https://pubsonline.informs.org/journal/ijoc

INFORMS JOURNAL ON COMPUTING
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0899-1499, EISSN 1526-5528

Submitted to *INFORMS Journal on Computing*

Efficient Nested Simulation Experiment Design via the Likelihood Ratio Method

Ben Mingbin Feng
Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, CANADA
ben.feng@uwaterloo.ca, https://www.math.uwaterloo.ca/ mbfeng/

Eunhye Song
H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA
eunhye.song@isye.gatech.edu, https://eunhyesong.info

Abstract. In nested simulation literature, a common assumption is that the experimenter can choose the number of outer scenarios to sample. This paper considers the case when the experimenter is given a fixed set of outer scenarios from an external entity. We propose a nested simulation experiment design that pools inner replications from one scenario to estimate another scenario's conditional mean via the likelihood ratio method. Given the outer scenarios, we decide how many inner replications to run at each outer scenario as well as how to pool the inner replications by solving a bi-level optimization problem that minimizes the total simulation effort. We provide asymptotic analyses on the convergence rates of the performance measure estimators computed from the optimized experiment design. Under some assumptions, the optimized design achieves $\mathcal{O}(\Gamma^{-1})$ mean squared error of the estimators given simulation budget Γ. Numerical experiments demonstrate that our design outperforms a state-of-the-art design that pools replications via regression.

Funding: This research was supported by [CMMI-2045400, National Science Foundation] and [RGPIN-2018–03755, Natural Sciences and Engineering Research Council of Canada].

Key words: Nested simulation, Design of experiment design, Likelihood ratio method, Enterprise Risk Management, Uncertainty Quantification

History: Received: Month DD, YYYY; Accepted: Month DD, YYYY; Published Online: Month DD, YYYY

1. Introduction

We consider a simulation experiment design problem known as *nested simulation* (Hong et al. 2017) whose objective is to estimate some functional of a conditional mean of a stochastic simulation output. Specifically, we are interested in quantities of the form

$$\rho\left(\mu(\boldsymbol{\theta})\right), \quad \mu(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}[g(\mathbf{X})] \equiv \mathrm{E}\left[g(\mathbf{X})|\boldsymbol{\theta}\right],$$

where $\boldsymbol{\theta}$ is a simulation input vector referred to as the *outer scenario*, $\mathbf{X}$ is a random vector whose distribution is parameterized by $\boldsymbol{\theta}$, and $g(\mathbf{X})$ is the stochastic simulation output. Each simulation run that generates $g(\mathbf{X})$ given $\boldsymbol{\theta}$ is referred to as an *inner replication*. Thus, $\mu(\boldsymbol{\theta})$ represents the conditional mean of the simulation output given the outer scenario, $\boldsymbol{\theta}$. Finally, ρ is a functional that maps the distribution of $\mu(\boldsymbol{\theta})$ to a real number. Two classes of ρ are considered in this study:

(i) $\rho(\mu(\boldsymbol{\theta})) = \mathrm{E}[\zeta(\mu(\boldsymbol{\theta}))]$ for some real-valued function ζ.

(ii) $\rho(\mu(\boldsymbol{\theta})) = \inf\{z \in \mathbb{R} : \mathbb{P}(\mu(\boldsymbol{\theta}) \leq z) \geq \alpha\}$ for some user-specified $0 < \alpha < 1$, i.e., the α-quantile or the α-level Value-at-Risk (VaR) of $\mu(\boldsymbol{\theta})$.

We refer to a generic $\rho(\mu(\boldsymbol{\theta}))$ as the *nested statistic* in this paper.

Nested simulation is a common tool for quantifying uncertainty of a stochastic simulation model. There is rich literature on simulation model risk analysis (Barton et al. 2022) that adopts the prediction interval of $\mu(\boldsymbol{\theta})$ (i.e., quantiles) to quantify the variability in the simulation output caused by uncertain input parameter $\boldsymbol{\theta}$. The enterprise risk management (ERM) literature (Broadie et al. 2015, Hong et al. 2017, Dang et al. 2022, Wang et al. 2022) also applies nested simulation to manage risk exposures of portfolios of complex financial instruments. In this case, $\mu(\boldsymbol{\theta})$ is typically a financial portfolio's profit/loss at a future time which depends on underlying risk factors such as equity prices and interest rates represented by $\boldsymbol{\theta}$. Popular choices for ρ include exceedance probability, variance, quantile (Value-at-Risk), and Conditional Value-at-Risk (CVaR) (Rockafellar and Uryasev 2002).

Standard nested simulation (SNS) is a classical experiment design that samples M outer scenarios, runs N inner replications at each sampled outer scenario $\boldsymbol{\theta}$ to estimate $\mu(\boldsymbol{\theta})$, and computes a nested statistic from them. Estimating a nested statistic with small bias and variance requires both M and N to be large. Thus, designing an efficient nested simulation experiment has attracted much attention from researchers and practitioners.

