Some Large Deviations Results For Latin Hypercube Sampling

Shane S. Drew
Tito Homem-de-Mello

Department of Industrial Engineering
and Management Sciences
Northwestern University
Evanston, IL 60208-3119, U.S.A.

Abstract

Large deviations theory is a well-studied area which has shown to have numerous applications. Broadly speaking, the theory deals with analytical approximations of probabilities of certain types of rare events. Moreover, the theory has recently proven instrumental in the study of complexity of approximations of stochastic optimization problems. The typical results, however, assume that the underlying random variables are either i.i.d. or exhibit some form of Markovian dependence. Our interest in this paper is to study the validity of large deviations results in the context of estimators built with *Latin Hypercube sampling*, a well-known sampling technique for variance reduction. We show that a large deviation principle holds for Latin Hypercube sampling for functions in one dimension and for separable multi-dimensional functions. Moreover, the upper bound of the probability of a large deviation in these cases is no higher under Latin Hypercube sampling than it is under Monte Carlo sampling. We extend the latter property to functions that are monotone in each argument. Numerical experiments illustrate the theoretical results presented in the paper.

1 Introduction

Suppose we wish to calculate $\mathbb{E}[g(X)]$ where $X = [X^1, \ldots, X^d]$ is a random vector in $\mathbb{R}^d$ and $g(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}$ is a measurable function. Further, suppose that the expected value is finite and cannot be written in closed form or be easily calculated, but that $g(X)$ can be easily computed for a given value of X. Let $\mathbb{E}[g(X)] = \mu \in (-\infty, \infty)$. To estimate the expected value, we can use the sample average approximation:

$$\frac{1}{n} S_n = \frac{1}{n} \sum_{i=1}^{n} g(X_i) \tag{1}$$

where the X_i are random samples from the distribution of X. When the X_i are i.i.d. (i.e. Monte Carlo sampling), by the law of large numbers the sample average approximation should approach the true mean μ (with probability one) as the number of samples n becomes large. In that context, large deviations theory identifies situations where the probability that the sample average approximation deviates from μ by a fixed amount $\delta > 0$ approaches zero exponentially fast as n goes to infinity.

Formally, this is expressed as

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\left| \frac{1}{n} S_n - \mu \right| > \delta \right) \;=\; -\beta_\delta, \tag{2}$$

where β_δ is a positive constant.

The above description, of course, is a small fraction of a much more general theory, but conveys a basic concept — that one obtains *exponential convergence* of probabilities involving sample average estimators under certain conditions. This idea has found applications in numerous areas, from statistical mechanics to telecommunications to physics. The theory is particularly useful as a tool to estimate probabilities of rare events, both analytically as well as via simulation; we refer to classical books in the area such as Shwartz and Weiss (1995), Dembo and Zeitouni (1998) and Bucklew (2004) for further discussions. Recently, the exponential bounds that lie at the core of large deviations theory have been used by Ökten, Tuffin, and Burago (2006) to show properties of the discrepancy of high-dimensional sequences constructed via padding quasi-Monte Carlo with Monte Carlo; again, we refer to that paper for details.

Despite the exponential convergence results mentioned above, it is well known that Monte Carlo methods have some drawbacks, particularly when one wants to calculate the *errors* corresponding to given estimates. Although the theory behind such calculations — notably the Central Limit Theorem — is solid, in practice the error may be large even for large sample sizes. That has led to the development of many variance reduction techniques as well as alternative sampling methods (see, e.g., Law and Kelton 2000 for a general discussion of this topic).

One alternative approach for sampling the X_i is called *Latin Hypercube sampling* (LHS, for short), introduced by McKay, Beckman, and Conover (1979). Broadly speaking, the method calls for splitting each dimension into n strata (yielding n^d hypercubes) and, for every dimension, sampling all n strata exactly once. This technique has been extensively used in practice, not only because of simplicity of implementation but also because of its nice properties. Indeed, McKay et al. (1979) show that if $g(\cdot)$ is monotone in all of its arguments, then the variance of the estimator obtained with LHS (call it Var^{LHS}) is no higher than the sample variance from Monte Carlo sampling (Var^{MC}). Hoshino and Takemura (2000) extend this result to the case where $g(\cdot)$ is monotone in all but one of its arguments. Stein (1987) writes the ANOVA decomposition of g, i.e.,

$$g(X) = \mu + g_1(X^1) + \cdots + g_d(X^d) + g_{resid}(X) \tag{3}$$

and shows that, asymptotically, the sample variance from Latin Hypercube sampling is just equal to the variance of the residual term and is no worse than the variance from Monte Carlo sampling. Loh (1996) extends this result to the multivariate case where $g : \mathbb{R}^d \mapsto \mathbb{R}^m$. Owen (1997) shows that for any n and any function g, $\mathrm{Var}^{LHS} \le \frac{n}{n-1}\mathrm{Var}^{MC}$. Also, Owen (1992) shows that LHS satisfies a Central Limit Theorem with the variance equal to the variance of the residual term.

The above discussion shows that the LHS method has been well studied and possesses many nice properties. However, to the best of our knowledge there have been no studies on the exponential convergence of probabilities involving estimators obtained with LHS. The difficulty lies in that, since the X_i are no longer i.i.d. under LHS, Cramér's Theorem (which is the basic pillar of the results for i.i.d. sampling) can no longer be applied.

In this paper, we study the above problem. We identify conditions under which large deviations results hold under Latin Hypercube sampling. More specifically, our results apply when the integrand function is of one of the following types: one-dimensional; multi-dimensional but separable (i.e. functions with no residual term); multi-dimensional with a bounded residual term; and multi-dimensional functions that are monotone in each component. In the case of functions with a bounded residual term, our results hold provided that the deviation we are measuring is large enough. Further, in all the above situations, we show that the upper bound for the large deviations probability is lower under LHS than under Monte Carlo sampling. Jin, Fu, and Xiong (2003) show this property holds when negatively dependent sampling is used to estimate a probability quantile of continuous distributions. A preliminary analysis of the cases discussed in the present paper is reported in Drew and Homem-de-Mello (2005).

The particular application that motivates our work arises in the context of stochastic optimization. For completeness, we briefly review the main concepts here. Consider a model of the form

$$\min_{y \in \mathcal{Y}} \left\{ \psi(y) := \mathbb{E}[\Psi(y, X)] \right\}, \tag{4}$$

where $\mathcal{Y}$ is a subset of $\mathbb{R}^m$, X is a random vector in $\mathbb{R}^d$ and $\Psi : \mathbb{R}^m \times \mathbb{R}^d \rightarrowtail \mathbb{R}$ is a real valued function. We refer to (4) as the "true" optimization problem. The class of problems of the form (4) is quite large and include, for example, many problems is statistics, finance and operations research, to name a few. Let y^* denote the optimal solution of (4) (assume for simplicity this solution is unique), and let ν^* denote the optimal value of (4).

Consider now a family $\{\hat{\psi}_n(\cdot)\}$ of random approximations of the function $\psi(\cdot)$, each $\hat{\psi}_n(\cdot)$ being defined as

$$\hat{\psi}_n(y) \ := \ \frac{1}{n} \sum_{j=1}^{n} \Psi(y, X_j), \tag{5}$$

where $X_1, \ldots, X_n$ are independent and identically distributed samples from the distribution of X. Then, one can construct the corresponding approximating program

$$\min_{y \in \mathcal{Y}} \hat{\psi}_n(y). \tag{6}$$

An optimal solution $\hat{y}_n$ of (6) provides an approximation (an estimator) of the optimal solution y^* of the true problem (4). Similarly, the optimal value $\hat{\nu}_n$ of (6) provides an approximation of the optimal value ν^* of (4).

Many results describing the convergence properties of $\{\hat{y}_n\}$ and $\{\hat{\nu}_n\}$ exist; see, for instance, Shapiro (1991), King and Rockafellar (1993), Kaniovski, King, and Wets (1995), Robinson (1996), Dai, Chen, and Birge (2000). Broadly speaking, these results ensure that, under mild conditions, $\hat{y}_n$ converges to y^* and $\hat{\nu}_n$ converges to ν^*. In case the function $\Psi(\cdot, x)$ is *convex* and *piecewise linear* for all x — which is setting in stochastic linear programs — and the distribution of X has finite support, a stronger property holds; namely, the probability that $\hat{y}_n$ *coincides with* y^* goes to one exponentially fast, i.e.,

$$\lim_{n\to\infty} \frac{1}{n} \log\left[\mathbb{P}(\hat{y}_n \neq y^*)\right] \;=\; -\beta \tag{7}$$

for some $\beta > 0$ — see Shapiro and Homem-de-Mello (2000), Shapiro, Homem-de-Mello, and Kim (2002). A similar property holds in case the feasibility set $\mathcal{Y}$ is finite (Kleywegt, Shapiro, and Homem-de-Mello, 2001). The importance of results of this type lies in that they allow for an estimation of a sample size that is large enough to ensure that one obtains an ε-optimal solution with a given probability; for example, in the case of finite feasibility set mentioned above, it is possible to show that if one takes

$$n \;\geq\; \frac{C}{\varepsilon^2} \log\left(\frac{|\mathcal{Y}|}{\alpha}\right)$$

— where C is a constant that depends on the variances of the random variables $\Psi(y, X)$ — then the probability that $|\psi(\hat{y}_n) - \psi(y^*)|$ is less than ε is at least $1 - \alpha$ (Kleywegt et al., 2001). Note that n depends only *logarithmically* on the size of the feasible set as well as on the "quality" parameter α. Besides its practical appeal, this conclusion has implications on the complexity of solving stochastic optimization problems; we refer to Shapiro (2006) and Shapiro and Nemirovski (2005) for details.

The results above described, although very helpful, require that the estimators in (5) be constructed from i.i.d. samples. However, it is natural to consider what happens in case those estimators are constructed from samples generated by other methods such as LHS. Numerical experiments reported in the literature invariably show that convergence of $\{\hat{y}_n\}$ and $\{\hat{\nu}_n\}$ improves when LHS is used, and indeed this is intuitively what one would expect.

We illustrate this point with a very simple example. Suppose that X is discrete uniform on $\{-2, -1, 0, 1, 2\}$. The *median* of X (which in this case is evidently equal to zero) can be expressed as the solution of the problem $\min_{y\in\mathbb{R}} \mathbb{E}|X - y|$. Note that this is a stochastic optimization problem of the form (4), with $\Psi(y, X) = |X - y|$ — which is a convex piecewise linear function. As before, let $X_1, \ldots, X_n$ be i.i.d. samples of X. Clearly, the approximating solution $\hat{y}_n$ is the median of $X_1, \ldots, X_n$. Note that, when n is odd, $\hat{y}_n$ is nonzero if and only if at least half of the sampled numbers are bigger than zero (or less than zero). That is, by defining a random variable Z with binomial distribution $B(n, 2/5)$ we have

$$\mathbb{P}(Z \geq n/2) \;\leq\; \mathbb{P}\left(\hat{y}_n \neq 0\right) \;\leq\; 2\mathbb{P}(Z \geq n/2)$$

and thus

$$\lim_{n\to\infty} \frac{1}{n} \log[\mathbb{P}(\hat{y}_n \neq 0)] \;=\; -I(1/2) \;=\; -0.0204,$$

where $I(\cdot)$ is the rate function of the Bernoulli distribution with parameter $p = 2/5$ (see Section 2.1 for a precise definition), which can be expressed as $I(x) = x\log(\frac{x}{p}) + (1-x)\log(\frac{1-x}{1-p})$ for $x \in [0,1]$.

