\tilde{y}$ has the same distribution as $\theta\theta^{\mathsf{T}}\boldsymbol{x}/\sqrt{2}$ which is also independent of $(\tilde{z} - \tilde{x})$ and $(I - \theta\theta^{\mathsf{T}})\boldsymbol{x}$. Making those substitutions in (4.2) yields (4.3). ∎

The Sobol' index in (4.3) matches the quantity optimized by the first active subspace apart from the divisor $\sqrt{2}$ affecting one of the d dimensions. We can also show directly that for $\boldsymbol{x} \sim \mathcal{N}(0, I)$ and $f(\boldsymbol{x}) = (1/2)\boldsymbol{x}^{\mathsf{T}}A\boldsymbol{x} + b^{\mathsf{T}}\boldsymbol{x}$ for a symmetric matrix A, the Sobol' criterion (4.1) reduces to $\theta^{\mathsf{T}}A^2\theta + (\theta^{\mathsf{T}}b)^2 - (1/2)(\theta^{\mathsf{T}}A\theta)^2$ compared to an active subspace criterion of $\theta^{\mathsf{T}}A^2\theta + (\theta^{\mathsf{T}}b)^2$.

4.2. Active subspace for preintegration. Because $C = \mathbb{E}(\nabla f(\boldsymbol{x})\nabla f(\boldsymbol{x})^{\mathsf{T}})$ is positive semidefinite, it has the eigen-decomposition $C = \Theta D\Theta^{\mathsf{T}}$, where $\Theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^{d\times d}$ is an orthogonal matrix consisting of eigenvectors of C, and $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$ with $\lambda_1 \geqslant \cdots \geqslant \lambda_d \geqslant 0$ being the eigenvalues. Constantine, Dow, and Wang [6] prove that there exists a constant c such that

$$\mathbb{E}\big(\big(f(\boldsymbol{x}) - \mathbb{E}(f(\boldsymbol{x})|\Theta[1{:}r]^{\mathsf{T}}\boldsymbol{x})\big)^2\big) \leqslant c(\lambda_{r+1} + \cdots + \lambda_d) \tag{4.4}$$

for all f with a square integrable gradient. In general, the Poincaré constant c depends on the support of the function and the probability measure. But for the multivariate standard Gaussian distribution, the Poincaré constant is always 1 [37].

In our problem, we take $r = 1$ and compute $\mathbb{E}(f(\boldsymbol{x}) \,|\, \theta^{\mathsf{T}}\boldsymbol{x})$, where θ is the first column of Θ. In practice, it is convenient to preintegrate the last variable of f_Θ, not the first. For instance, one would use the first $d-1$ components in a Sobol' sequence not components 2 through d. Taking θ to be the first column of Θ, we compute

$$g(\boldsymbol{x}_{-d}) = \int_{-\infty}^{\infty} f(\theta x_d + \Psi\boldsymbol{x}_{-d})\varphi(x_d)\,\mathrm{d}x_d,$$

where Ψ is orthogonal to θ (i.e., $\theta^{\mathsf{T}}\Psi = \mathbf{0}$), using a quadrature rule of negligible error or, if possible, a closed form. We then integrate this g over $d-1$ variables by RQMC. We can use $\Psi = \Theta[2{:}d]$. Or if we want to avoid the cost of computing the full eigendecomposition of C we can find θ by a power iteration and then use a Householder transformation $\tilde{\Theta} = I - 2\boldsymbol{w}\boldsymbol{w}^{\mathsf{T}}$, where $\boldsymbol{w} = (\theta - e_1)/\|\theta - e_1\|$ and $e_1 = (1, 0, 0, \ldots, 0)^{\mathsf{T}}$. This $\tilde{\Theta}$ is an orthogonal matrix whose first column is θ. In our numerical work, we have used $\Psi = \Theta[2{:}d]$ instead of the Householder transformation because of the effective dimension motivation for those eigenvectors given by [47].

We note that active subspaces use an uncentered PCA analysis of the matrix of sample gradients. One could use instead a centered analysis of $\mathbb{E}((\nabla f(\boldsymbol{x}) - \eta)(\nabla f(\boldsymbol{x}) - \eta)^{\mathsf{T}})$, where $\eta = \mathbb{E}(\nabla f(\boldsymbol{x}))$. The potential advantage of this is that $\nabla f(\boldsymbol{x}) - \eta$ is the gradient of $f(\boldsymbol{x}) - \eta^{\mathsf{T}}\boldsymbol{x}$ which subtracts a linear approximation from f before searching for θ. The rationale for this alternative is that RQMC might already do well integrating a linear function and we would then want to choose a direction θ that performs well for the nonlinear part of f. In our examples, we found very little difference between the two methods and so we proposed the simpler uncentered active subspace preintegration.

In practice, we must estimate C. In the above description, we replace C by

$$\widehat{C} = \frac{1}{M}\sum_{i=0}^{M-1} \nabla f(\boldsymbol{x}_i)\nabla f(\boldsymbol{x}_i)^{\mathsf{T}}, \tag{4.5}$$

Algorithm 4.1 Preintegration with active subspace

Input: Integrand f, number of samples M to compute $\widehat{C}$, number of samples n to compute $\hat{\mu}$

▷ Find active subspaces

Take $\boldsymbol{x}_0, \dots, \boldsymbol{x}_{M-1} \sim \mathcal{N}(0, I_d)$ by RQMC.