Our objective is to propose an experiment design from which various nested statistics can be computed with estimation errors comparable to SNS while spending significantly fewer replications. We propose an experiment design that has different numbers of inner replications at different outer scenario. Furthermore, we focus on the case where the experimenter has no control over outer scenario generation and *must work with a given set of outer scenarios*. Indeed, such a case arises frequently in practice. For instance, Song et al. (2020) describe the nested simulation experiment conducted at General Motors (GM) to quantify uncertainty in their market simulator. In their context, $\boldsymbol{\theta}$ is a consumer utility parameter vector for vehicle features whose distribution is estimated from a conjoint analysis. Specifically, $\boldsymbol{\theta}$ has a few thousand dimensions, where each represents a survey respondent's preference for a vehicle feature (e.g., color-black). They impose a Bayesian prior on $\boldsymbol{\theta}$ and updates it to the posterior by incorporating the conjoint data assuming that the respondents make their choices following a discrete choice model. Since there is no conjugacy to

exploit in this model, a Markov chain Monte Carlo algorithm is applied to generate an approximately independent and identically distributed (i.i.d.) sample from the posterior, which discards many realized $\boldsymbol{\theta}$s to avoid high positive correlations among them. This procedure takes significantly longer than inner replications (market simulation). Thus, it is more natural to consider the computational budget for generating $\boldsymbol{\theta}$s separately from the cost of inner replications. Moreover, because GM runs several other market analyses using a common set of $\boldsymbol{\theta}$s across different business divisions, it is required to work with the same given set of $\boldsymbol{\theta}$s for uncertainty quantification for consistency.

In practical ERM applications, the underlying risk factors (outer scenarios) may be provided by regulators or by a separate modeling team in a firm, so the experimenter has little control over them. Risk and Ludkovski (2018) point out that "the outer scenarios are usually treated by practitioners as a fixed object (namely coming from an exogenous economic scenario generator (ESG))...one cannot add (or subtract) further scenarios" Li and Feng (2021) also mention that "(in nested simulation) it is often necessary to run a pre-determined set of economic scenarios to project the institution's cash flows." Ha and Bauer (2022) mention that "the same firm-wide scenarios are typically used for evaluating the company's asset and liability portfolios." Therefore, it has an important practical impact to study the design where the number of outer scenarios is fixed.

A long-studied experiment design question for nested simulation is how to allocate given simulation budget (the number of inner replications) Γ to minimize estimation error. For SNS, we have $\Gamma = NM$ and the question boils down to choosing M and N so that the nested statistic estimator's mean squared error (MSE) is minimized. Lee (1998) shows that for an exceedance probability estimator, its bias and variance respectively diminish in $\mathcal{O}(N^{-1})$ and $\mathcal{O}(M^{-1})$.[1] This analysis leads to the asymptotically optimal allocation, $N = \mathcal{O}(\Gamma^{1/3})$ and $M = \mathcal{O}(\Gamma^{2/3})$. Gordy and Juneja (2010) establish the same result for a quantile estimator and analyze the bias of CVaR. Sun et al. (2011) study estimating the variance of $\mu(\boldsymbol{\theta})$ using analysis of variance and shows that as M increases, the choice of N minimizing the asymptotic variance of the variance estimator converges to a constant.

Recent studies depart from the M-outer-N-inner experiment design to enhance efficiency. Broadie et al. (2011) propose a sequential experiment design to estimate an exceedance probability, which selects outer scenarios adaptively and allocates inner replications to the scenarios more critical for reducing the MSE of the estimator; they achieve asymptotic MSE of $\mathcal{O}(\Gamma^{-4/5+\epsilon})$ for any positive ϵ. Giles and Haji-Ali (2019) extend their idea to propose a multi-level Monte Carlo estimator. Another line of research considers *pooling* inner replications from some or all outer scenarios to estimate $\mu(\boldsymbol{\theta})$ for each outer scenario $\boldsymbol{\theta}$. These approaches fit a metamodel for $\mu(\boldsymbol{\theta})$ by running an initial experiment design, then use the fitted model to estimate a nested statistic with no additional inner replications. Liu and Staum (2010) adopts a stochastic

[1] See Section 2 for the definition of $\mathcal{O}(\cdot)$ and other asymptotic notation used throughout the paper.

kriging model to estimate CVaR. Broadie et al. (2015) applies a regression model to estimate Type-(i) nested statistic. With the optimal choice of $M = \Gamma$ and $N = 1$ for the initial experiment, they show that the MSE of the nested statistic estimator converges in $\mathcal{O}(\Gamma^{-1+\delta})$ for any $\delta > 0$ until it reaches an asymptotic bias level. Hong et al. (2017) adopts nonparametric kernel regression and the k-nearest-neighbor estimator to pool inner replications from nearby outer scenarios to estimate each $\mu(\boldsymbol{\theta})$. For Type-(i) nested statistic, their optimal parameter choice leads to MSE convergence rate of $\mathcal{O}(\Gamma^{-\min\{1,4/(d+2)\}})$, where d is the dimension of $\boldsymbol{\theta}$. Wang et al. (2022) considers kernel ridge regression, which results in the MSE convergence rate between $\mathcal{O}(\Gamma^{-1})$ and $\mathcal{O}(\Gamma^{-2/3})$, depending on the smoothness of $\mu(\cdot)$.