Suppose now that the samples $X_1, \ldots, X_n$ are generated using LHS. It is easy to see that there are at least $\lfloor \frac{2n}{5} \rfloor$ and at most $\lfloor \frac{2n}{5} \rfloor + 1$ numbers in $\{1,2\}$ (similarly for $\{-2,-1\}$). Thus, we have that $\hat{y}_n \neq 0$ only if

$$\lfloor \frac{2n}{5} \rfloor + 1 \;\geq\; n/2.$$

Clearly, this is impossible for $n \geq 9$, so $\mathbb{P}(\hat{y}_n \neq 0) = 0$ for n large enough. That is, the asymptotic rate of convergence of $\{\hat{y}_n\}$ to the true value is *infinite*, as opposed to the exponential rate with positive constant obtained with Monte Carlo. A natural question that arises is, does such result hold in more general settings?

An answer to the above question has been provided (at least partially) by Homem-de-Mello (2006), who shows that if a large deviations property such as (2) holds for the pointwise estimators $\hat{\psi}_n(y)$ at each $y \in \mathcal{Y}$, then the convergence result (7) holds under similar assumptions to those in the Monte Carlo case. Moreover, an improvement in the rate of convergence of pointwise estimators will lead to an improvement of the rate β in (7). This motivates the study of large deviations results under alternative sampling methods — for instance, under LHS as in the present paper.

The remainder of the paper is organized as follows. In Section 2, we give some background on large deviations theory and Latin Hypercube sampling. We also state some of the calculus results needed later in the paper. In Section 3, we show our results for functions in one-dimension. In Section 4, we extend the one-dimensional results to separable functions with multi-dimensional domain, multi-dimensional functions with bounded residual term, and multi-dimensional functions that are monotone in each argument. In Section 5 we show some examples of our results and in Section 6 we present concluding remarks.

2 Background

2.1 Large Deviations

We begin with a brief overview of some of the basic results from large deviations theory (although Proposition 1 below is a new result). For more comprehensive discussions, we refer to books such as Dembo and Zeitouni (1998) or den Hollander (2000). In our presentation, we emphasize the cases where independence assumptions are not required.

Suppose Y is a real-valued variable with mean $\mu = \mathbb{E}[Y] < \infty$ and let $S_n = \sum_{i=1}^n Y_i$, where $Y_1, \ldots, Y_n$ are (not necessarily i.i.d.) *unbiased* samples from the distribution of Y, i.e., $\mathbb{E}[Y_i] = \mu$.

Clearly, S_n/n is an unbiased estimator of μ. Define the extended real-valued function

$$\phi_n(\theta) \ := \ \frac{1}{n} \log \mathbb{E}[\exp(\theta S_n)]. \tag{8}$$

It is easy to check that $\phi_n(\cdot)$ is convex with $\phi_n(0) = 0$.

Let (a, b) be an interval on the real line containing μ. We wish to calculate the probability that the estimator S_n/n deviates from μ, i.e. $\mathbb{P}\left(\frac{1}{n}S_n \notin (a,b)\right) = \mathbb{P}\left(\frac{1}{n}S_n \leq a\right) + \mathbb{P}\left(\frac{1}{n}S_n \geq b\right)$. Standard calculations show that

$$\frac{1}{n}\log\left[\mathbb{P}\left(\frac{1}{n}S_n \geq b\right)\right] \ \leq \ \inf_{\theta\geq 0} -[\theta b - \phi_n(\theta)] \ = \ -\sup_{\theta\geq 0}\,[\theta b - \phi_n(\theta)]. \tag{9}$$

Note that this inequality holds regardless of any independence assumptions on the Y_is. Also, by Jensen's inequality we have $\mathbb{E}[\exp(\theta S_n)] \geq \exp(\theta \mathbb{E}[S_n]) = \exp(\theta n \mu)$ for any $\theta \in \mathbb{R}$ and hence

$$\phi_n(\theta) \ \geq \ \theta\mu \qquad \text{for all } \theta \in \mathbb{R}. \tag{10}$$

It follows that $\theta b - \phi_n(\theta) \leq \theta(b - \mu)$. Since $b > \mu$, we can take the supremum in (9) over $\theta \in \mathbb{R}$.

By doing a similar calculation for $\mathbb{P}(\frac{1}{n}S_n \leq a)$ we conclude that

$$\frac{1}{n}\log\left[\mathbb{P}\left(\frac{1}{n}S_n \geq b\right)\right] \leq -I(n,b) \tag{11a}$$

$$\frac{1}{n}\log\left[\mathbb{P}\left(\frac{1}{n}S_n \leq a\right)\right] \leq -I(n,a), \tag{11b}$$

where the function $I(n, z)$ is defined as

$$I(n,z) \ := \ \sup_{\theta\in\mathbb{R}}\,[\theta z - \phi_n(\theta)]. \tag{12}$$

Note that (11) holds for all $n \geq 1$. Also, $I(n, z) \geq 0$ for all n and all z. We would like, however, to establish that $I(n, z) > 0$ for $z \neq \mu$ and all n, in which case the deviation probabilities in (11) yield an exponential decay (note that, since $\theta z - \phi_n(\theta) \leq \theta(z - \mu)$ for all θ and z, it follows that $I(n, \mu) = 0$ for all n — a natural conclusion since we cannot expect to have an exponential decay for the probability $\mathbb{P}(S_n/n \geq \mu)$).

We proceed now in that direction. Suppose that the functions $\{\phi_n(\cdot)\}$ are bounded above by a common function $\phi^*(\cdot)$. Then, by defining $I^*(z) := \sup_{\theta\in\mathbb{R}}\,[\theta z - \phi^*(\theta)]$ we have that (11) holds with $I^*(a)$ and $I^*(b)$ in place of respectively $I(n, a)$ and $I(n, b)$. Since those quantities do not depend on n, it follows that $\mathbb{P}(S_n/n \leq a)$ and $\mathbb{P}(S_n/n \geq b)$ converge to zero *at least as fast as* the exponential functions $\exp(-nI^*(a))$ and $\exp(-nI^*(b))$, i.e.,

$$\mathbb{P}\left(\frac{1}{n}S_n \geq b\right) \ \leq \ \exp(-nI^*(b)), \qquad \mathbb{P}\left(\frac{1}{n}S_n \leq a\right) \ \leq \ \exp(-nI^*(a)). \tag{13}$$

The proposition below establishes further conditions on ϕ^* in order for I^* to have some desired properties. In particular, under those conditions I^* is a *rate function* (in the sense of Dembo and Zeitouni 1998) that satisfies $I^*(z) > 0$ for all $z \neq \mu$.

Proposition 1 *Consider the functions* $\{\phi_n(\cdot)\}$ *defined in* (8)*. Suppose that there exists an extended real-valued function* $\phi^*(\cdot)$ *such that* $\phi_n(\cdot) \leq \phi^*(\cdot)$ *for all* n*, with* ϕ^* *satisfying the following properties:* (i) $\phi^*(0) = 0$*;* (ii) $\phi^*(\cdot)$ *is continuously differentiable and strictly convex on a neighborhood of zero; and* (iii) $(\phi^*)'(0) = \mu$*.*

Then, the function $I^*(z) := \sup_{\theta\in\mathbb{R}} [\theta z - \phi^*(\theta)]$ *is lower semi-continuous, convex everywhere and strictly convex on a neighborhood of* μ*. Moreover,* $I^*(\cdot)$ *is non-negative and* $I^*(\mu) = 0$*.*

Proof. From (10), we have $\phi_n(\theta) \geq \theta\mu$ for all $\theta \in \mathbb{R}$ and hence $\phi^*(\theta) \geq \theta\mu$ for all $\theta \in \mathbb{R}$. It follows from Theorem X.1.1.2 in Hiriart-Urruty and Lemarechal (1993) that I^* (the conjugate function of ϕ^*) is convex and lower semi-continuous.

Next, condition (ii) implies that $(\phi^*)'(\cdot)$ is continuous and strictly increasing on a neighborhood of zero. Since $(\phi^*)'(0) = \mu$ by condition (iii), there exists some $\varepsilon > 0$ such that, given any $z_0 \in [\mu - \varepsilon, \mu + \varepsilon]$, there exists θ_0 satisfying $(\phi^*)'(\theta_0) = z_0$. It follows from Theorem X.4.1.3 in Hiriart-Urruty and Lemarechal (1993) that I^* is strictly convex on $[\mu - \varepsilon, \mu + \varepsilon]$.

Non-negativity of $I^*(\cdot)$ follows immediately from $\phi^*(0) = 0$. Finally, since $\theta\mu - \phi^*(\theta) \leq 0$ for all $\theta \in \mathbb{R}$, it follows from the definition of I^* that $I^*(\mu) = 0$. □

A simple setting where the conditions of Proposition 1 are satisfied is when the functions ϕ_n are bounded by the log-moment generating function of some random variable W (i.e., $\phi^*(\theta) = \log \mathbb{E}[\exp(\theta W)]$) such that $\mathbb{E}[W] = \mu$. Clearly, condition (i) holds in that case. Moreover, if there exists a neighborhood $\mathcal{N}$ of zero such that $\phi^*(\cdot)$ is finite on $\mathcal{N}$, then it is well known that ϕ^* is infinitely differentiable on $\mathcal{N}$ and (iii) holds. In that case, Proposition 1 in Shapiro et al. (2002) ensures that ϕ^* is strictly convex on $\mathcal{N}$.

The above developments are valid regardless of any i.i.d. assumption on the samples $\{Y_i\}$. When such an assumption is imposed, we have

$$\phi_n(\theta) \;=\; \frac{1}{n}\log(\mathbb{E}[\exp(\theta S_n)]) \;=\frac{1}{n}\log(\{\mathbb{E}[\exp(\theta Y_1)]\}^n) \;=\log(\mathbb{E}[\exp(\theta Y_1)]) \;=\log M_{Y_1}(\theta), \quad (14)$$

where $M_{Y_1}(\theta)$ is the moment generating function of Y_1 evaluated at θ. In that case, of course, we have $\phi_n(\theta) = \phi^*(\theta) := \log M_{Y_1}(\theta)$ for all n, and the resulting function I^* is the rate function associated with Y_1. The inequalities in (13) then yield the so-called Chernoff upper bounds on the deviation probabilities.

Inequalities such as (13), while useful in their own, do not fully characterize the deviation probabilities since they only provide an upper bound on the decay. One of the main contributions of large deviations theory is the verification that, in many cases, the decay rate given by those inequalities is *asymptotically exact*, in the sense that (11) holds with equality as n goes to infinity. One such case is when $\{Y_i\}$ is i.i.d.; that, of course, is the conclusion of the well-known Cramér's Theorem.

In general, the idea of an asymptotically exact decay rate is formalized as follows. The estimator $\frac{1}{n}S_n$ — calculated from possibly non-i.i.d. random variables — is said to satisfy a *large deviation principle (LDP)* with rate function $I(\cdot)$ if the following conditions hold:

1. $I(\cdot)$ is lower semi-continuous, i.e., it has closed level sets;
2. For every closed subset $F \subseteq \mathbb{R}$,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}S_n \in F\right) \leq -\inf_{x\in F} I(x) \tag{15}$$

3. For every open subset $G \subseteq \mathbb{R}$,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}S_n \in G\right) \geq -\inf_{x\in G} I(x). \tag{16}$$

$I(\cdot)$ is said to be a *good rate function* if it has compact level sets. Note that this implies that there exists some point x such that $I(x) = 0$.

As mentioned above, in the i.i.d. case an LDP holds with I^* — the rate function of Y_1 — in place of I. The main tool for the general case is the *Gärtner-Ellis Theorem*, which we describe next. Let ϕ_n be as defined in (8), and define

$$\phi(\theta) \ := \ \lim_{n\to\infty} \phi_n(\theta) \quad \text{ when the limit exists.} \tag{17}$$

Roughly speaking, the theorem asserts that, under proper conditions, a large deviation principle holds for the estimator $\frac{1}{n}S_n$, with the rate defined in terms of the limiting $\phi(\theta)$ defined in (17). For a complete statement we refer to Dembo and Zeitouni (1998) or den Hollander (2000).