Compute $\widehat{C} = \frac{1}{M}\sum_{i=0}^{M-1} \nabla f(\boldsymbol{x}_i)\nabla f(\boldsymbol{x}_i)^\mathsf{T}$.

Compute the eigen-decomposition $\widehat{C} = \widehat{\Theta}\widehat{D}\widehat{\Theta}^\mathsf{T}$.

▷ Preintegration

Let $\theta = \widehat{\Theta}[1]$ and $\Psi = \widehat{\Theta}[2{:}d]$

Define $g : \mathbb{R}^{d-1} \to \mathbb{R}$ by $g(\boldsymbol{x}) = \int_{-\infty}^{\infty} f(\theta x_d + \Psi \boldsymbol{x}_{-d})\varphi(x_d)\,\mathrm{d}x_d$

▷ RQMC integration

Take $\boldsymbol{x}_0, \dots, \boldsymbol{x}_{n-1} \sim \mathcal{N}(0, I_{d-1})$ by RQMC and compute $\Psi x_0, \dots, \Psi x_{n-1}$.

return $\hat{\mu} = \frac{1}{n}\sum_{i=0}^{n-1} g(\boldsymbol{x}_i)$.

for an RQMC generated sample with $\boldsymbol{x}_i \sim \mathcal{N}(0, I_d)$ and then define θ and Θ_{-1} using $\widehat{C}$ in place of C. We summarize the procedure in Algorithm 4.1.

Using our prior notation we can now describe the approach of [46] more precisely. They first preintegrate one variable in closed form, producing a $d-1$ dimensional integrand. They then apply GPCA to the preintegrated function to find a good $d-1$ dimensional rotation matrix [6]. That is, they first find $h(\boldsymbol{x}_{2:d}) := \mathbb{E}(f(\boldsymbol{x})|\boldsymbol{x}_{2:d})$, then compute

$$\widehat{\widehat{C}} = \frac{1}{M}\sum_{i=0}^{M-1} \nabla h(\boldsymbol{x}_i)\nabla h(\boldsymbol{x}_i)^\mathsf{T} \in \mathbb{R}^{(d-1)\times(d-1)}, \quad \boldsymbol{x}_i \sim \mathcal{N}(0, I_{d-1}), \tag{4.6}$$

using RQMC points $\boldsymbol{x}_i$. Then they find the eigen-decomposition $\widehat{\widehat{C}} = \widehat{V}\widehat{\Lambda}\widehat{V}^\mathsf{T}$. Finally, they use RQMC to integrate the function $h(\widehat{V}\boldsymbol{x})$, where $\boldsymbol{x} \sim \mathcal{N}(0, I_{d-1})$. The main difference is that they apply preintegration to the original integrand $f(\boldsymbol{x})$, while we apply preintegration to the rotated integrand $f_\Theta(\boldsymbol{x}) = f(\Theta\boldsymbol{x})$. They conduct GPCA in the end as an approach to reduce effective dimension, while we conduct a similar GPCA (active subspace method) at the beginning to find the important subspace.

5. Application to option pricing. Here we study some Gaussian integrals arising from financial valuation. We assume that an asset price S_t, such as a stock, follows a geometric Brownian motion satisfying the stochastic differential equation (SDE)

$$\mathrm{d}S_t = rS_t\,\mathrm{d}t + \sigma S_t\,\mathrm{d}B_t,$$

where B_t is a Brownian motion under the risk-neutral measure. Here, r is the interest rate and $\sigma > 0$ is the constant volatility for the asset. For an initial price S_0, the SDE above has a unique solution

$$S_t = S_0 \exp\Big(\Big(r - \frac{\sigma^2}{2}\Big)t + \sigma B_t\Big).$$

Suppose that the maturity time of the option is T. In practice, we simulate a discrete Brownian motion. We call B a d-dimensional discrete Brownian motion if B follows a multivariate Gaussian distribution with mean zero and covariance Σ with $\Sigma_{ij} =$

$\Delta t \min(i,j)$, where $\Delta t = T/d$ is the length of each time interval and $1 \leqslant i,j \leqslant d$. To sample a discrete Brownian motion, we can first find a $d \times d$ matrix R such that $RR^{\mathsf{T}} = \Sigma$, then generate a standard Gaussian variable $\boldsymbol{z} \sim \mathcal{N}(0, I_d)$, and let $B = R\boldsymbol{z}$. Taking R to be the lower triangular matrix in the Cholesky decomposition of Σ yields the *standard construction*. Using the usual eigen-decomposition $\Sigma = U\Lambda U^{\mathsf{T}}$, we can take $R = U\Lambda^{1/2}$. This is called the *principal component analysis (PCA)* construction. For explicit forms of both these choices of R, see [12].

5.1. Option with one asset. When we use the matrix R, we can approximate $S_{j\Delta t}$ by

$$S_j = S_0 \exp\Big(\Big(r - \frac{\sigma^2}{2}\Big) j\Delta t + \sigma B_j\Big), \quad 1 \leqslant j \leqslant d$$

where $B = R\boldsymbol{z}$ is the discrete Brownian motion and B_j is the jth coordinate of B. The arithmetic average of the stock price is given by

$$\bar{S}(R, \boldsymbol{z}) = \frac{S_0}{d} \sum_{j=1}^{d} \exp\left(\left(r - \frac{\sigma^2}{2}\right) j\Delta t + \sigma \sum_{k=1}^{d} R_{jk} z_k\right).$$

Then the expected payoff of the arithmetic average Asian call option with strike price K is $\mathbb{E}((\bar{S}(R, \boldsymbol{z}) - K)_+)$, where the expectation is taken over $\boldsymbol{z} \sim \mathcal{N}(0, I_d)$.