A common assumption in aforementioned studies is that given budget Γ, the experimenter can choose the values of M and N. Hence, it is sensible to seek for estimation-error-minimizing M (and N) as a function of Γ. However, the same assumption does not apply to when the M outer scenarios are fixed. Another common feature of the reviewed studies is that an equal number of inner replications are spent at all outer scenarios; Broadie et al. (2011) and Giles and Haji-Ali (2019) are exceptions, however, they study sequential experiment designs whereas we propose a single-stage experiment design. By single stage, we mean that no initial experiment is required to fit a metamodel as in other work.

We identify two key experiment design decisions for our pooling scheme: (1) *At which outer scenarios to run inner simulation and how many replications to run?* (2) *How to pool these replications* when estimating the conditional mean at each outer scenario? The former is referred to as the *sampling decision* and the latter as the *pooling decision*.

We propose a new single-stage experiment design that makes both optimal sampling and pooling decisions via the likelihood ratio (LR) method by solving a bi-level optimization problem. We estimate each $\mu(\boldsymbol{\theta})$ by taking a weighted average of the LR estimators defined between $\boldsymbol{\theta}$ and all M outer scenarios. However, some pairwise LR estimators may have infinite variances. We adopt an LR estimator variant that allows us to (approximately) assess its variance and hence evaluate the efficiency of pooling each pair of the outer scenarios *prior to running any inner replications*.

Utilizing this feature, our bi-level optimization problem has a constraint that requires the pooled estimator of $\mu(\boldsymbol{\theta})$ for each $\boldsymbol{\theta}$ to have variance (approximately) no larger than that of a standard Monte Carlo (MC) estimator with N i.i.d. samples. The upper-level problem makes the sampling decision that minimizes the total simulation budget. The lower-level problem makes the pooling decision by finding the variance-minimizing weights for each $\boldsymbol{\theta}$ given any sampling decision. Unlike other pooled nested simulation designs, ours may result in running no replication at some outer scenarios. We reformulate the bi-level problem into a linear program (LP) with guaranteed feasibility. The total simulation budget for our design is the optimal objective value of the LP.

While our method works for any N that the experimenter provides, a question arises on how to choose N. When M is fixed, even if N is infinity, the nested statistic estimator has nonzero MSE. Hence, we argue

that it is sensible to increase N just enough so that the MSE convergence rate is not slowed down by N. The functional relationship between N and M is difficult to derive for finite M in general. However, by analyzing the convergence rate of the estimator as $M, N \to \infty$, we obtain a guidance on how to choose N relative to M. We show that the MSE convergence rate of the Type-(i) nested statistic estimator computed from our experiment design is $\mathcal{O}(N^{-1}) + \mathcal{O}(M^{-1})$. Similarly, the Type-(ii) statistic (quantile) converges to the true quantile in $\mathcal{O}_p(M^{-1/2}) + \mathcal{O}_p(N^{-1/2})$. Both results suggest $N = \Theta(M)$.

Caution must be taken when interpreting our asymptotic analysis results to avoid the misconception that it is less efficient than SNS, which has an MSE convergence rate of $\mathcal{O}(N^{-2}) + \mathcal{O}(M^{-1})$. Recall that in our design, N is merely a parameter in the bi-level program. This parameter ensures that the pooled LR estimator's approximate variance is no larger than the variance of the standard MC estimator with N i.i.d. replications. Our design does not actually run N replications at the M scenarios, and its simulation budget tends to be much less than MN. Indeed, under some mild conditions, we show that the optimized simulation budget grows in $\mathcal{O}(M)$ when $N = M$ is adopted. This result combined with the MSE convergence rate analyses suggests that the MSE of a type-(i) statistic estimated from our design converges in $\mathcal{O}(\Gamma^{-1})$ given a simulation budget Γ.

Lastly, to place our contribution in the literature, we provide a brief review on two closely related methods: importance sampling (IS) and the LR method. These are mathematically identical but differ in their goals. IS concerns finding a variance-minimizing sampling distribution for a single estimator *prior to* running simulations. While we also consider sampling decisions, our study differ from classical IS (Hesterberg 1988, Owen 2013, Glasserman 2003) because we consider multiple estimators at the same time. The LR method aims to save computation by reusing outputs *after* the simulations are run. This method has been applied to improve efficiency of metamodeling (Dong et al. 2018), gradient estimation (L'Ecuyer 1990, 1993, Glasserman and Xu 2014), and sensitivity analysis and optimization (Rubinstein and Shapiro 1993, Kleijnen and Rubinstein 1996, Fu 2015, Maggiar et al. 2018). The LR method is also applied in the "green simulation" experiment design (Feng and Staum 2015, 2017), which recycles and reuses simulation outputs previously made to save computations and improve precision. Similar to these approaches, our pooling decision is also guided by the LR method once the sampling decision is made.