Our main goal is to derive conditions under which the above results can be applied under Latin Hypercube sampling. In Sections 3 and 4 we will show that, under those conditions, the upper bound (13) holds with the same function I^* as the standard Monte Carlo (which suggests that LHS can do no worse than i.i.d. sampling). In some cases, we will be able to apply the Gärtner-Ellis Theorem to show that a large deviation principle holds. Before stating those results, we review in detail the basic ideas of LHS.

2.2 Latin Hypercube Sampling

Let $X = [X^1, X^2, \ldots, X^d]$ be a vector of d independent input variables of a simulation and let $Y = g(X) = g(X^1, X^2, \ldots, X^d)$ be the output of the simulation. Let $F_j(\cdot)$ be the marginal cumulative distribution function for X^j. Suppose the quantity of interest is $\mathbb{E}[Y]$.

One possible sampling method to estimate $\mathbb{E}[Y]$ is to randomly sample n points in the sample space (Monte Carlo sampling). For each replication i from 1 to n, Uniform(0,1) random numbers $U_i = [U_i^1, \ldots, U_i^d]$ are generated (one per dimension) which, assuming we can use the inverse transform method, yield the input random vector $X_i = [F_1^{-1}(U_i^1), \ldots, F_d^{-1}(U_i^d)]$ and the output $Y_i = g(X_i)$.

One problem with Monte Carlo sampling is that there is no guarantee that all sections of the sample space will be equally represented; input points could be clustered in one particular region. This is, of course, a well-known issue, and a considerable body of literature — notably on *quasi-Monte Carlo* methods — exists dealing with that topic. Latin Hypercube sampling, first proposed by McKay et al. (1979), falls into that category. The method splits each dimension of the sample space into n sections (or strata) each with probability $\frac{1}{n}$, and samples one observation from each stratum. The algorithm is comprised of three steps — it generates some uniform random numbers, then some random permutations and finally these elements are put together to yield the samples. We present below a detailed algorithm. Although the algorithm is usually described in a more succinct form in the literature, our formulation introduces some notation that will be used later on.

1. *Generate uniform random numbers:* Generate a $n \times d$ matrix V of independent random numbers such that each $(i,j)^{th}$ entry V_i^j is uniformly distributed on the interval $[\frac{i-1}{n}, \frac{i}{n}]$.

2. *Generate random permutations:*

   (a) Let $\mathcal{P}$ be the set of $n \times d$ matrices where each column is a permutation of the numbers $1, 2, \ldots, n$. There are $(n!)^d$ elements in $\mathcal{P}$.

   (b) Let K be uniformly distributed on $\{1, 2, \ldots, (n!)^d\}$, and let $\Pi(K)$ denote the Kth element of $\mathcal{P}$. Let $\pi_i^j(K)$ be the $(i,j)^{th}$ entry of this matrix. Note that the permutation matrix $\Pi(K)$ is independent of the random number matrix V.

   (c) In Latin Hypercube sampling, only n of the n^d strata are sampled. The rows of the $\Pi(K)$ matrix determine which hypercubes get sampled. Let $\pi_i(K) = [\pi_i^1(K), \ldots, \pi_i^d(K)]$ be the i^{th} row of $\Pi(K)$. This corresponds to the hypercube that covers the $\pi_i^1(K)^{th}$ stratum of X^1, the $\pi_i^2(K)^{th}$ stratum of X^2, ..., and the $\pi_i^d(K)^{th}$ stratum of X^d.

3. *Determine the randomly sampled point within each hypercube.*

   (a) Create matrix $Z = Z(V, K)$ with $(i,j)^{th}$ entry $Z_i^j = V_{\pi_i^j(K)}^j$, i.e., the j^{th} column V^j of the random number matrix V is permuted according to the j^{th} column of the permutation matrix $\Pi(K)$.

   (b) Let $X_i^j = F_j^{-1}[Z_i^j]$. Then $X_i = [X_i^1, \ldots, X_i^d]$ and $Y_i = g(X_i)$.

The above algorithm generates n random vectors $Z_i = [Z_i^1, \ldots, Z_i^d]$, each of which is uniformly distributed on $[0,1]^d$. Unlike standard Monte Carlo, of course, the vectors $Z_1, \ldots, Z_n$ are *not* independent. These vectors are mapped via inverse transform into vectors $X_1, \ldots, X_n$, which then are used to generate the samples $Y_1, \dots, Y_n$. It is well known that each Y_i generated by the LHS method is an unbiased estimate of $\mathbb{E}[Y]$ (see, e.g., the appendix in McKay et al. 1979). Since the application if inverse transform is quite standard, without loss of generality we will assume that the outputs Y_i are functions of random vectors that are uniformly distributed on $[0,1]^d$.

2.3 Calculus Results

For the remaining sections of this paper, we will need to define some notation and recall some results from analysis. These results are known but we state them for later reference. The discussion below follows mostly Bartle (1987) and Royden (1988).

Let $P := \{z_0, z_1, \dots, z_n\}$ be a partition of the interval $[a,b]$ with $a = z_0 < z_1 < \cdots < z_{n-1} < z_n = b$ and let $|P|$ denote the norm of that partition (the maximum distance between any two consecutive points of the partition). Let $h : [a,b] \mapsto \mathbb{R}$ be a *bounded* function. A *Riemann sum* of h corresponding to P is a real number of the form $R(P,h) = \sum_{i=1}^n h(\xi_i)(z_i - z_{i-1})$, where $\xi_i \in [z_{i-1}, z_i]$, $i = 1, \dots, n$. Riemann integrability ensures that, given $\varepsilon > 0$, there exists $\eta > 0$ such that $|R(P,h) - \int_a^b h(z)dz| < \varepsilon$ for any Riemann sum $R(P,h)$ such that $|P| < \eta$. At this point it is worthwhile recalling that a bounded function h is Riemann integrable if and only if the set of points at which h is discontinuous has Lebesgue measure zero (Royden, 1988, p. 85).

In our setting we will often deal with functions that are *not* bounded. In that case, we say that $h : [a,b] \mapsto \mathbb{R}$ is *integrable* if h is Lebesgue integrable. Of course, when h is bounded and Riemann integrable both the Lebesgue and the Riemann integrals coincide.

The next lemma is a trivial application of dominated convergence, but we state it for later reference. Below, I_A denotes the indicator function of a set A.

Lemma 1 *Suppose $h : [a,b] \mapsto \mathbb{R}$ is integrable, and let $\{A_n\}$ be a sequence of subsets of $[a,b]$ such that $\mathrm{I}_{A_n} \to 0$ almost surely. Then, $\int_{A_n} h(z)dz \to 0$.*

3 The One-Dimensional Case

We study now large deviations properties of the estimators generated by LHS. In order to facilitate the analysis, we start by considering the one-dimensional case.

Let $h : [0,1] \mapsto \mathbb{R}$ be a real-valued function in one variable, and suppose we want to estimate $\mathbb{E}[h(Z)]$, where Z has Uniform(0,1) distribution. In standard Monte Carlo sampling, the samples Z_i are all independent Uniform(0,1) random variables. In that case, we have from (14) that

$$\phi^{MC}(\theta) \;:=\; \phi_n^{MC}(\theta) \;=\; \log(\mathbb{E}[\exp(\theta h(Z_1))]) \;=\; \log\left[\int_0^1 \exp(\theta h(z))dz\right],$$

which is independent of n.

In LHS, when the interval $[0,1]$ is split into n strata of equal probability $\frac{1}{n}$, the intervals are all of the form $[\frac{j-1}{n}, \frac{j}{n}]$ and each random variable Z_i is now uniform on some interval of length $\frac{1}{n}$. Further, independence no longer holds.

We make the following assumptions about the function $h(z) : [0,1] \mapsto \mathbb{R}$:

Assumption 1

(a) *$h(\cdot)$ has at most a finite number of singularities.*

(b) *$h(\cdot)$ has a finite moment generating function (i.e. $\int_0^1 \exp(\theta h(z))dz < \infty$ for all $\theta \in \mathbb{R}$).*

(c) *The set of points at which $h(\cdot)$ is discontinuous has Lebesgue measure zero.*

A simple situation where the above assumptions are satisfied is when h is a bounded measurable function with at most countably many discontinuities; however, we do allow h to be unbounded.

To show that LHS satisfies a large deviation principle, we will apply the Gärtner-Ellis Theorem. In what follows, $Z_1, \ldots, Z_n$ are samples from a Uniform(0,1) distribution, generated by the Latin Hypercube sampling algorithm, and ϕ_n^{LHS} is defined as in (8), with $Y_i = h(Z_i)$.

Our main result in this section is Theorem 1 below. Before stating that result, we introduce some auxiliary results.

Lemma 2 *Suppose Assumption 1 holds. Then,*

$$\phi_n^{LHS}(\theta) \;=\; \theta \frac{1}{n} \sum_{i=1}^{n} c_i(n,\theta), \tag{18}$$

where, for $\theta \neq 0$, $c_i(n,\theta)$ is defined as

$$c_i(n,\theta) \;:=\; \frac{1}{\theta} \log \left(n \int_{\frac{i-1}{n}}^{\frac{i}{n}} \exp(\theta h(z)) dz \right) \tag{19}$$

(and $c_i(n,0) := 0$).

Proof. Fix $\theta \neq 0$ (for $\theta = 0$ the result is trivial). Following the notation defined in the LHS algorithm described above, let us denote the LH samples by $Z_1(V,K), \ldots, Z_n(V,K)$. Let $\mathbb{E}_V[\cdot]$ denote the expectation with respect to the random number matrix V, and $\mathbb{E}[\cdot]$ with no subscripts denote the expectation with respect to both V and K. We have

$$\begin{aligned}\exp(n\phi_n^{LHS}(\theta)) \;&=\; \mathbb{E}\left[\exp(\theta S_n)\right] \;=\; \mathbb{E}\left[\exp\left(\theta \sum_{i=1}^{n} h(Z_i(V,K))\right)\right] \\ &=\; \sum_{k=1}^{n!} \mathbb{E}_V \left[\exp\left(\theta \sum_{i=1}^{n} h(Z_i(V,K))\right) \middle| K = k \right] \mathbb{P}(K = k).\end{aligned}$$

Since each permutation is equally likely, $\mathbb{P}(K = k) = \frac{1}{n!}$ for all permutations k. For each permutation k, one of the $Z_i(V, k)$ is uniform on stratum $[0, \frac{1}{n}]$, another is uniform on stratum $[\frac{1}{n}, \frac{2}{n}]$, etc., and every interval $[\frac{i-1}{n}, \frac{i}{n}]$ is sampled exactly once. It is easy to see then that the random variables $Z_1(V, K), \ldots, Z_n(V, K)$ are exchangeable and thus the conditional distribution of $\sum_{i=1}^n h(Z_i(V, K))$ on $K = k$ is the same for all k. Further, once the permutation has been fixed, the samples in each stratum are independent. It follows that

$$\begin{aligned}
\exp(n\phi_n^{LHS}(\theta)) &= \frac{1}{n!}\sum_{k=1}^{n!}\mathbb{E}_V\left[\exp\left(\theta\sum_{i=1}^n h(Z_i(V,K))\right)\middle| K=k\right] \\
&= \mathbb{E}_V\left[\prod_{i=1}^n \exp(\theta h(Z_i(V,K)))\middle| K=1\right] \\
&= \prod_{i=1}^n \mathbb{E}_V\left[\exp(\theta h(Z_i(V,K)))\middle| K=1\right] \\
&= \prod_{i=1}^n \int_{\frac{i-1}{n}}^{\frac{i}{n}} \exp(\theta h(z))\, n\, dz. \qquad (20)
\end{aligned}$$

To get the latter equation, we have assumed (without loss of generality) that the permutation $k = 1$ is $(1, 2, \ldots, n)$. Also, we have used the fact that the density function of a Uniform$(\frac{i-1}{n}, \frac{i}{n})$ random variable Z_i is $n\, dz$.