Suppose that we want to preintegrate with respect to z_1 before computing the expectation $\mathbb{E}((\bar{S}(R, \boldsymbol{z}) - K)_+)$. If $R_{j1} > 0$ for all $1 \leqslant j \leqslant d$, then $\bar{S}(R, \boldsymbol{z})$ is increasing in z_1 for any value of $\boldsymbol{z}_{2:d}$. If we can find $\gamma = \gamma(\boldsymbol{z}_{2:d})$ such that

$$\bar{S}(R, (\gamma, \boldsymbol{z}_{2:d})) = K, \tag{5.1}$$

then the preintegration step becomes

$$\begin{aligned}
&\mathbb{E}((\bar{S}(R, \boldsymbol{z}) - K)_+ \,|\, \boldsymbol{z}_{2:d}) \\
&= \int_{z_1 \geqslant \gamma(\boldsymbol{z}_{2:d})} (\bar{S}(R, (z_1, \boldsymbol{z}_{2:d})) - K) \varphi(z_1)\, \mathrm{d}z_1 \\
&= \frac{S_0}{d} \sum_{j=1}^{d} \exp\left(\left(r - \frac{\sigma^2}{2}\right) j\Delta t + \sigma \sum_{k=2}^{d} R_{jk} z_k + \frac{\sigma^2 R_{j1}^2}{2}\right) \bar{\Phi}(\gamma - \sigma R_{j1}) - K\bar{\Phi}(\gamma),
\end{aligned} \tag{5.2}$$

where $\bar{\Phi}(x) = 1 - \Phi(x)$. In practice, (5.1) can be solved by a root finding algorithm [15]. For example, Newton iteration usually converges in only a few steps. The monotonicity in z_1 is important for the preintegration step to be carried out easily. Without monotonicity, there might exist multiple roots for (5.1). For instance, if there are two roots γ_1, γ_2, we need to integrate over $[\gamma_1, \gamma_2]$ or $(-\infty, \gamma_1) \cup (\gamma_2, +\infty)$. In more general cases, we might have to use a high-precision quadrature rule for the preintegration step.

The condition that $R_{j1} \geqslant 0$ for $1 \leqslant j \leqslant d$ is satisfied when we use the standard construction or PCA construction of Brownian motion. With the active subspace method, we are using $\tilde{R} = R\Theta$ in the place of R where Θ consists of the eigenvectors of $C = \mathbb{E}(\nabla f(\boldsymbol{z}) \nabla f(\boldsymbol{z})^{\mathsf{T}})$, and here $f(\boldsymbol{z}) = (\bar{S}(R, \boldsymbol{z}) - K)_+$. In this example, f is not differentiable at those $\boldsymbol{z}$ where $\bar{S}(R, \boldsymbol{z}) = K$. But those nondifferentiable points have zero probability, so with probability one $\nabla f(\boldsymbol{z})$ exists. We use the analytic form of the gradients when computing $\widehat{C}$ in (4.5). Now we show that $\tilde{R}_{j1} \geqslant 0$ for all j. We start from the standard construction, meaning that $R_{ij} = \sqrt{\Delta_t}\mathbf{1}\{i \geqslant j\}$; then $f(\boldsymbol{z})$ is

increasing in all coordinates of $\boldsymbol{z}$. So $\nabla f(\boldsymbol{z})$, whenever it exists, is always nonnegative in all coordinates, which means that $\widehat{C}_{ij} \geqslant 0$ for all $1 \leqslant i, j \leqslant d$. Therefore, the first eigenvector of $\widehat{C}$ has the same sign in all of its components. Without loss of generality, we can take $\Theta_{j1} \geqslant 0$. So the first column of $\tilde{R} = R\Theta$ is also nonnegative. This proves that the preintegration step for the active subspace integrand can also be easily conducted by a Newton search.

The arXiv preprint of this paper [24] includes numerical results for some of the Greeks for the Asian option. Those Greeks have integrands that are not only nondifferentiable but actually discontinuous at $\boldsymbol{z}$, where $\bar{S}(R, \boldsymbol{z}) = K$, so the motivation for active subspaces does not apply to them. Our method in that report estimated C by the expected value of $\nabla f(\boldsymbol{z}) \nabla f(\boldsymbol{z})^\mathsf{T}$ taken over $\boldsymbol{z}$ where $\nabla f(\boldsymbol{z})$ exists. This leads to a direction θ influenced by the gradient information while ignoring the jump discontinuities. Preintegrating over the resulting active subspace direction cannot raise variance and it produced very competitive results for those examples but this may be because the options ended up in the money (i.e., with $\bar{S}(R, \boldsymbol{z}) > K$) about half of the time. Choosing a direction θ to preintegrate over for discontinuous integrands is out of the scope of the present paper.

We compare Algorithm 4.1 with other methods in the option pricing example considered in [46] and [17]. We take the parameters $d = 50$, $T = 1$, $\sigma = 0.4$, $r = 0.1$, $S_0 = K = 100$ the same as in [17]. We consider five methods:

- `AS + preint`: our proposed active subspace preintegration method (Algorithm 4.1), which applies the active subspace method to find the direction to preintegrate;
- `preint+DimRed`: the method proposed in [46], which first preintegrates z_1 and applies GPCA to conduct dimension reduction for the other $d - 1$ variables;
- `preint`: preintegrating z_1 with no dimension reduction;
- `RQMC`: usual RQMC; and
- `MC`: plain Monte Carlo.