In recent development of simulation model risk quantification literature, Zhou and Liu (2019) apply the LR method to pool inner replications of the nested simulation with equal weights and demonstrate significant computational savings. Zhang et al. (2022) also propose an LR-based approach that pools inner replications simulated from one common sampling distribution. Neither of these studies optimizes the sampling decision. Moreover, both designs have a risk of making some outer scenarios' conditional mean estimators have large or even unbounded variances. Our optimal experiment design avoids such a case by assigning a small (or zero) pooling weight to an outer scenario pair, if their LR estimator is deemed to have a large variance.

Our preliminary work (Feng and Song 2019) explores application of various kinds of LR estimators to make the simulation risk analysis more efficient. However, we have not considered optimizing the sampling and pooling decisions in the earlier work. Moreover, the theoretical convergence analyses on different types of nested statistics are new and we focus exclusively on the self-normalized LR estimator in the current paper.

The remainder of the paper is organized as follows. Section 2 provides a mathematical framework for the nested simulation problem. Section 3 defines our pooled LR estimator and discusses its properties. In Section 4, we propose the experiment design framework to optimize sampling and pooling decisions. Asymptotic properties of the nested simulation statistics computed from the experiment design are analyzed in Section 5 followed by the analysis on the simulation budget growth rate in Section 6. Empirical evaluations and conclusions are provided in Sections 7 and 8, respectively.

2. Problem Statement

Consider a nested simulation experiment, where the outer scenarios are M i.i.d. multidimensional random vectors, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M$, whose support is $\boldsymbol{\Theta} \subseteq \mathbb{R}^p$. With slight abuse of notation, we treat $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M$ as realized outer scenarios rather than random vectors until Section 5. For any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, let $h(\mathbf{x}; \boldsymbol{\theta})$, $\mu(\boldsymbol{\theta})$, and $\mathrm{V}_{\boldsymbol{\theta}}[g(\mathbf{X})]$ represent the conditional density function, mean, and variance of the input vector $\mathbf{X}$ given $\boldsymbol{\theta}$, respectively. When no ambiguity arises, we also use the shorthand notation $\mu_i = \mu(\boldsymbol{\theta}_i)$ for convenience. The following assumption facilitates our discussion.

ASSUMPTION 1. *For all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $h(\mathbf{x}; \boldsymbol{\theta})$ is a well-defined probability distribution function and has a common support $\mathcal{X} \subseteq \mathbb{R}^d$ for a fixed dimension d. Furthermore, the simulation output function g does not depend on $\boldsymbol{\theta}$ and $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathrm{V}_{\boldsymbol{\theta}}[g(\mathbf{X})] < \infty$.*

The common support, $\mathcal{X}$, ensures that the LR between any two outer scenarios is well-defined. The common g allows us to reuse $\mathbf{X}$ generated from $\boldsymbol{\theta}_j \neq \boldsymbol{\theta}_i$ to estimate μ_i via the LR method. The fixed input dimension, d, is a limitation of our method. We refer the readers to Feng and Song (2019) on applying the LR method to improve nested simulation efficiency when the dimension of $\mathbf{X}$ is random. Lastly, $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathrm{V}_{\boldsymbol{\theta}}[g(\mathbf{X})] < \infty$ ensures that the standard Monte Carlo (MC) estimator for $\mu(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ has finite variance.

Recall that we consider two types of nested statistics. For Type-(i) statistic, we consider three choices for ζ: (1) The indicator function, $\zeta(\mu_i) = I(\mu_i \leq \xi)$ for some fixed $\xi \in \mathbb{R}$; the indicator function can be used to estimate the exceedance probability that a portfolio's loss beyond ξ, that is, $\Pr(\mu_i > \xi)$. (2) The hockey stick function, $\zeta(\mu_i) = \max\{\mu_i - \xi, 0\} = (\mu_i - \xi) I(\mu_i > \xi)$, which is commonly used in ERM applications to price derivatives or compound options (Glasserman 2003). (3) Smooth function ζ with a bounded second derivative. An example of such ζ is the squared loss function given target ξ, $\zeta(\mu_i) = (\mu_i - \xi)^2$.

In SNS with N inner replications for each outer scenario, the MC estimator of μ_i, $\bar{\mu}_i \equiv \sum_{k=1}^{N} g(\mathbf{X}_k)/N$, is computed for each i, where $\mathbf{X}_k \overset{i.i.d.}{\sim} h(\mathbf{x}; \boldsymbol{\theta}_i)$ is the input vector generated within the kth inner replication

made at $\boldsymbol{\theta}_i$. Under Assumption 1, $\mathrm{V}_{\boldsymbol{\theta}_i}[\bar{\mu}_i] = \mathrm{V}_{\boldsymbol{\theta}_i}[g(\mathbf{X})]/N$ for any $\boldsymbol{\theta}_i \in \boldsymbol{\Theta}$. From $\bar{\mu}_1, \ldots, \bar{\mu}_M$, $\mathrm{E}[\zeta(\mu_i)]$ can be estimated by $\sum_{i=1}^{M} \zeta(\bar{\mu}_i)/M$. Also, the α-quantile of μ_i can be estimated by the empirical quantile $\bar{\mu}_{(\lceil M\alpha \rceil)}$, where $\bar{\mu}_{(i)}$ is the ith order statistic of $\bar{\mu}_1, \ldots, \bar{\mu}_M$.