By the finiteness of the moment generating function, $\exp(\theta h(\cdot))$ has some finite average value on the interval $[\frac{i-1}{n}, \frac{i}{n}]$, which is given by $1/(1/n)\int_{(i-1)/n}^{i/n}\exp(\theta h(z))dz$. This average value is equal to $\exp(\theta c_i(n, \theta))$, where $c_i(n, \theta)$ is defined in (19). Substituting the above expression back into (20) we obtain

$$\exp(n\phi_n^{LHS}(\theta)) = \prod_{i=1}^n n\int_{\frac{i-1}{n}}^{\frac{i}{n}}\exp(\theta h(z))dz = \prod_{i=1}^n \exp(\theta c_i(n,\theta)) = \exp\left(\theta\sum_{i=1}^n c_i(n,\theta)\right) \qquad (21)$$

and hence

$$\phi_n^{LHS}(\theta) = \theta\frac{1}{n}\sum_{i=1}^n c_i(n,\theta). \qquad \square$$

Lemma 3 *For any $\theta \in \mathbb{R}$, the quantities $c_i(n, \theta)$ defined in Lemma 2 satisfy*

$$\theta\int_{\frac{i-1}{n}}^{\frac{i}{n}} h(z)dz \le \theta\frac{1}{n}c_i(n,\theta) \le \int_{\frac{i-1}{n}}^{\frac{i}{n}}\exp(\theta h(z))dz. \qquad (22)$$

Proof. Using Jensen's inequality, we have, for each $i = 1, \ldots, n$,

$$\theta\frac{1}{n}c_i(n,\theta) = \frac{1}{n}\log\left(\int_{\frac{i-1}{n}}^{\frac{i}{n}}\exp(\theta h(z))\, n\, dz\right) \ge \frac{1}{n}\int_{\frac{i-1}{n}}^{\frac{i}{n}}\log\left(\exp(\theta h(z))\right) n\, dz = \theta\int_{\frac{i-1}{n}}^{\frac{i}{n}} h(z)dz.$$

On the other hand, since $x \leq \exp(x)$ for any x, we have

$$\frac{1}{n}\theta c_i(n,\theta) \;\leq\; \frac{1}{n}\exp(\theta c_i(n,\theta)) \;=\; \frac{1}{n}\int_{\frac{i-1}{n}}^{\frac{i}{n}} n\exp(\theta h(z))dz \;=\; \int_{\frac{i-1}{n}}^{\frac{i}{n}} \exp(\theta h(z))dz. \qquad \square$$

The proposition below provides a key result. It shows that $\{\phi_n(\cdot)\}$ converges to a *linear* function in θ.

Proposition 2 *Suppose Assumption 1 holds. Then, for all $\theta \in \mathbb{R}$ we have*

$$\lim_{n\to\infty} \phi_n^{LHS}(\theta) \;=\; \theta\int_0^1 h(z)dz. \tag{23}$$

Proof. Fix $\theta \neq 0$ (for $\theta = 0$ the proof is trivial). Our goal is to show that the limit (as $n \to \infty$) of the expression on the right-hand side of (18) exists and is equal to $\theta\int_0^1 h(z)dz$. Although we do not assume that either $h(\cdot)$ or $\exp(\theta h(\cdot))$ is bounded, by assuming a finite number of singularities (cf. Assumption 1a) we can decompose the function into regions that are bounded plus neighborhoods around the singularity points.

Without loss of generality, let us assume that the function $h(\cdot)$ has just one singularity at $z = s$ with $s \in (0,1)$. The case with more than one singularity — but finitely many ones — is a straightforward extension of the one-singularity case. Also, if $s = 0$ the argument presented in the next paragraphs can be readily adapted by splitting the domain into two pieces, namely $[0,\delta)$ and $[\delta,1]$ (similarly, if $s = 1$ we split the domain into $[0, 1-\delta]$ and $(1-\delta, 1]$).

Fix an arbitrary $\varepsilon > 0$. From Lemma 1, we can take $\delta > 0$ so that $|\int_{s-\delta}^{s+\delta} h(z)dz| \leq \varepsilon$ and $|\int_{s-\delta}^{s+\delta}\exp(\theta h(z))dz| \leq \varepsilon|\theta|$. We can then split the domain into three pieces: $[0, s-\delta], (s-\delta, s+\delta)$, and $[s+\delta, 1]$. Let $P(n)$ be the partition of the interval $[0,1]$ into equal subintervals of length $\frac{1}{n}$. $|P(n)|$ is just $\frac{1}{n}$. Denote by $\tilde{P}(n) = \{0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{\ell_1}{n}, s-\delta\}\bigcup\{s+\delta, \frac{\ell_2}{n}, \frac{\ell_2+1}{n}, \ldots, 1\}$ the corresponding partition of $[0, s-\delta]\bigcup[s+\delta, 1]$ with $\ell_1 = \max\{\ell \in \mathbb{N} : \frac{\ell}{n} < s-\delta\}$ and $\ell_2 = \min\{\ell \in \mathbb{N} : \frac{\ell}{n} > s+\delta\}$. Note that both ℓ_1 and ℓ_2 depend on n, but we omit the dependence to ease the notation. The partition $\tilde{P}(n)$ also has norm $\frac{1}{n}$.

The function h is bounded on both $[0, s-\delta]$ and $[s+\delta, 1]$. It follows that the quantities

$$m_i(n) \;:=\; \inf_{z\in[\frac{i-1}{n},\frac{i}{n}]} h(z)$$

$$M_i(n) \;:=\; \sup_{z\in[\frac{i-1}{n},\frac{i}{n}]} h(z)$$

are finite for all i such that $\left[\frac{i-1}{n}, \frac{i}{n}\right] \subseteq [0, s-\delta]$ (or $\subseteq [s+\delta, 1]$). Moreover, it is easy to check that $m_i(n) \leq c_i(n,\theta) \leq M_i(n)$, where $c_i(n,\theta)$ is defined in (19).

Next, define

$$\tilde{c}_{\ell_1+1}(n,\theta) \;:=\; \frac{1}{\theta}\log\left(\frac{1}{(s-\delta)-\frac{\ell_1}{n}}\int_{\frac{\ell_1}{n}}^{s-\delta}\exp(\theta h(z))dz\right)$$

and

$$\tilde{c}_{\ell_2}(n,\theta) \;:=\; \frac{1}{\theta}\log\left(\frac{1}{\frac{\ell_2}{n}-(s+\delta)}\int_{s+\delta}^{\frac{\ell_2}{n}}\exp(\theta h(z))dz\right),$$

so that $\exp(\theta\tilde{c}_{\ell_1+1}(n,\theta))$ and $\exp(\theta\tilde{c}_{\ell_2}(n,\theta))$ are equal to the average value of $\exp(\theta h(\cdot))$ on the intervals $[\frac{\ell_1}{n}, s-\delta]$ and $[s+\delta, \frac{\ell_2}{n}]$ respectively. Define now the sum

$$R(n) \;:=\; \frac{1}{n}\left[\sum_{i=1}^{\ell_1} c_i(n,\theta) + \sum_{i=\ell_2+1}^{n} c_i(n,\theta)\right] + \left((s-\delta)-\frac{\ell_1}{n}\right)\tilde{c}_{\ell_1+1}(n,\theta) + \left(\frac{\ell_2}{n}-(s+\delta)\right)\tilde{c}_{\ell_2}(n,\theta).$$

Assumption 1c, together with boundedness of h on $[0, s-\delta]\bigcup[s+\delta, 1]$, implies Riemann integrability of h on that region. The construction of the $c_i(n,\theta)$ and $\tilde{c}_j(n,\theta)$ terms and the properties derived above then ensure we can find some $n > 0$ such that

$$\left|R(n) - \left(\int_0^{s-\delta} h(z)dz + \int_{s+\delta}^{1} h(z)dz\right)\right| \;\leq\; \varepsilon. \tag{24}$$

Thus,

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n} c_i(n,\theta) - \int_0^1 h(z)dz\right| \\
&\quad= \Bigg|R(n) - \left[\int_0^{s-\delta} h(z)dz + \int_{s+\delta}^{1} h(z)dz\right] + \left[\frac{1}{n}c_{\ell_1+1}(n,\theta) - \left((s-\delta)-\frac{\ell_1}{n}\right)\tilde{c}_{\ell_1+1}(n,\theta)\right] \\
&\qquad + \left[\frac{1}{n}c_{\ell_2}(n,\theta) - \left(\frac{\ell_2}{n}-(s+\delta)\right)\tilde{c}_{\ell_2}(n,\theta)\right] + \left[\frac{1}{n}\sum_{i=\ell_1+2}^{\ell_2-1} c_i(n,\theta) - \int_{s-\delta}^{s+\delta} h(z)dz\right]\Bigg| \\
&\quad\leq \left|R(n) - \left(\int_0^{s-\delta} h(z)dz + \int_{s+\delta}^{1} h(z)dz\right)\right| + \left|\frac{1}{n}c_{\ell_1+1}(n,\theta)\right| + \left|\left((s-\delta)-\frac{\ell_1}{n}\right)\tilde{c}_{\ell_1+1}(n,\theta)\right| \\
&\qquad + \left|\frac{1}{n}c_{\ell_2}(n,\theta)\right| + \left|\left(\frac{\ell_2}{n}-(s+\delta)\right)\tilde{c}_{\ell_2}(n,\theta))\right| + \left|\frac{1}{n}\sum_{i=\ell_1+2}^{\ell_2-1} c_i(n,\theta)\right| + \left|\int_{s-\delta}^{s+\delta} h(z)dz\right|
\end{aligned}$$

The first term on the right-hand side of the above inequality is less than ε for n large enough (cf. (24)). Moreover, Lemma 3, together with Assumption 1 and Lemma 1 show that, for each i, we have that $|\frac{1}{n}c_i(n,\theta)| \leq \varepsilon$ for n large enough. A similar argument holds for the $\tilde{c}_j(n,\theta)$ terms. Lemma 3, Assumption 1 and Lemma 1 (together with the choice of δ) also imply that

$$\frac{1}{n}\sum_{i=\ell_1+2}^{\ell_2-1} |c_i(n,\theta)| \;\leq\; \max\left(\frac{1}{|\theta|}\int_{\frac{\ell_1+1}{n}}^{\frac{\ell_2-1}{n}} \exp(\theta h(z))\,dz,\; \int_{\frac{\ell_1+1}{n}}^{\frac{\ell_2-1}{n}} |h(z)|\,dz\right) \;\leq\; \varepsilon, \tag{25}$$

since $\left[\frac{\ell_1+1}{n}, \frac{\ell_2-1}{n}\right] \subseteq [s-\delta, s+\delta]$ for any n.