The first three methods, `AS + preint`, `preint+DimRed`, and `preint`, all use RQMC sampling unless otherwise specified.

For each sample size n, we repeat the simulation 50 times and compute the standard error across the 50 replicates. The standard error (std. err.) is plotted versus the sample size (n) on the log-log scale. For the methods `AS + preint` and `preint+DimRed`, we use $M = 128$ samples to estimate C as in (4.5). We approximate the gradients of the preintegrated integrand g by the finite difference

$$\nabla g(\boldsymbol{x}) \approx \left(\frac{g(\boldsymbol{x} + \varepsilon \boldsymbol{e}_1) - g(\boldsymbol{x})}{\varepsilon}, \ldots, \frac{g(\boldsymbol{x} + \varepsilon \boldsymbol{e}_{d-1}) - g(\boldsymbol{x})}{\varepsilon} \right)^\mathsf{T}, \quad \varepsilon = 10^{-6},$$

matching the choice in [46]. We chose a small value of M to keep the costs comparable to plain RQMC. Also because θ is a local maximizer of $\theta^\mathsf{T} C \theta$, we have $\hat{\theta}^\mathsf{T} C \hat{\theta} = \theta^\mathsf{T} C \theta + O(\|\hat{\theta} - \theta\|^2)$, so there are diminishing benefits to accurate estimation of θ. Finally, any θ where f varies along $\theta^\mathsf{T} \boldsymbol{x}$ brings a variance reduction.

In Figure 1, we show the results under the standard construction (left) and the PCA construction (right) of the Brownian motion. We observe that `AS + preint` makes a huge improvement under the standard construction and is about the same as `preint+DimRed` under the PCA construction. The performance of active subspace preintegration is the same under either the standard or the PCA construction by invariance. For the Asian call option, it is already well known that the PCA construction is especially effective. Active subspace preintegration finds something almost as good without special knowledge.

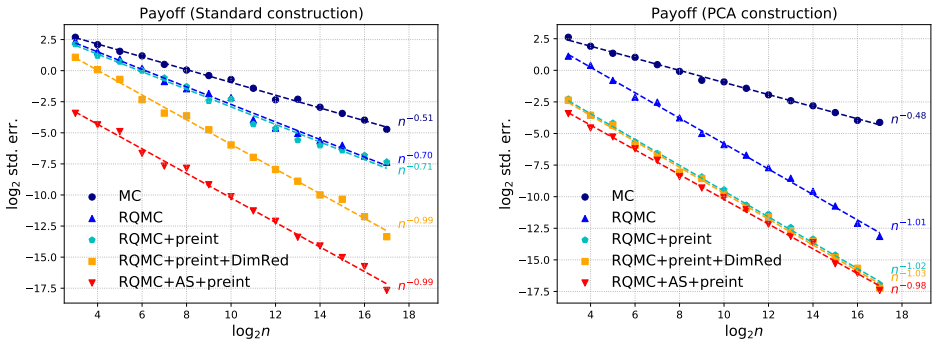

FIG 1. *Single-asset call option.*

Running time is another measure of efficiency. In our examples, preintegration increases the running time by roughly 12-fold compared to methods that do not preintegrate. But that is not enough to offset the gains from preintegrating the first principal component or the first active subspace component for large n. In Figure 1, the improvement of `AS + preint` over plain RQMC is about 1000-fold under standard construction and 30-fold under PCA construction for large n. We also observe that the proposed `AS + preint` method is faster than `preint+DimRed` in terms of computing $\widehat{C}$ and $\widehat{\widetilde{C}}$. This is because we only need to evaluate the original integrand to compute $\widehat{C}$ but we need to evaluate the preintegrated integrand to compute $\widehat{\widetilde{C}}$. Evaluating the preintegrated integrand is more expensive because we need to apply Newton iterations for root-finding. The cost of computing estimates of C is negligible because it is done by inspecting only $M \ll n$ points in the domain of f. We put more details of the timings in the supplement (SM4).

5.2. Basket option. A basket option depends on a weighted average of several assets. Suppose that under the risk-neutral measure the L assets $S^{(1)}, \dots, S^{(L)}$ follow the SDE

$$\mathrm{d}S_t^{(\ell)} = rS_t^{(\ell)}\,\mathrm{d}t + \sigma_\ell S_t^{(\ell)}\,\mathrm{d}B_t^{(\ell)},$$

where $\{B^{(\ell)}\}_{1\leqslant\ell\leqslant L}$ are standard Brownian motions with correlation $\mathrm{Corr}(B_t^{(\ell)}, B_t^{(k)}) = \rho_{\ell k}$ for all $t > 0$. For some nonnegative weights $w_1 + \cdots + w_L = 1$, the payoff function of the Asian basket call option is given by

$$\left(\sum_{\ell=1}^{L} w_\ell \bar{S}^{(\ell)} - K\right)_+$$

where $\bar{S}^{(\ell)}$ is the arithmetic average of $S_t^{(\ell)}$ in the time interval $[0, T]$. Here, we only consider $L = 2$ assets. To generate $B^{(1)}, B^{(2)}$ with correlation ρ, we can generate two independent standard Brownian motions $W^{(1)}, W^{(2)}$ and let $B^{(1)} = W^{(1)}$, $B^{(2)} = \rho W^{(1)} + \sqrt{1-\rho^2}W^{(2)}$. Following the same discretization as before, we can generate $(\boldsymbol{z}^\mathsf{T}, \tilde{\boldsymbol{z}}^\mathsf{T}) \sim \mathcal{N}(0, I_{2d})$. Then for time steps $j = 1, \dots, d$, let