Throughout the paper, for positive sequences $\{a_n\}$ and $\{b_n\}$, $a_n = \mathcal{O}(b_n)$ means that there exists constant $\bar{c} > 0$ such that $a_n \leq \bar{c} b_n$ for all $n \in \mathbb{Z}^+$ and $a_n = o(b_n)$ implies $a_n/b_n \to 0$ as $n \to \infty$. Moreover, $a_n = \Theta(b_n)$ means that there exist constants $\bar{c}, \underline{c} > 0$ such that $\underline{c} b_n \leq a_n \leq \bar{c} b_n$ for all $n \in \mathbb{Z}^+$. For sequence of random variables $\{X_n\}$, we say $X_n = \mathcal{O}_p(b_n)$, if for any $\varepsilon > 0$ there exist $\underline{c} > 0$ and $\bar{n} > 0$ such that $\Pr\{|X_n/b_n| > \underline{c}\} < \varepsilon$ for all $n > \bar{n}$.

3. Conditional Mean Estimator via Likelihood Ratio Method

In this section, we continue to treat $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M$ as realized outer scenarios. We apply the LR method to pool inner replications from several scenarios to estimate the conditional mean at each scenario. Consider two scenarios $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$, where $\boldsymbol{\theta}_i$ is the target scenario whose conditional mean μ_i is desired and $\boldsymbol{\theta}_j$ is the sampling scenario at which inner replications are generated. Then, from the change of measure, we have

$$\mu_i = \mathrm{E}_{\boldsymbol{\theta}_i}[g(\mathbf{X})] = \mathrm{E}_{\boldsymbol{\theta}_j}\left[g(\mathbf{X})\frac{h(\mathbf{X};\boldsymbol{\theta}_i)}{h(\mathbf{X};\boldsymbol{\theta}_j)}\right] = \mathrm{E}_{\boldsymbol{\theta}_j}[g(\mathbf{X})W_{ij}(\mathbf{X})], \tag{1}$$

where $W_{ij}(\mathbf{X}) \equiv \frac{h(\mathbf{X};\boldsymbol{\theta}_i)}{h(\mathbf{X};\boldsymbol{\theta}_j)}$ is the LR of the input vector, $\mathbf{X}$, between $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$. Under Assumption 1, $W_{ij}(\mathbf{X})$ is well-defined and $\mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}(\mathbf{X})] = 1$ for any $\boldsymbol{\theta}_i, \boldsymbol{\theta}_j \in \boldsymbol{\Theta}$. From (1), one can define an unbiased LR estimator of μ_i, $\widehat{\mu}_{ij} \equiv \sum_{k=1}^{N_j} g(\mathbf{X}_k)W_{ij}(\mathbf{X}_k)/N_j$, where N_j is the number of inner replications made at $\boldsymbol{\theta}_j$. We refer to $\widehat{\mu}_{ij}$ as the nominal LR estimator.

A diagnostic measure for assessing efficiency of an LR estimator is the *effective sample size* (ESS), which is defined as the number of replications required for the MC estimator to achieve the same variance as the LR estimator. So, the larger the ESS, the more precise the LR estimator is. Because $\mathrm{V}[\widehat{\mu}_{ij}] = \mathrm{V}_{\boldsymbol{\theta}_j}[g(\mathbf{X})W_{ij}(\mathbf{X})]/N_j$, the ESS of the nominal LR estimator is $\frac{\mathrm{V}_{\boldsymbol{\theta}_j}[g(\mathbf{X})]}{\mathrm{V}_{\boldsymbol{\theta}_j}[g(\mathbf{X})W_{ij}(\mathbf{X})]}N_j$. Due to its dependence on $g(\mathbf{X})$, this ESS cannot be evaluated analytically in general and must be estimated via simulation.

In this study, we adopt the self-normalized LR estimator:

$$\widetilde{\mu}_{ij} = {N_j}^{-1}\sum_{k=1}^{N_j} g(\mathbf{X}_k)\widetilde{W}_{ij,k}, \quad \mathbf{X}_k \overset{i.i.d.}{\sim} h(\mathbf{x};\boldsymbol{\theta}_j), \forall k = 1, \ldots, N_j, \tag{2}$$

where $\widetilde{W}_{ij,k} = \frac{W_{ij}(\mathbf{X}_k)}{\sum_{\ell=1}^{N_j} W_{ij}(\mathbf{X}_\ell)/N_j}$ is the self-normalized LR from the kth replication so that the sample average of $\widetilde{W}_{ij,1}, \widetilde{W}_{ij,2}, \ldots, \widetilde{W}_{ij,N_j}$ becomes 1. Lemma 1 states asymptotic properties of $\widetilde{\mu}_{ij}$. Note that $\mathbf{X}$ is dropped from $W_{ij}(\mathbf{X})$ henceforth for notational convenience.