It follows from the above developments that for n large enough we have

$$\left|\frac{1}{n}\sum_{i=1}^{n} c_i(n,\theta) - \int_0^1 h(z)dz\right| \;\leq\; 7\varepsilon.$$

Since ε was chosen arbitrarily, it follows that $\theta \frac{1}{n}\sum_{i=1}^n c_i(n,\theta) \to \theta \int_0^1 h(z)dz$ as we wanted to show. $\square$

The main result of this section is the following:

Theorem 1 *Let $h : [0,1] \mapsto \mathbb{R}$ and suppose that Assumption 1 holds. Let Z be a Uniform(0,1) random variable and define $\mu_1 := \mathbb{E}[h(Z)] = \int_0^1 h(z)dz$. Then, the LHS estimator of μ_1 satisfies a large deviation principle with good rate function*

$$I^{LHS}(x) = \begin{cases} \infty, & \text{if } x \neq \mu_1 \\ 0, & \text{if } x = \mu_1. \end{cases}$$

Proof. Let $\phi(\theta) := \theta \int_0^1 h(z)dz = \theta\mu_1$. Proposition 2 ensures that $\phi_n^{LHS}(\theta) \to \phi(\theta)$ for all θ. Since the limiting function $\phi(\cdot)$ is linear, the assumptions of the Gärtner-Ellis Theorem hold, so that result can be applied. The resulting rate function is

$$I^{LHS}(x) \;=\; \sup_\theta\,[\theta x - \phi(\theta)] \;=\; \sup_\theta\,[\theta(x-\mu_1)] \;=\; \begin{cases} \infty, & \text{if } x \neq \mu_1 \\ 0, & \text{if } x = \mu_1 \end{cases}$$

which is a good rate function since $\{x : I^{LHS}(x) \leq \alpha\} = \{\mu_1\}$ for any $\alpha \geq 0$. $\square$

Theorem 1 implies that, for any closed subset F of $\mathbb{R}$, as long as $\mu_1 \notin F$ we have that

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}S_n \in F\right) \leq -\inf_{x\in F} I^{LHS}(x) = -\infty.$$

That is, we have a *superexponential decay rate*, as opposed to the exponential rate obtained with standard Monte Carlo. This shows that, asymptotically, LHS is much more precise than Monte Carlo. It is interesting to note that Theorem 1 explains the median example discussed in Section 1 — indeed, the results in Homem-de-Mello (2006) ensure that the superexponential rate obtained with LHS in one dimension carries over to the optimization setting (i.e. $\mathbb{P}(\hat{y}_n \neq y^*)$), which in that case corresponds to the probability that the sample median is different from the true median.

The next result suggests that superiority of LHS (in the context of deviation probabilities) in fact holds for any finite n.

Proposition 3 *Consider the setting of Theorem 1. Let $I^{MC}(x)$ and $I^{LHS}(n,x)$ denote the (non-asymptotic) functions defined in* (12) *respectively for Monte Carlo and for LHS. Then, for* any *sample size n and all x we have that $I^{LHS}(n,x) \geq I^{MC}(x)$.*

Proof. For LHS we have, from Lemma 2,

$$\begin{aligned}\phi_n^{LHS}(\theta) \;&=\; \theta\frac{1}{n}\sum_{i=1}^n c_i(n,\theta) \;=\; \frac{1}{n}\sum_{i=1}^n \log\left(n\int_{\frac{i-1}{n}}^{\frac{i}{n}} \exp(\theta h(z))\,dz\right) \\ &\leq\; \log\left[\frac{1}{n}\sum_{i=1}^n n\int_{\frac{i-1}{n}}^{\frac{i}{n}} \exp(\theta h(z))dz\right] \;=\; \log\left[\int_0^1 \exp(\theta h(z))dz\right] \;=\; \phi^{MC}(\theta),\end{aligned}$$

where the inequality follows from Jensen's inequality. Thus, $\phi_n^{LHS}(\theta) \le \phi^{MC}(\theta)$ for all n and θ. Equivalently, $I^{LHS}(n,x) \ge I^{MC}(x)$ for all x. $\square$

Note that the above result fits the framework of the discussion around (13), i.e., the upper bound for the probability of a large deviation is smaller under Latin Hypercube sampling than under Monte Carlo sampling for any sample size n. Although in general a comparison of upper bounds is not particularly useful, the importance of Proposition 3 lies in the fact that the Monte Carlo upper bound is *tight* asymptotically. This suggests that even for small sample sizes the deviation probabilities under LHS may be smaller than under Monte Carlo — a fact that is corroborated in the examples of Section 5.

4 The Multi-Dimensional Case

We consider now the multi-dimensional case $h : [0,1]^d \mapsto \mathbb{R}$. That is, we want to estimate $\mathbb{E}[h(Z)]$, where $Z = (Z^1, \ldots, Z^d)$ is uniformly distributed on $[0,1]^d$. As before, let $Z_1, \ldots, Z_n$ denote samples from the vector Z, so that $Z_i = [Z_i^1, \ldots, Z_i^d]$.

For Monte Carlo sampling, a large deviation principle holds, and from (14) we have that

$$\phi^{MC}(\theta) = \log\left[\int_{[0,1]^d} \exp(\theta h(z))dz\right]. \tag{26}$$

Again, we would like to show that a large deviation principle holds for Latin Hypercube sampling in the multi-dimensional case and that the upper bound for the probability of a large deviation under LHS is lower than it is for Monte Carlo sampling. While these assertions may not be true in general for multidimensional functions, we will focus on three special cases: (1) $h(\cdot)$ is a *separable* function, (2) $h(\cdot)$ has a bounded residual term in its ANOVA decomposition, and (3) $h(\cdot)$ is a multi-dimensional function which is monotone in each component.

In the multi-dimensional case, each Latin Hypercube permutation is equally likely with probability $\mathbb{P}(K=k) = \frac{1}{(n!)^d}$ (recall that the permutation matrices are indexed by k, and that K is a random index). As in the one-dimensional case, given a particular permutation $\Pi(k)$, the point sampled from each strata is independent of the point sampled from any other strata, so the product and the expectation can be switched. Thus, we can write

$$\begin{aligned}
\exp(n\phi_n^{LHS}(\theta)) &= \mathbb{E}\left[\prod_{i=1}^n \exp(\theta h(Z_i(V,K)))\right] \\
&= \sum_{k=1}^{(n!)^d} \mathbb{E}\left[\prod_{i=1}^n \exp(\theta h(Z_i(V,K))) \middle| K=k\right] \mathbb{P}(K=k) \\
&= \frac{1}{(n!)^d} \sum_{k=1}^{(n!)^d} \prod_{i=1}^n \mathbb{E}\left[\exp(\theta h(Z_i(V,K)))\middle| K=k\right].
\end{aligned} \tag{27}$$

Also, given a particular permutation index k, for each sample i we have that

$$Z_i(V,k) \in \left[\frac{\pi_i^1(k)-1}{n}, \frac{\pi_i^1(k)}{n}\right] \times \cdots \times \left[\frac{\pi_i^d(k)-1}{n}, \frac{\pi_i^d(k)}{n}\right],$$

where the $\pi_i^j(k)$ indicate which strata are sampled, as defined in Section 2.2.

For notational convenience, define $a_i^j(k) := \frac{\pi_i^j(k)-1}{n}$ and $b_i^j(k) := \frac{\pi_i^j(k)}{n}$. Also, let $z := (z^1, \ldots, z^d)$ and $dz := dz^1 \cdots dz^d$. Note that $Z_i^j(V,k)$ is uniformly distributed on the interval $(a_i^j(k), b_i^j(k))$. Then, (27) becomes

$$\exp(n\phi_n^{LHS}(\theta)) \;=\; \frac{1}{(n!)^d} \sum_{k=1}^{(n!)^d} \prod_{i=1}^{n} n^d \int_{a_i^d(k)}^{b_i^d(k)} \cdots \int_{a_i^1(k)}^{b_i^1(k)} \exp(\theta h(z))dz. \tag{28}$$

We now specialize the calculations for the three cases mentioned above.

4.1 Case 1: The Separable Function Case

We shall consider initially the case where the function h is *separable*, i.e., there exist one-dimensional functions $h^1, \ldots, h^d$ such that $h(z^1, \ldots, z^d) = h^1(z^1) + \ldots + h^d(z^d)$ almost surely. Note that this is equivalent to saying that the ANOVA decomposition of h (cf. (3)) has residual part equal to zero. Clearly, when h is separable we have

$$\int_{[0,1]^d} h(z)dz = \int_0^1 h^1(z^1)dz^1 + \cdots + \int_0^1 h^d(z^d)dz^d.$$

Since a separable multidimensional function can be decomposed into a sum of one dimensional functions, it is intuitive that our results from the one-dimensional case can be extended to this case. The theorem below states precisely that:

Theorem 2 *Suppose that $h : [0,1]^d \mapsto \mathbb{R}$ is a separable function and that each component h^j of h satisfies Assumption 1. Let Z be a random vector uniformly distributed on $[0,1]^d$, and define $\mu_d := \mathbb{E}[h(Z)] = \int_{[0,1]^d} h(z)dz$. Then, the LHS estimator of μ_d satisfies a large deviation principle with good rate function*

$$I^{LHS}(x) = \begin{cases} \infty, & \text{if } x \neq \mu_d \\ 0, & \text{if } x = \mu_d. \end{cases}$$

Proof. From the proof of Theorem 1, it suffices to show that the functions $\{\phi_n^{LHS}(\theta)\}$ converge to the linear function $\phi^{LHS}(\theta) := \theta\mu_d$.

Using the special property of Latin Hypercube Sampling that in each dimension, each stratum is sampled from exactly once, we get:

$$\begin{aligned}
\exp(n\phi_n^{LHS}(\theta)) &= \frac{1}{(n!)^d}\sum_{k=1}^{(n!)^d}\prod_{i=1}^{n} n^d \int_{a_i^d(k)}^{b_i^d(k)} \cdots \int_{a_i^1(k)}^{b_i^1(k)} \exp(\theta h(z))\, dz \\
&= \frac{1}{(n!)^d}\sum_{k=1}^{(n!)^d}\prod_{j=1}^{d}\prod_{i=1}^{n} n \int_{a_i^j(k)}^{b_i^j(k)} \exp(\theta h^j(z^j))\, dz^j \\
&= \frac{1}{(n!)^d}\sum_{k=1}^{(n!)^d}\prod_{j=1}^{d}\prod_{i=1}^{n} n \int_{\frac{i-1}{n}}^{\frac{i}{n}} \exp(\theta h^j(z^j))\, dz^j \\
&= \prod_{j=1}^{d}\prod_{i=1}^{n} n \int_{\frac{i-1}{n}}^{\frac{i}{n}} \exp(\theta h^j(z^j))\, dz^j \\
&= \exp\left(\theta \sum_{i=1}^{n} c_i^1(n,\theta)\right) \cdots \exp\left(\theta \sum_{i=1}^{n} c_i^d(n,\theta)\right) \\
&= \exp\left(\theta \sum_{j=1}^{d}\sum_{i=1}^{n} c_i^j(n,\theta)\right),
\end{aligned}$$

where the $c_i^j(n,\theta)$ are defined as in (19) (with h^j in place of h). Then,

$$\phi_n^{LHS}(\theta) = \sum_{j=1}^{d} \theta \frac{1}{n} \sum_{i=1}^{n} c_i^j(n,\theta).$$

This is just the sum of d one-dimensional cases. Thus, it follows from Proposition 2 that

$$\phi^{LHS}(\theta) \;=\; \lim_{n\to\infty} \phi_n^{LHS}(\theta) \;=\; \theta \sum_{j=1}^{d} \int_0^1 h^j(z^j) dz^j \;=\; \theta \int_{[0,1]^d} h(z) dz = \theta\mu_d. \qquad \square$$

As before, for any closed subset F of $\mathbb{R}$, as long as $\mu_d \notin F$ we have a decay with superexponential rate, i.e.,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} S_n \in F\right) \le -\inf_{x\in F} I^{LHS}(x) = -\infty.$$

Moreover, as in the one-dimensional case, when $h(\cdot)$ is separable the upper bound for the probability of a large deviation is smaller under Latin Hypercube sampling than under Monte Carlo sampling for any sample size n, i.e., we have an extension of Proposition 3:

Proposition 4 *Consider the setting of Theorem 2. Let $I^{MC}(x)$ and $I^{LHS}(n,x)$ denote the (non-asymptotic) functions defined in* (12) *respectively for Monte Carlo and for LHS. Then, for* any *sample size n and all x we have that $I^{LHS}(n,x) \ge I^{MC}(x)$.*

Proof. This follows immediately from Proposition 3, noticing that both $\phi_n^{LHS}(\cdot)$ and $\phi^{MC}(\cdot)$ can be written in the form $\phi_n(\theta) = \sum_{j=1}^{d} \phi_n^{(j)}(\theta)$, where $\phi_n^{(j)}$ is defined as in (8) for the jth component. $\square$

4.2 Case 2: The Bounded Residual Case

We now turn to the case where $h(\cdot)$ is not separable, but its residual term in the ANOVA decomposition is bounded. Recall that we can decompose h as $h(z) = \mu + h_{add}(z) + h_{resid}(z)$, where $\mu = \mathbb{E}[h(Z)]$, $h_{add}(z) = \sum_{j=1}^{d} h^j(z^j)$ and $\mathbb{E}[h^j(Z^j)] = \mathbb{E}[h_{resid}(Z)] = 0$ for all j. The bounded residual case assumes that $-m \leq h_{resid}(\cdot) \leq M$ where $m, M \geq 0$. Note that this class of functions includes all functions that are bounded themselves. Of course, the results below are useful only in case the bounds on the residual are significantly smaller than the bounds on the whole function, i.e. the function may not be separable but must have a strong additive component.