$$S_j^{(1)} = S_0^{(1)} \exp\left(\left(r - \frac{\sigma_1^2}{2}\right) j\Delta t + \sigma_1 \sum_{k=1}^{d} R_{jk} z_k\right), \quad \text{and}$$

$$S_j^{(2)} = S_0^{(2)} \exp\left(\left(r - \frac{\sigma_2^2}{2}\right) j\Delta t + \sigma_2 \left(\rho \sum_{k=1}^{d} R_{jk} z_k + \sqrt{1-\rho^2} \sum_{k=1}^{d} R_{jk} \tilde{z}_k\right)\right).$$

Again, the matrix R can be constructed by the standard construction or the PCA construction. We call these methods the ordinary standard construction and ordinary PCA construction.

A sharper principal components analysis would merge the two Brownian motions into a single $2d$-dimensional process and use the principal components from their joint covariance matrix. The processes $B^{(1),\sigma_1} = \sigma_1 B^{(1)}$ and $B^{(2),\sigma_2} = \sigma_2 B^{(2)}$ have joint distribution

$$\begin{pmatrix} B^{(1),\sigma_1} \\ B^{(2),\sigma_2} \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} \sigma_1^2\Sigma & \rho\sigma_1\sigma_2\Sigma \\ \rho\sigma_1\sigma_2\Sigma & \sigma_2^2\Sigma \end{pmatrix}\right).$$

Let $\widetilde{\Sigma}$ be the joint covariance matrix above. We can pick $\tilde{R}$ with $\tilde{R}\tilde{R}^{\mathsf{T}} = \tilde{\Sigma}$ and let

$$\begin{pmatrix} B^{(1),\sigma_1} \\ B^{(2),\sigma_2} \end{pmatrix} = \tilde{R}\begin{pmatrix} \boldsymbol{z} \\ \tilde{\boldsymbol{z}} \end{pmatrix}, \quad \text{where} \quad \begin{pmatrix} \boldsymbol{z} \\ \tilde{\boldsymbol{z}} \end{pmatrix} \sim \mathcal{N}(0, I_{2d}).$$

The matrix $\tilde{R}$ can be found by either a Cholesky decomposition or eigendecomposition of $\tilde{R}$. We call this method joint standard construction or joint PCA construction. With $B^{(1),\sigma_1}$ and $B^{(2),\sigma_1}$ generated, we can compute

$$S_j^{(\ell)} = S_0^{(\ell)} \exp\left(\left(r - \frac{\sigma_\ell^2}{2}\right) j\Delta t + \sum_{k=1}^{d} B_k^{(1),\sigma_\ell}\right).$$

In the preintegration step, we choose to integrate out z_1. This can be easily carried out similarly as in (5.1) and (5.2) provided that the first column of $\tilde{R}$ is nonnegative. This is true if $\tilde{R}$ is found by the active subspaces method, following a similar argument in the previous example. We take $d = 50$, $T = 1$, $\rho = 0.5$, $S_0^{(1)} = S_0^{(2)} = 100$, $r = 0.1$, $\sigma_1 = 0.1$, $\sigma_2 = 0.2$, $w_1 = 0.8$, $w_2 = 0.2$. The standard errors under ordinary standard construction and PCA construction are plotted in Figures 2 and 3. The results for joint standard and PCA constructions are in Figure 4. The three columns correspond to $K = 80$, 100, and 120, respectively. A few observations are in order:

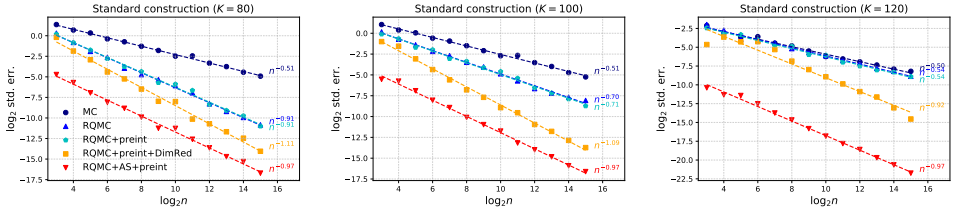

FIG 2. *Basket options under ordinary standard construction.*

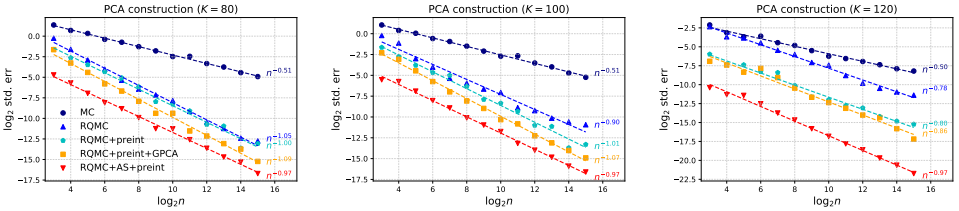

FIG 3. *Basket options under ordinary PCA construction.*

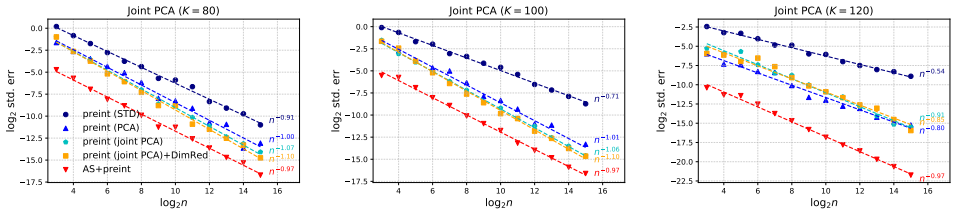

FIG 4. *Basket options under joint standard or joint PCA construction with preintegration.*