LEMMA 1. *Consider any given target and sampling scenarios $\boldsymbol{\theta}_i, \boldsymbol{\theta}_j \in \boldsymbol{\Theta}$. Under Assumption 1, (i) $\widetilde{\mu}_{ij}$ converges almost surely to μ_i as N_j increases; and (ii) with additional regularity conditions in Assumption OS.1 of the Online Supplement,*

$$\mathrm{E}_{\boldsymbol{\theta}_j}[\widetilde{\mu}_{ij}] - \mu_i = \frac{-\mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2(g(\mathbf{X})-\mu_i)]}{N_j} + o(N_j^{-1}); \mathrm{V}_{\boldsymbol{\theta}_j}[\widetilde{\mu}_{ij}] = \frac{\mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2(g(\mathbf{X})-\mu_i)^2]}{N_j} + o(N_j^{-1}). \tag{3}$$

Part (i) of Lemma 1 is proved in Theorem 9.2 of Owen (2013). Results similar to Part (ii) can be found in Hesterberg (1988) and Owen (2013).

As shown in Lemma 1, the self-normalized LR estimator is biased but consistent. We adopt the self-normalized LR estimator in our study because it has a convenient ESS approximation. Kong (1992) shows that the variance of $\widetilde{\mu}_{ij}$ can be approximated as

$$\mathrm{V}_{\boldsymbol{\theta}_j}[\widetilde{\mu}_{ij}] \approx \mathrm{V}_{\boldsymbol{\theta}_i}[g(\mathbf{X})]\frac{\mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2]}{N_j}. \tag{4}$$

Then, the approximate ESS of $\widetilde{\mu}_{ij}$ is $n_{ij}^e \equiv N_j/\mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2]$, which is free of $g(\mathbf{X})$ and can be computed analytically in some cases. In Lemma 3 (Section 6), we derive a closed-form expression for $\mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2]$ when $h(\mathbf{x};\boldsymbol{\theta})$ belongs to the exponential family.

In a more detailed derivation, Liu (1996) shows the approximation error of (4) is

$$\mathrm{V}_{\boldsymbol{\theta}_j}[\widetilde{\mu}_{ij}] = \left(\mathrm{V}_{\boldsymbol{\theta}_i}[g(\mathbf{X})]\mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2] + \mathrm{E}_{\boldsymbol{\theta}_i}[(W_{ij}-\mathrm{E}_{\boldsymbol{\theta}_i}[W_{ij}])(g(\mathbf{X})-\mu_i)^2]\right) N_j^{-1} + o(N_j^{-1}), \tag{5}$$

which holds under Assumption OS.1 and comments that $\mathrm{E}_{\boldsymbol{\theta}_i}[(W_{ij}-\mathrm{E}_{\boldsymbol{\theta}_i}[W_{ij}])(g(\mathbf{X})-\mu_i)^2]N_j^{-1}$ omitted from (4) may be small when $g(\mathbf{X})$ is relatively flat, however, when it is large, the approximation error of (4) can be significant. Nevertheless, Liu (1996) recommends using (4) as a rule of thumb. We adopt this recommendation and evaluate efficiency of $\widetilde{\mu}_{ij}$ using n_{ij}^e for each $(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j)$ pair prior to running any inner simulations. In Section 7.1, we empirically demonstrate that Approximation (4) performs well with an ERM example.

Remark 1. Some studies (Martino et al. 2017, Elvira et al. 2018) define the ESS as the number of replications such that the LR estimator's MSE matches $\mathrm{V}[\bar{\mu}_i]$. As seen in Lemma 1, the squared bias diminishes faster than the variance asymptotically, thus we consider matching the variances.

4. Optimal Nested Simulation Experiment Design

Suppose M realized outer scenarios, $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M$, are given. We estimate the conditional means at all M scenarios by pooling the self-normalized LR estimators. We require that, for each $\boldsymbol{\theta}_i$, the pooled estimator's variance is no larger than $\mathrm{V}[\bar{\mu}_i]/N$, which is the variance of the SNS estimator with N inner replications in each scenario; we refer to this requirement as the *precision requirement*. Our goal is to minimize the total number of inner replications run at the M scenarios, that is, the simulation budget. In this section, we assume that the experimenter chooses an appropriate N.