The superexponential rate of decay for the deviation probability no longer holds in general when the separability condition is removed. However, we will show that the superexponential rate does still hold if the deviation is sufficiently large. This is stated in the proposition below:

Proposition 5 *Suppose $h : [0,1]^d \mapsto \mathbb{R}$ is a function such that its residual component satisfies $-m \leq h_{resid}(\cdot) \leq M$ for some $m, M \geq 0$. Suppose also that each term h^j of the additive component h_{add} satisfies Assumption 1. Let Z be a random vector with independent components uniformly distributed on $[0,1]^d$, and define $\mu := \mathbb{E}[h(Z)] = \int_{[0,1]^d} h(z)dz$.*

Then, for any a, b such that $a < \mu < b$, the LHS estimator S_n/n of μ satisfies

$$\lim_{n\to\infty} \frac{1}{n} \log \left[\mathbb{P}\left(\frac{1}{n} S_n \geq b \right) \right] = -\infty \quad \text{if } b > \mu + M \tag{29}$$

$$\lim_{n\to\infty} \frac{1}{n} \log \left[\mathbb{P}\left(\frac{1}{n} S_n \leq a \right) \right] = -\infty \quad \text{if } a < \mu - m. \tag{30}$$

Proof. By assumption, we have

$$\mu + \sum_{j=1}^{d} h^j(z^j) - m \;\leq\; h(z) \;\leq\; \mu + \sum_{j=1}^{d} h^j(z^j) + M.$$

For $\theta > 0$,

$$\exp\left(\theta \left[\mu + \sum_{j=1}^{d} h^j(z^j) - m \right] \right) \;\leq\; \exp(\theta h(z)) \;\leq\; \exp\left(\theta \left[\mu + \sum_{j=1}^{d} h^j(z^j) + M \right] \right).$$

Also, by the properties of the integral, for any $i \in 1, 2, \ldots, n$ and any permutation index k, we have that

$$\begin{aligned}
&\int_{a_i^d(k)}^{b_i^d(k)} \cdots \int_{a_i^1(k)}^{b_i^1(k)} n^d \exp\left(\theta \left[\mu + \sum_{j=1}^{d} h^j(z^j) - m \right] \right) dz \\
&\qquad \leq \int_{a_i^d(k)}^{b_i^d(k)} \cdots \int_{a_i^1(k)}^{b_i^1(k)} n^d \exp(\theta h(z))\, dz \\
&\qquad \leq \int_{a_i^d(k)}^{b_i^d(k)} \cdots \int_{a_i^1(k)}^{b_i^1(k)} n^d \exp\left(\theta \left[\mu + \sum_{j=1}^{d} h^j(z^j) + M \right] \right) dz.
\end{aligned}$$

Then, since all of these integrals are positive, we have

$$\frac{1}{(n!)^d}\sum_{k=1}^{(n!)^d}\prod_{i=1}^{n}\int_{a_i^d(k)}^{b_i^d(k)}\cdots\int_{a_i^1(k)}^{b_i^1(k)} n^d \exp\left(\theta\left[\mu+\sum_{j=1}^{d}h^j(z^j)-m\right]\right)\,dz \tag{31}$$
$$\le \exp(n\phi_n(\theta))$$
$$\le \frac{1}{(n!)^d}\sum_{k=1}^{(n!)^d}\prod_{i=1}^{n}\int_{a_i^d(k)}^{b_i^d(k)}\cdots\int_{a_i^1(k)}^{b_i^1(k)} n^d \exp\left(\theta\left[\mu+\sum_{j=1}^{d}h^j(z^j)+M\right]\right)\,dz,$$

where we know from (28) that

$$\exp(n\phi_n^{LHS}(\theta)) = \frac{1}{(n!)^d}\sum_{k=1}^{(n!)^d}\prod_{i=1}^{n} n^d \int_{a_i^d(k)}^{b_i^d(k)}\cdots\int_{a_i^1(k)}^{b_i^1(k)} \exp(\theta h(z))dz.$$

By manipulating (31) as in the proof of Theorem 2 we obtain

$$\prod_{i=1}^{n}\exp(\theta(\mu-m))\prod_{j=1}^{d}\int_{\frac{i-1}{n}}^{\frac{i}{n}} n\exp\left(\theta h^j(z^j)\right)\,dz^j \;\le\; \exp(n\phi_n(\theta))$$
$$\le \prod_{i=1}^{n}\exp(\theta(\mu+M))\prod_{j=1}^{d}\int_{\frac{i-1}{n}}^{\frac{i}{n}} n\exp\left(\theta h^j(z^j)\right)\,dz^j.$$

By defining quantities $c_i^j(n,\theta)$ as in (19) (with h^j in place of h) we can rewrite the above inequalities as

$$\prod_{i=1}^{n}\left[\exp(\theta(\mu-m))\prod_{j=1}^{d}\exp(\theta c_i^j(n,\theta))\right] \;\le\; \exp(n\phi_n(\theta)) \;\le\; \prod_{i=1}^{n}\left[\exp(\theta(\mu+M))\prod_{j=1}^{d}\exp(\theta c_i^j(n,\theta))\right],$$

which further simplifies to

$$\exp\left(\theta\sum_{j=1}^{d}\sum_{i=1}^{n}c_i^j(n,\theta)+\theta n(\mu-m)\right) \;\le\; \exp(n\phi_n(\theta)) \;\le\; \exp\left(\theta\sum_{j=1}^{d}\sum_{i=1}^{n}c_i^j(n,\theta)+\theta n(\mu+M)\right),$$

and so

$$\frac{1}{n}\theta\sum_{j=1}^{d}\sum_{i=1}^{n}c_i^j(n,\theta)+\theta(\mu-m) \;\le\; \phi_n(\theta) \;\le\; \frac{1}{n}\theta\sum_{j=1}^{d}\sum_{i=1}^{n}c_i^j(n,\theta)+\theta(\mu+M). \tag{32}$$

Note that, as shown in the proof of Proposition 2, the term $\frac{1}{n}\sum_{j=1}^{d}\sum_{i=1}^{n}c_i^j(n,\theta)$ converges to $\int_{[0,1]^d} h_{add}(z)\,dz = \mathbb{E}[h_{add}(Z)]$, which is equal to zero. Hence, given $\varepsilon>0$ we have that

$$\theta\left(\mu-m-\varepsilon\right) \;\le\; \phi_n(\theta) \;\le\; \theta\left(\mu+M+\varepsilon\right)$$

for n large enough and thus

$$\sup_{\theta\ge 0}\left[\theta x-\theta(\mu+M+\varepsilon)\right] \;\le\; \sup_{\theta\ge 0}\left[\theta x-\phi_n(\theta)\right] \;\le\; \sup_{\theta\ge 0}\left[\theta x-\theta(\mu-m-\varepsilon)\right]$$

for any x. Clearly, when $x > \mu + M + \varepsilon$ the left-most term is infinite. Since ε was chosen arbitrarily, it follows from (9) that

$$\lim_n \frac{1}{n} \log \left[\mathbb{P}\left(\frac{1}{n} S_n \geq b \right) \right] = -\infty \quad \text{if } b > \mu + M.$$

For $\theta < 0$, we can use a similar argument to conclude that

$$\lim_n \frac{1}{n} \log \left[\mathbb{P}\left(\frac{1}{n} S_n \leq a \right) \right] = -\infty \quad \text{if } a < \mu - m. \qquad \square$$

Note that this result generalizes the separable case, since in that context we have $h_{resid} \equiv 0$ and so $m = M = 0$.

4.3 Case 3: The Monotone Case

We move now to the case of functions that possess a certain form of monotonicity, in the specific sense defined below:

Definition 1 *A function $h : [0,1]^d \mapsto \mathbb{R}$ is said to be* monotone *if it is monotone in each argument when the other arguments are held fixed, i.e., if for all $z \in [0,1]^d$ and all $j = 1, \ldots, d$ we have that either*

$$h(z^1, \ldots, z^{j-1}, z^j, z^{j+1}, \ldots, z^d) \;\leq\; h(z^1, \ldots, z^{j-1}, w^j, z^{j+1}, \ldots, z^d) \quad \textit{for all } w^j \in [0,1],\; z^j \leq w^j$$

or

$$h(z^1, \ldots, z^{j-1}, z^j, z^{j+1}, \ldots, z^d) \;\geq\; h(z^1, \ldots, z^{j-1}, w^j, z^{j+1}, \ldots, z^d) \quad \textit{for all } w^j \in [0,1],\; z^j \leq w^j.$$

The relevance of this case is due to the fact that monotone functions preserve a property called negative dependence, which we for completeness we define below:

Definition 2 *Random variables $Y_i, i = i \ldots n$ are called* negatively dependent *if*

$$\mathbb{P}(Y_1 \leq y_1, \ldots, Y_n \leq y_n) \;\leq\; \mathbb{P}(Y_1 \leq y_1) \cdots \mathbb{P}(Y_n \leq y_n).$$

The following lemma from Jin et al. (2003) gives an important property of negatively dependent random variables.

Lemma 4 *If $Y_i, i = 1, \ldots n$ are nonnegative and negatively dependent and if $\mathbb{E}[Y_i] < \infty, i = 1 \ldots n$ and $\mathbb{E}[Y_1 \cdot \cdots \cdot Y_n] < \infty$, then $\mathbb{E}[Y_1 \cdot \cdots \cdot Y_n] \leq \mathbb{E}[Y_1] \cdots \mathbb{E}[Y_n]$.*

In our context, we are interested in the case where $Y_i = g(Z_i^1, \ldots, Z_i^d)$, where g is nonnegative monotone and the vectors Z_i, $i = 1, \ldots, n$ are LH samples of a Uniform($[0,1]^d$) random vector Z. Jin et al. (2003) show that (i) Latin Hypercube samples are negatively dependent, and (ii) monotone functions preserve negative dependence. Hence, if g is monotone then the random variables $Y_1, \ldots, Y_n$ are negatively dependent. In what follows we will use such a property repeatedly.

Unfortunately, in the present case it is not clear whether we can apply the Gärtner-Ellis Theorem to derive large deviations results — the reason being that we do not know if negative dependence suffices to ensure convergence of the functions $\{\phi_n^{LHS}(\theta)\}$. We must note, however, that the Gärtner-Ellis Theorem only provides *sufficient* (but not necessary) conditions for the validity of a large deviation principle; that is, it is possible that a large deviation principle holds in the present case even if the assumptions of the theorem are violated. A definite answer to that question is still an open problem.

Nevertheless, we can still derive results that fit the framework of the discussion around (13). Proposition 6 below provides results that are analogous to Propositions 3 and 4, i.e., it shows that the Monte Carlo rate $I^{MC}(x)$ is dominated by $I^{LHS}(n, x)$. Jin et al. (2003) show that, when the quantity to be estimated is a *quantile*, the upper bound on a deviation probability with negatively dependent sampling is less than that from Monte Carlo sampling. Here we show a similar result but in the context of estimation of the mean.