1. For the standard construction, preintegrating over z_1 brings little improvement over plain RQMC. But for the PCA construction, preintegrating over z_1 brings a big variance reduction for $K = 100$ and 120.
2. The dimension reduction technique from [46] largely improves the error rate compared to preintegration without dimension reduction under the ordinary standard construction and ordinary PCA construction. This improvement is particularly significant for the ordinary standard construction and for smaller K.
3. The proposed method `AS+preint` has the best performance in all settings. It is even better than preintegrating out the first joint principal component of the Brownian motion with dimension reduction. In this example, the active subspace method is able to find a better direction than the first principal component over which to preintegrate.
4. For the $K = 120$ case, the probability that the option is exercised is about 2%. When we compute $\widehat{C}$ with 128 samples, we only get five gradients that are not zero vector. But it is enough to find a better preintegration direction.

5.3. Rainbow option. A rainbow option is also a multiasset option whose payoff depends on the maximum or the minimum price across the assets. Here we consider the Asian *call on max* and *call on min* options, whose payoffs are defined as

$$\text{call on max: } \mathbb{E}\left[\left(\max(\bar{S}^{(1)}, \bar{S}^{(2)}) - K\right)_+\right] \quad \text{and}$$
$$\text{call on min: } \mathbb{E}\left[\left(\min(\bar{S}^{(1)}, \bar{S}^{(2)}) - K\right)_+\right],$$

respectively. Here, $\bar{S}^{(1)}$ and $\bar{S}^{(2)}$ are the same as in the basket option.

For the above two integrands, we still have the monotonicity in z_1 if $R_{\cdot 1} \geqslant 0$; thus we can find the threshold $\gamma = \gamma(z_{-1})$ such that $\max(\bar{S}^{(1)}, \bar{S}^{(2)}) = K$ or $\min(\bar{S}^{(1)}, \bar{S}^{(2)}) = K$ by Newton iterations. After finding γ, the expectation over x_1 can be written as

$$\int_{\gamma}^{\infty} \max(\bar{S}^{(1)}, \bar{S}^{(2)}) \varphi(x_1)\, \mathrm{d}x_1 - K\bar{\Phi}(\gamma).$$

For the call on min option, "max" is replaced by "min" in the last display. However, the above integral does not have a closed form. Therefore, we use a Gaussian quadrature to compute the integral over the interval $[\gamma, 10]$ using the quadrature function from SciPy [41].

We take $r = 0.1$, $\sigma_1 = 0.2$, $\sigma_2 = 0.1$, $S_0^{(1)} = 120$, $S_0^{(2)} = 80$. For the call on max option, we take $K \in \{100, 120\}$. For the call on min option, we take $K \in \{80, 100\}$. The results are shown in Figure 5. The call on min option with $K = 100$ is extremely out-of-money so we increase M to 2048 when computing $\widehat{C}$ for the active subspace

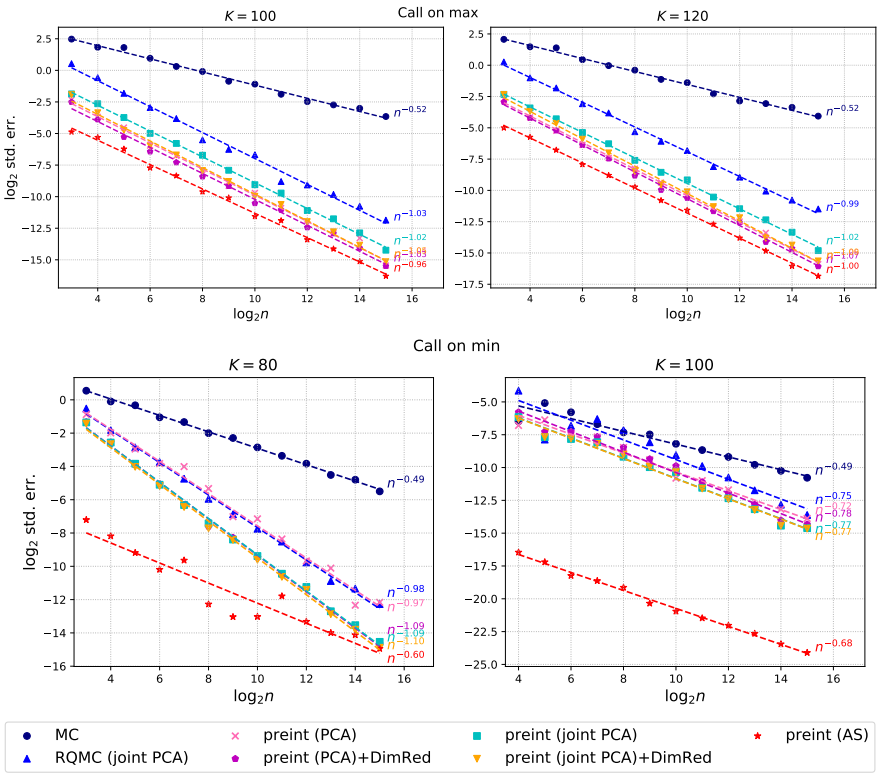

FIG 5. *Rainbow options. The top panel is for the call on max option with* $K \in \{100, 120\}$ *and the bottom panel is for the call on min option with* $K \in \{80, 100\}$.

method. We get three nonzero gradients. We also increase the number of replicates to 100 for this integrand. In the plot, n starts from 2^4 because for $n = 2^3$ some methods only get 0 in all 100 replicates.