Suppose N_j i.i.d. inner replications are run at each $\boldsymbol{\theta}_j$. Some outer scenarios may have zero replications, i.e., $N_j = 0$. Then, $\widetilde{\mu}_{ij}$ is well-defined for any $(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j), 1\le i,j\le M$, if $N_j > 0$. For each i, consider the following pooled LR estimator of μ_i:

$$\widetilde{\mu}_i \equiv \sum\nolimits_{j=1,N_j>0}^{M} \gamma_{ij}\widetilde{\mu}_{ij}, \quad \sum\nolimits_{j=1,N_j>0}^{M}\gamma_{ij} = 1, \tag{6}$$

where $\gamma_{ij}, j = 1, \ldots, M$, are the pooling weights. In SNS, $\bar{\mu}_i$ only uses the inner replications simulated at $\boldsymbol{\theta}_i$. In contrast, the pooled LR estimator $\widetilde{\mu}_i$ pools all inner replications from all sampling scenarios with weights $\{\gamma_{ij} : j = 1, \ldots, M\}$. As all inner replications are run independently, the variance of the pooled LR estimator is given by $\mathrm{V}[\widetilde{\mu}_i] = \sum_{j=1, N_j>0}^{M} \gamma_{ij}^2 \mathrm{V}_{\boldsymbol{\theta}_j}[\widetilde{\mu}_{ij}]$.

Recall that our precision requirement for each $\boldsymbol{\theta}_i$ is $\mathrm{V}[\widetilde{\mu}_i] \le \mathrm{V}[\bar{\mu}_i]$. From (4), we have

$$\sum_{j=1, N_j>0}^{M} \gamma_{ij}^2 \mathrm{V}_{\boldsymbol{\theta}_j}[\widetilde{\mu}_{ij}] = \mathrm{V}[\widetilde{\mu}_i] \le \mathrm{V}[\bar{\mu}_i] = \frac{\mathrm{V}_{\boldsymbol{\theta}_i}[g(\mathbf{X})]}{N} \overset{(4)}{\Rightarrow} \sum_{j=1, N_j>0}^{M} \frac{\gamma_{ij}^2 \mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2]}{N_j} \le \frac{1}{N}. \tag{7}$$

For given sampling decisions $\{N_j\}$, there may be infinitely many feasible weights $\{\gamma_{ij}\}$ that satisfy the constraints. Among them, it is sensible to choose $\{\gamma_{ij}\}$ that minimizes $\mathrm{V}[\widetilde{\mu}_i]$. From this insight, we formulate the following bi-level optimization problem:

$$\min_{N_j \ge 0, \gamma_{ij}} \quad \sum_{j=1}^{M} N_j \tag{8}$$

$$\text{subject to} \quad \sum_{j=1, N_j>0}^{M} \frac{\gamma_{ij}^2 \mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2]}{N_j} \le \frac{1}{N}, \qquad \forall i = 1, \ldots, M$$

$$\{\gamma_{ij}\} \in \arg\min_{\gamma_{ij}} \left\{ \sum_{j=1, N_j>0}^{M} \frac{\gamma_{ij}^2 \mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2]}{N_j} : \sum_{j=1, N_j>0}^{M} \gamma_{ij} = 1 \right\}, \quad \forall i = 1, \ldots, M \tag{9}$$

A brief overview of bi-level optimization is provided in Section OS.4. The upper-level problem (8) makes the *sampling decision* by finding the optimal $\{N_j\}$. The sampling decision determines not only at which $\boldsymbol{\theta}_j$s to run replications at, but also how many replications to run at each. The lower-level problem (9) defined for each target scenario i makes the *pooling decision* to find $\{\gamma_{ij}\}$ that minimizes the (approximate) variance of $\widetilde{\mu}_i$ given the sampling decision $\{N_j\}$. We ignore integrality constraints for $\{N_j\}$ here. In numerical experiments, we assign $\lceil N_j \rceil$ inner replications to $\boldsymbol{\theta}_j$.

Given the sampling decision $\{N_j\}$, for each i the lower-level problem (9) is a simple quadratic program that can be solved analytically via the Karush-Kuhn-Tucker (KKT) conditions: The Lagrangian function of the ith lower-level problem is $\mathcal{L}(\gamma_{ij}, \lambda) = \sum_{j=1, N_j>0}^{M} \frac{\gamma_{ij}^2 \mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2]}{N_j} + \lambda(1 - \sum_{j=1, N_j>0}^{M} \gamma_{ij})$. The corresponding KKT condition is $\frac{2\gamma_{ij} \mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2]}{N_j} - \lambda = 0$ for all $j = 1, \ldots, M$ such that $N_j > 0$, which implies $\gamma_{ij} = \frac{\lambda N_j}{2\mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2]}$. Hence, the optimal pooling decision, $\gamma_{ij}^\star$, is proportional to $\frac{N_j}{\mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2]}$. Considering $\sum_{j=1, N_j>0}^{M} \gamma_{ij} = 1$, we have $\gamma_{ij}^\star = \frac{N_j / \mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2]}{\sum_{k=1}^{M} N_k / \mathrm{E}_{\boldsymbol{\theta}_k}[W_{ik}^2]}$, for all $j = 1, \ldots, M$ such that $N_j > 0$. Note that we drop the condition, $N_j > 0$ since the same expression produces $\gamma_{ij}^\star = 0$ when $N_j = 0$, as desired. Notice that $\gamma_{ij}^\star = 0$ when $\mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2] = \infty$ even if $N_j > 0$. This is the case when $\boldsymbol{\theta}_i$ does not pool from the inner replications at $\boldsymbol{\theta}_j$ because $\mathrm{V}[\widetilde{\mu}_{ij}]$ is large. Plugging in $\gamma_{ij}^\star$ into (9), the ith lower-level problem's optimal objective value is

$$\sum_{j=1}^{M} \frac{(\gamma_{ij}^\star)^2 \mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2]}{N_j} = \left(\sum_{j=1}^{M} \frac{N_j}{\mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2]} \right)^{-1}. \tag{10}$$