Proposition 6 *Suppose* $h : [0,1]^d \mapsto \mathbb{R}$ *is a monotone function in the sense of Definition 1. Let* Z *be a random vector uniformly distributed on* $[0,1]^d$*, and assume that* $\mathbb{E}[\exp(\theta h(Z)] < \infty$ *for all* $\theta \in \mathbb{R}$*. Let* $I^{MC}(x)$ *and* $I^{LHS}(n, x)$ *denote the (non-asymptotic) functions defined in* (12) *respectively for Monte Carlo and for LHS. Then, for* any *sample size* n *and all* x *we have that* $I^{LHS}(n, x) \geq I^{MC}(x)$.

Proof. We will show that, if h is monotone, then $\phi_n^{LHS}(\theta) \leq \phi^{MC}(\theta)$ for all n and all θ. Note initially that from (8) we have

$$\exp(n\phi_n^{LHS}(\theta)) \;=\; \mathbb{E}\left[\prod_{i=1}^{n} \exp(\theta h(Z_i))\right]. \tag{33}$$

Clearly, the monotonicity of the exponential function (in the standard sense) implies that $\exp(\theta h(\cdot))$ is a monotone function in the sense of Definition 1 and so $\exp(\theta h(\cdot))$ preserves negative dependence. By assumption, $\mathbb{E}[\exp(\theta h(Z)] < \infty$ for all $\theta \in \mathbb{R}$, and hence a direct application of Cauchy-Schwarz inequality shows that $\mathbb{E}\left[\prod_{i=1}^{n} \exp(\theta h(Z_i))\right] < \infty$. Thus, we can apply Lemma 4 to conclude that, for any θ,

$$\mathbb{E}\left[\prod_{i=1}^{n} \exp(\theta h(Z_i))\right] \;\leq\; \prod_{i=1}^{n} \mathbb{E}\left[\exp(\theta h(Z_i))\right] \;=\; \left(\mathbb{E}\left[\exp(\theta h(Z))\right]\right)^n, \tag{34}$$

the latter equality being a consequence of the unbiasedness of LHS estimators. Combining (33) and (34), it follows that

$$\phi_n^{LHS}(\theta) \;\leq\; \log \mathbb{E}\left[\exp(\theta h(Z))\right] \;=\; \phi^{MC}(\theta). \qquad \square$$

5 Examples

We now show examples comparing the probability of a large deviation under Latin Hypercube sampling and Monte Carlo sampling on six different functions. For each function, we generated (unless stated otherwise) both Monte Carlo and Latin Hypercube samples for various sample sizes n (n = 50, 100, 500, 1000, 5000, 10000). For each sampling method and each n, we estimated the probability that the estimator S_n/n deviates by more than 0.1% of the true mean. This was calculated by doing 1000 independent replications and counting the number of occurrences in which the sample mean was not within 0.1% of the true mean, divided by 1000.

In each graph below, the x-axis represents the different sample sizes while the y-axis shows the estimated large deviations probabilities for each sample size. Estimates for both Latin Hypercube and Monte Carlo sampling are graphed as well as the upper and lower 95% confidence intervals for each estimate (represented by the dashed lines).

Example 1: $h(z) = \log(\frac{1}{\sqrt{z_1}})$. This is a one-dimensional function with a singularity at $z_1 = 0$. Its integral on $[0, 1]$ is equal to $\frac{1}{2}$. LHS considerably outperforms Monte Carlo sampling with a large deviation probability of essentially zero when $n = 5000$. Meanwhile the probability of a large deviation is still roughly 0.9 for Monte Carlo sampling with $n = 10000$. This is shown in Figure 1 (left).

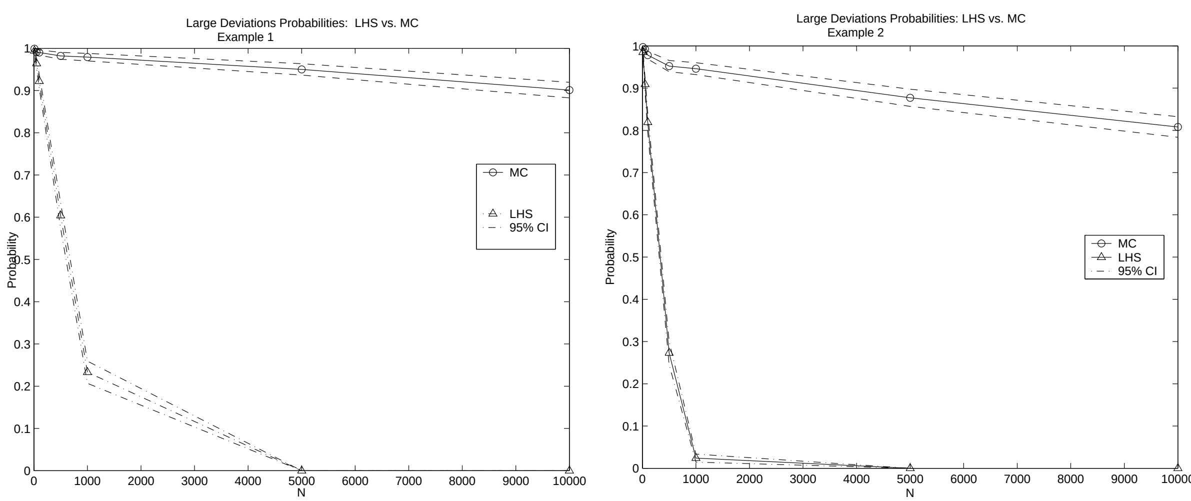

Figure 1: Examples 1 (left) and 2 (right).

Example 2: $h(z) = \log(z_1 z_2 z_3 z_4 z_5)$. This function is separable, so by Theorem 2 we expect the large deviation probability to be essentially zero under LHS with large n. The integral of the function is -5. Again LHS dominates the Monte Carlo sampling which has a large deviation probability of nearly 0.8 at $n = 10000$. This is also shown in Figure 1 (right).

Example 3: $h(z) = \log(\frac{1}{\sqrt{z_1}} + \frac{1}{\sqrt{z_2}})$. While not separable, this function is monotone in both z_1 and z_2. Its integral is $\frac{5}{4}$. From Proposition 6, we know that the upper bound for the large deviations probability is guaranteed to be smaller under LHS than under Monte Carlo for each value of n, and indeed we see that LHS again dominates Monte Carlo. This is shown in Figure 2 (left).

Example 4: $h(z) = \log\left[2 + \sin(2\pi z_1)\cos(2\pi z_2^2)\right]$. This function is neither separable nor monotone — in fact, it is highly non-separable. We have no guarantee that LHS will produce a lower probability of a large deviation than Monte Carlo sampling. This function has integral 0.6532, which was calculated numerically. As shown in Figure 2 (right), the two sampling methods yield similar results for this function. In fact, from the graph we see that it is possible for Monte Carlo sampling to have a lower probability of large deviation than LHS, even at $n = 10000$.

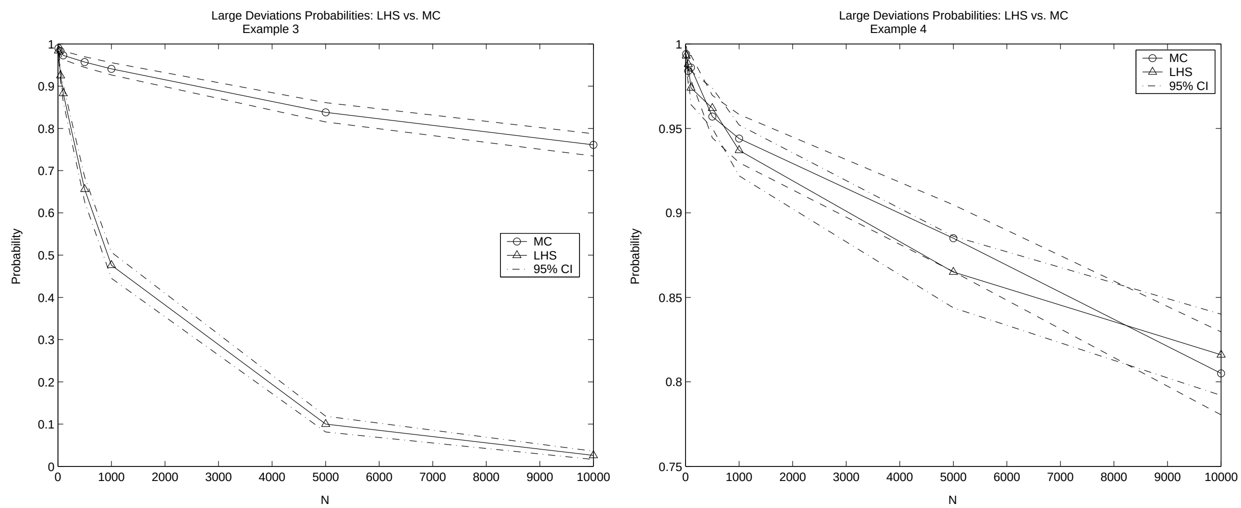

Figure 2: Examples 3 (left) and 4 (right).

Example 5: $h(z) = (z_1 - \frac{1}{2})^2(z_2 - \frac{1}{2})^2$. Again, this function is neither separable nor monotone. When considering deviations of 0.1% from its mean of $\frac{1}{144}$, we can see that neither deviation probability approaches zero very quickly. See Figure 3 (left). However, if we measure larger deviations such as any value of the sample mean outside of the interval $[0, \frac{3}{144}]$ (note that the function itself is bounded below by zero), the deviation probability approaches zero more rapidly for LHS. This is shown in Figure 3 (right) for sample sizes from 1 to 10 (10000 replications).

Example 6: In order to connect the results derived in this paper with the stochastic optimization concepts discussed in Section 1, our final example is the objective function of a two-stage stochastic linear program. Models of this type are extensively used in operations research (see, e.g., Birge and Louveaux 1997 for a comprehensive discussion). The basic idea is that decisions are made in two stages: an initial decision (say, y) made before any uncertainty is realized, and a "correcting" decision (say, u) which is made after the uncertain factors are known. A typical example is that of

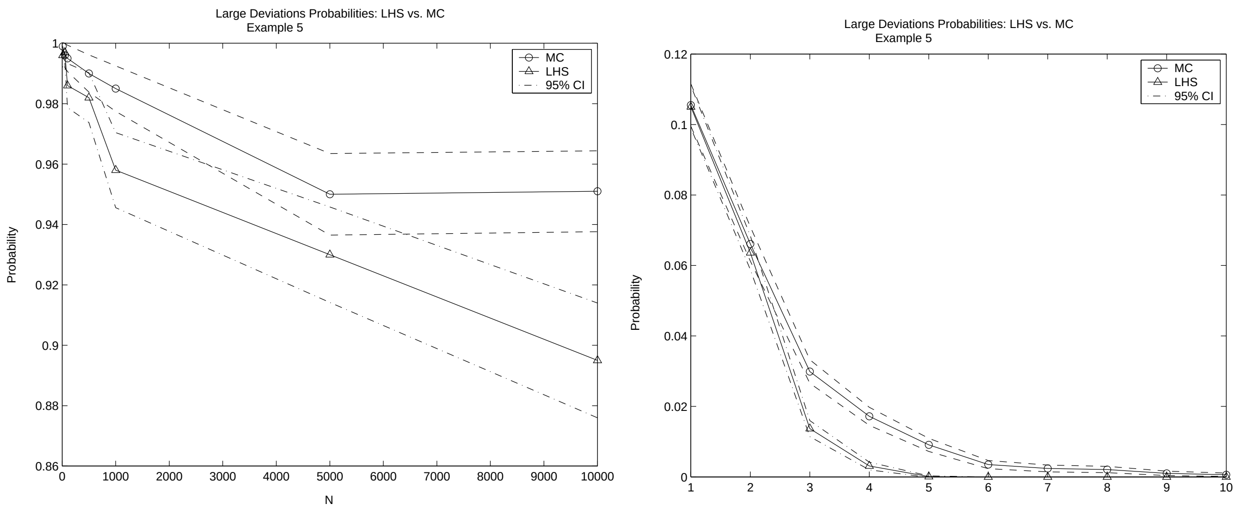

Figure 3: Example 5.

an inventory system, where one must decide on the order quantity before observing the demand. Within the class of two-stage problems, we consider those of the form

$$\min_{y \in \mathcal{Y}} c^t y + \mathbb{E}[Q(y, X)], \tag{35}$$

where $\mathcal{Y}$ is a convex polyhedral set and

$$Q(y, X) = \inf \left\{ q^t u \, : \, Wu \leq X - Ty, \ u \geq 0 \right\} \tag{36}$$

In the above, X is a d-dimensional random vector with independent components and corresponding finite support cdfs $F_1, \ldots, F_d$. Let $\Psi(y, X)$ denote the function $c^t y + Q(y, X)$; then, we see that the above problem falls in the framework of (4).