Several observations are in order:

1. For the call on max option, preintegration with ordinary PCA is slightly better than preintegration with joint PCA. This is because for the call on max option, the maximum is most likely achieved by $\bar{S}^{(1)}$. So the integrand behaves like the single asset payoff in section 5.1. When we preintegrate z_1 under the ordinary PCA construction, we are preintegrating the first principal component of the Brownian motion for $\bar{S}^{(1)}$. And this is known to be quite effective.
2. For the call on min option, preintegration with joint PCA is better than preintegration with ordinary PCA. This is because the minimum is mostly achieved by $\bar{S}^{(2)}$. So to make preintegration under ordinary PCA more effective, one might consider exchanging the order of $\bar{S}^{(1)}$ and $\bar{S}^{(2)}$. But it is still not as effective as the active subspace method. See Figure SM4 in the supplement.
3. The proposed method, preintegration with the active subspace integrand, outperforms all other methods without using any specific knowledge about the integrands.

6. Discussion. In this paper we have studied a kind of conditional RQMC known as preintegration. We found that, just like conditional MC, the procedure can reduce variance but cannot increase it. We proposed to preintegrate over the first component in the active subspace. We showed a close relationship between this choice of preintegration variable and what one would get using a computationally infeasible but well-motivated choice by maximizing the Sobol' index of a linear combination of variables. The proposed method outperforms previous methods in some option pricing problems, especially when there is no strong incumbent construction of the Brownian motion.

Acknowledgments. We thank Dirk Nuyens and two anonymous reviewers for comments that have helped us improve this paper.

REFERENCES

[1] N. Achtsis, R. Cools, and D. Nuyens, *Conditional sampling for barrier option pricing under the LT method*, SIAM J. Financial Math., 4 (2013), pp. 327–352.
[2] P. Acworth, M. Broadie, and P. Glasserman, *A comparison of some Monte Carlo techniques for option pricing*, in Monte Carlo and Quasi-Monte Carlo Methods '96, Springer, New York, 1997, pp. 1–18.
[3] D. Bigoni, Y. Marzouk, C. Prieur, and O. Zahm, *Nonlinear Dimension Reduction for Surrogate Modeling Using Gradient Information*, Tech. report, https://arxiv.org/abs/2102.10351, 2021.
[4] W. Chen, A. Srivastav, and G. Travaglini, eds., *A Panorama of Discrepancy Theory*, Springer, Cham, Switzerland, 2014.
[5] P. G. Constantine, *Active Subspaces: Emerging Ideas for Dimension Reduction in Parameter Studies*, SIAM, Philadelphia, 2015.
[6] P. G. Constantine, E. Dow, and Q. Wang, *Active subspace methods in theory and practice: Applications to Kriging surfaces*, SIAM J. Sci. Comput., 36 (2014), pp. A1500–A1524.
[7] P. J. Davis and P. Rabinowitz, *Methods of Numerical Integration*, 2nd ed., Academic Press, San Diego, 1984.
[8] J. Dick and F. Pillichshammer, *Digital Sequences, Discrepancy and Quasi-Monte Carlo Integration*, Cambridge University Press, Cambridge, UK, 2010.
[9] B. Efron and C. Stein, *The jackknife estimate of variance*, Ann. Statist., 9 (1981), pp. 586–596.
[10] C. J. Geyer, *Conditioning in Markov chain Monte Carlo*, J. Comput. Graph. Statist., 4 (1995), pp. 148–154.
[11] A. D. Gilbert, F. Y. Kuo, and I. H. Sloan, *Preintegration Is Not Smoothing when Monotonicity Fails,* Tech. report, https://arxiv.org/ abs/2112.11621, 2021.
[12] P. Glasserman, *Monte Carlo Methods in Financial Engineering*, Stoch. Model. Appl. Probab. 53, Springer, New York, 2004.
[13] P. Glasserman, P. Heidelberger, and P. Shahabuddin, *Asymptotically optimal importance sampling and stratification for pricing path-dependent options*, Math. Finance, 9 (1999), pp. 117–152.
[14] M. Griebel, F. Y. Kuo, and I. H. Sloan, *The smoothing effect of the ANOVA decomposition*, J. Complexity, 26 (2010), pp. 523–551.
[15] A. Griewank, F. Y. Kuo, H. Leövey, and I. H. Sloan, *High dimensional integration of kinks and jumps—smoothing by preintegration*, J. Comput. Appl. Math., 344 (2018), pp. 259–274.
[16] J. M. Hammersley, *Conditional Monte Carlo*, J. ACM, 3 (1956), pp. 73–76.
[17] Z. He, *On the error rate of conditional quasi–Monte Carlo for discontinuous functions*, SIAM J. Numer. Anal., 57 (2019), pp. 854–874.
[18] W. Hoeffding, *A class of statistics with asymptotically normal distribution*, Ann. Math. Stat., 19 (1948), pp. 293–325.
[19] C. R. Hoyt and A. B. Owen, *Mean dimension of ridge functions*, SIAM J. Numer. Anal., 58 (2020), pp. 1195–1216.
[20] J. Imai and K. S. Tan, *Minimizing effective dimension using linear transformation*, in Monte Carlo and Quasi-Monte Carlo Methods 2002, Springer, New York, 2004, pp. 275–292.
[21] M. J. W. Jansen, *Analysis of variance designs for model output*, Comput. Phys. Commun., 117 (1999), pp. 35–43.

[22] I. T. Jolliffe, *A note on the use of principal components in regression*, J. R. Stat. Soc. Ser. C. Appl. Stat., 31 (1982), pp. 300–303.
[23] P. L'Ecuyer and C. Lemieux, *Variance reduction via lattice rules*, Management Sci., 46 (2000), pp. 1214–1235.
[24] S. Liu and A. B. Owen, *Pre-integration via Active Subspaces,* Tech. report, https://arxiv.org/abs/2202.02682, 2022.
[25] J. Matoušek, *On the L^2–discrepancy for anchored boxes*, J. Complexity, 14 (1998), pp. 527–556.
[26] B. Moskowitz and R. E. Caflisch, *Smoothness and dimension reduction in quasi-Monte Carlo methods*, Math. Comput. Model., 23 (1996), pp. 37–54.
[27] H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods*, SIAM, Philadelphia, 1992.
[28] G. Ökten and A. Göncü, *Generating low-discrepancy sequences from the normal distribution: Box-Muller or inverse transform?*, Math. Comput. Model., 53 (2011), pp. 1268–1281.
[29] A. B. Owen, *Randomly permuted (t, m, s)-nets and (t, s)-sequences*, in Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing, Springer-Verlag, New York, 1995, pp. 299–317.
[30] A. B. Owen, *Monte Carlo variance of scrambled net quadrature*, SIAM J. Numer. Anal., 34 (1997), pp. 1884–1910.
[31] A. B. Owen, *Scrambled net variance for integrals of smooth functions*, Ann. Statist., 25 (1997), pp. 1541–1562.
[32] A. B. Owen, *Scrambling Sobol' and Niederreiter-Xing points*, J. Complexity, 14 (1998), pp. 466–489.
[33] A. B. Owen, *Local antithetic sampling with scrambled nets*, Ann. Statist, 36 (2008), pp. 2319–2343.
[34] A. B. Owen, *Monte Carlo Theory, Methods and Examples*, statweb.stanford.edu/∼owen/mc, 2013.
[35] A. B. Owen and D. Rudolf, *A strong law of large numbers for scrambled net integration*, SIAM Rev., 63 (2021), pp. 360–372.
[36] A. Papageorgiou, *The Brownian bridge does not offer a consistent advantage in quasi-Monte Carlo integration*, J. Complexity, 18 (2002), pp. 171–186.
[37] M. T. Parente, J. Wallin, and B. Wohlmuth, *Generalized bounds for active subspaces*, Electron. J. Stat., 14 (2020), pp. 917–943.
[38] G. Pirsic, *Schnell konvergierende Walshreihen über gruppen*, Master's thesis, University of Salzburg, Institute for Mathematics, 1995.
[39] S. Razavi et al., *The future of sensitivity analysis: An essential discipline for systems modeling and policy support*, Environ. Modell. Soft., 137 (2021), 104954.
[40] C. P. Robert and G. O. Roberts, *Rao-Blackwellization in the MCMC Era*, University of Warwick, Tech. report, https://arxiv.org/abs/2101.01011, 2021.
[41] *SciPy* 1.0*: Fundamental algorithms for scientific computing in Python*, Nat. Met., 17 (2020), pp. 261–272, https://doi.org/10.1038/s41592-019-0686-2.
[42] I. M. Sobol', *Multidimensional Quadrature Formulas and Haar Functions*, Nauka, Moscow, 1969 (in Russian).
[43] I. M. Sobol', *Sensitivity estimates for nonlinear mathematical models*, Math. Model. Comput. Experiment, 1 (1993), pp. 407–414.
[44] I. M. Sobol' and S. Kucherenko, *Derivative based global sensitivity measures and their link with global sensitivity indices*, Math. Comput. Simulation, 79 (2009), pp. 3009–3017.
[45] H. F. Trotter and J. W. Tukey. *Conditional Monte Carlo for normal samples,* in Symposium on Monte Carlo Methods, Wiley, New York, 1956, pp. 64–79.
[46] Y. Xiao and X. Wang, *Conditional quasi-Monte Carlo methods and dimension reduction for option pricing and hedging with discontinuous functions*, J. Comput. Appl. Math., 343 (2018), pp. 289–308.
[47] Y. Xiao and X. Wang, *Enhancing quasi-Monte Carlo simulation by minimizing effective dimension for derivative pricing*, Comput. Econ., 54 (2019), pp. 343–366.
[48] R.-X. Yue and S.-S. Mao, *On the variance of quadrature over scrambled nets and sequences*, Statist. Probab. Lett., 44 (1999), pp. 267–280.
[49] O. Zahm, P. Constantine, C. Prieur, and Y. Marzouk, *Gradient-based dimension reduction of multivariate vector-valued functions*, SIAM J. Sci. Comput., 42 (2020), pp. A534–A558.