Next, for each $y \in \mathcal{Y}$ and $z \in [0, 1]^d$ let $h_y(z) = \Psi(y, [F_1^{-1}(z_1), \ldots, F_d^{-1}(z_d)])$. Note that $h_y(\cdot)$ is monotone in the sense of Definition 1, since increasing the right-hand side of the minimization problem in (36) enlarges the feasible set and hence the optimal value decreases, i.e., $Q(y, \cdot)$ (and therefore $\Psi(y, \cdot)$) is decreasing. Also, each F_j^{-1} is a monotone function. Thus, in light of the results discussed in Section 4, we expect LHS to perform well in terms of the deviation probabilities of the pointwise estimators $\hat{\psi}_n(y) := \frac{1}{n} \sum_{j=1}^n \Psi(y, X_j)$.

Figure 4 shows the results obtained for a two-stage problem from the literature related to capacity expansion. The problem has the formulation (35)-(36). The figure on the left depicts a deviation probability for the pointwise estimator $\hat{\psi}_n(y^*)$, where y^* is the optimal solution of the problem. More specifically, the figure shows the estimated value (over 1000 replications) of $\mathbb{P}(|\hat{\psi}_n(y^*) - \nu^*| > 0.1\nu^*)$ (i.e., a 10% deviation from the mean), where $\nu^* = \psi(y^*) = \mathbb{E}[\hat{\psi}_n(y^*)]$

is the corresponding objective function value. As expected, that probability goes to zero faster under LHS than under Monte Carlo. The figure on the right depicts the estimated probability of obtaining an incorrect solution, i.e. $\mathbb{P}(\hat{y}_n \neq y^*)$. As discussed in Shapiro et al. (2002), this is a *well-conditioned* problem, in the sense that the latter probability goes to zero quite fast under Monte Carlo. We can see however that convergence of that quantity under LHS is even faster. Here we see a confirmation of the results in Homem-de-Mello (2006) — the faster convergence of deviation probabilities for pointwise estimators leads to a faster convergence of the probability of obtaining an incorrect solution.

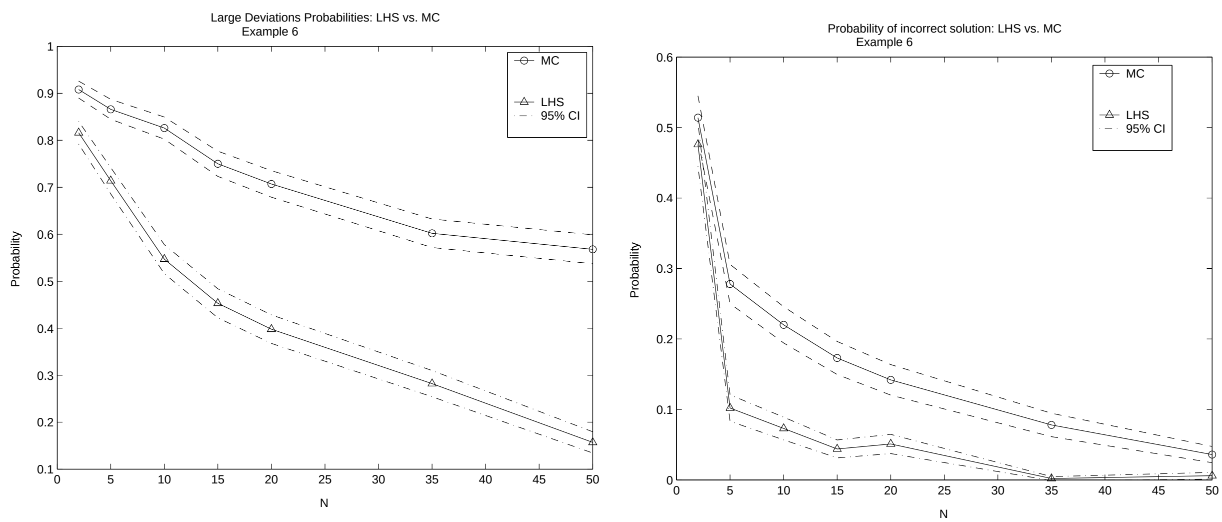

Figure 4: Deviation probabilities (left) and probabilities of obtaining and incorrect solution (right) for Example 6.

6 Conclusions

We have studied large deviations properties of estimators obtained with Latin Hypercube sampling. We have shown that LHS satisfies a large deviation principle for real-valued functions of one variable and for separable real-valued functions in multiple variables, with the rate being superexponential. We have also shown that the upper bound of the probability of a large deviation is smaller under LHS than it is for Monte Carlo sampling in these cases regardless of the sample size. This is analogous to the result that Latin Hypercube sampling gives a smaller variance than Monte Carlo sampling in these same cases since Var^{LHS} approaches the variance of the residual term, which in these cases is nonexistent.

We have also shown that, if the underlying function is monotone in each component, then the upper bound for the large deviation probability is again less than that of Monte Carlo sampling regardless of the sample size. Again, this is analogous to the fact that the variance from LHS is no greater than that of Monte Carlo sampling when the function is monotone in all arguments. Unfortunately we do not know whether the large deviations rate is superexponential, as it is in the separable case.

Large deviations results for LHS for general functions still remain to be shown, though the Latin Hypercube variance results found in the literature seem to provide a good direction. In general, the variance of a Latin Hypercube estimator may not be smaller than that of a Monte Carlo estimate (recall the bound $\mathrm{Var}^{LHS} \leq \frac{n}{n-1}\mathrm{Var}^{MC}$ proven by Owen (1997)); however, asymptotically it is no worse. This might also be the case for the upper bound of the large deviations probability. Also, just as Stein (1987) has shown that, asymptotically, Var^{LHS} is equal to just the variance of the residual term, it is possible that the rate of convergence of large deviations probabilities for LHS depend only on the residual terms — indeed, we have shown that, in case the residual term is bounded, the rate of convergence depends directly on the values of such bounds.

As discussed earlier, large deviations theory plays an important role in the study of approximation methods for stochastic optimization problems. The results in the paper show that LHS can significantly help in that regard, at least for a class of problems. It is also possible that these results will help with the derivation of properties of discrepancies of high-dimensional sequences constructed via padding quasi-Monte Carlo with LHS (instead of padding with standard Monte Carlo as in Ökten et al. (2006)). Research on that topic is underway.

References

R. Bartle. *The Elements of Real Analysis.* Wiley, New York, 2nd. edition, 1987.

J. R. Birge and F. Louveaux. *Introduction to Stochastic Programming.* Springer Series in Operations Research. Springer-Verlag, New York, NY, 1997.

J. A. Bucklew. *Introduction to Rare Event Simulation.* Springer-Verlag, New York, 2004.

L. Dai, C. H. Chen, and J. R. Birge. Convergence properties of two-stage stochastic programming. *J. Optim. Theory Appl.*, 106(3):489–509, 2000.

A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications.* Springer-Verlag, New York, NY, 2nd. edition, 1998.

F. den Hollander. *Large Deviations.* Number 14 in Fields Institute Monographs. American Mathematical Society, Providence, RI, 2000.

S. S. Drew and T. Homem-de-Mello. Some large deviations results for Latin Hypercube Sampling. In M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, editors, *Proceedings of the 2005 Winter Simulation Conference*, pages 673–681. IEEE Press, 2005.

J. B. Hiriart-Urruty and C. Lemarechal. *Convex Analysis and Minimization Algorithms*, volume II. Springer-Verlag, Berlin, Germany, 1993.

T. Homem-de-Mello. On rates of convergence for stochastic optimization problems under non-i.i.d. sampling. Manuscript, available on *Optimization Online* (`www.optimization-online.org`). Submitted for publication, 2006.

N. Hoshino and A. Takemura. On reduction of finite sample variance by extended Latin hypercube sampling. *Bernoulli*, 6(6):1035–1050, 2000.

X. Jin, M. C. Fu, and X. Xiong. Probabilistic error bounds for simulation quantile estimators. *Management Science*, 49(2):230–246, 2003.

Y. M. Kaniovski, A. J. King, and R. J.-B. Wets. Probabilistic bounds (via large deviations) for the solutions of stochastic programming problems. *Ann. Oper. Res.*, 56:189–208, 1995.

A. J. King and R. T. Rockafellar. Asymptotic theory for solutions in statistical estimation and stochastic programming. *Mathematics of Operations Research*, 18:148–162, 1993.

A. Kleywegt, A. Shapiro, and T. Homem-de-Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2):479–502, 2001.

A. M. Law and W. D. Kelton. *Simulation Modeling and Analysis*. McGraw-Hill, New York, NY, 3rd. edition, 2000.

W. Loh. On Latin hypercube sampling. *The Annals of Statistics*, 24(5):2058–2080, 1996.

M. D. McKay, R. J. Beckman, and W. J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21: 239–245, 1979.

G. Ökten, B. Tuffin, and V. Burago. A central limit theorem and improved error bounds for a hybrid-Monte Carlo sequence with applications in computational finance. *Journal of Complexity*, 22(4):435–458, 2006.

A. B. Owen. A central limit theorem for Latin hypercube sampling. *J. Roy. Statist. Soc. Ser. B*, 54:541–551, 1992.

A. B. Owen. Monte Carlo variance of scrambled net quadrature. *SIAM J. Numer. Anal.*, 34(5): 1884–1910, 1997.

S. M. Robinson. Analysis of sample-path optimization. *Mathematics of Operations Research*, 21: 513–528, 1996.

H. L. Royden. *Real Analysis*. Macmillan, New York, 3rd edition, 1988.

A. Shapiro. Asymptotic analysis of stochastic programs. *Annals of Operations Research*, 30:169–186, 1991.

A. Shapiro. On complexity of multistage stochastic programs. *Operations Research Letters*, 34:1–8, 2006.

A. Shapiro and T. Homem-de-Mello. On rate of convergence of Monte Carlo approximations of stochastic programs. *SIAM Journal on Optimization*, 11:70–86, 2000.

A. Shapiro, T. Homem-de-Mello, and J. C. Kim. Conditioning of convex piecewise linear stochastic programs. *Mathematical Programming*, 94:1–19, 2002.

A. Shapiro and A. Nemirovski. On complexity of stochastic programming problems. In V. Jeyakumar and A. M. Rubinov, editors, *Continuous Optimization: Current Trends and Applications*, pages 111–144. Springer, 2005.

A. Shwartz and A. Weiss. *Large Deviations for Performance Analysis*. Chapman & Hall, London, 1995.

M. L. Stein. Large sample properties of simulations using Latin hypercube sampling. *Technometrics*, 29:143–151, 1987.