Plugging (10) into the first constraint of the bi-level program, we have:

$$\min_{N_j \geq 0} \ \sum_{j=1}^{M} N_j \quad \text{subject to} \ \sum_{j=1}^{M} \frac{N_j}{\mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2]} \geq N, \qquad \forall i = 1, \ldots, M. \tag{11}$$

Note that this linear program (LP) is always feasible because $\mathrm{E}_{\boldsymbol{\theta}_j}[W_{jj}^2] = 1$, so $\{N_j = N, \forall j = 1, \ldots, M\}$ satisfies all constraints. Also note that the constraints can be written as $\sum_{j=1}^{M} n_{ij}^e \geq N$ for all $i = 1, \ldots, M$. Namely, these constraints require the total ESS of the pooled estimator to be at least N for each target scenario. Provided that $\mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2]$ can be computed a priori, e.g., exponential family, (11) can be solved prior to running any inner replications. Numerical studies in Section 7 show that the optimal objective value of (11) is orders of magnitude smaller than MN, which demonstrates that our experiment design significantly reduces the simulation budget compared to SNS.

Let $\{c_j^\star\}$ be an optimal solution of (11) when $N = 1$. Proposition 1 shows that, for fixed outer scenarios, one can solve (11) for $N = 1$ and obtain the optimal solution for arbitrary N by simply scaling the solution from the former by N. Therefore, even if the target precision, N, is changed *post hoc*, there is no need to resolve (11). Such proportionality is also useful for showing asymptotic properties of the pooled estimator in Section 5.

PROPOSITION 1. *Given $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M$, let $\{c_j\}$ and $\{c_j^\star\}$ be a feasible solution and an optimal solution of* (11)*, respectively, for $N = 1$. Then, $\{N_j = N \cdot c_j\}$ and $\{N_j^\star = N \cdot c_j^\star\}$ are a feasible solution and an optimal solution of* (11)*, respectively, for arbitrary $N > 0$.*

Proof. As $\{c_j\}$ is feasible for (11) when $N = 1$, multiplying both sides of the inequality constraint of (11) by arbitrary N shows that $\{N_j = N \cdot c_j\}$ is a feasible solution of the revised problem. By means of contradiction, suppose that there exists a feasible solution $\{N_j'\}$ of the revised problem such that $\sum_{j=1}^{M} N_j' < \sum_{j=1}^{M} N_j^\star$. Dividing both sides of the constraints by N, it is clear that $\{c_j' = N_j'/N\}$ is a feasible solution of (11) when $N = 1$. However, by construction $\sum_{j=1}^{M} c_j' = (\sum_{j=1}^{M} N_j')/N < (\sum_{j=1}^{M} N_j^\star)/N = \sum_{j=1}^{M} c_j^\star$, which contradicts the premise that $\{c_j^\star\}$ is an optimal solution of (11) when $N = 1$. □

In light of Proposition 1, the optimal pooling weights corresponding to $\{N_j^\star\}$ satisfy

$$\gamma_{ij}^\star = \frac{c_j^\star / \mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2]}{\sum_{k=1}^{M} c_k^\star / \mathrm{E}_{\boldsymbol{\theta}_k}[W_{ik}^2]}, \quad \forall j = 1, \ldots, M, \tag{12}$$

which implies $\{\gamma_{ij}^\star\}$ does not depend on N.

We note that both the optimal pooling decision (12) and the objective value (10) have meaningful interpretations: Recall that the ESS of $\widetilde{\mu}_i$ is approximated by $n_{ij}^e = N_j / E_{\boldsymbol{\theta}_j}[W_{ij}^2]$. For any outer scenario $\boldsymbol{\theta}_i$, (12) can be written as $\gamma_{ij}^\star = n_{ij}^e / (\sum_{k=1}^{M} n_{ik}^e)$. So, given any sampling decision $\{N_j\}$, the optimal way to pool the estimators $\widetilde{\mu}_{ij}$, $j = 1, \ldots, M$, is to weight them proportionally to their ESS. Moreover, plugging $\gamma_{ij}^\star =$

$n_{ij}^e/(\sum_{k=1}^M n_{ik}^e)$ into the first equality of (7), we have $\mathrm{V}[\widetilde{\mu}_i] \approx \mathrm{V}_{\boldsymbol{\theta}_i}[g(\mathbf{X})]/(\sum_{j=1}^M n_{ij}^e)$. This suggests that the ESS of the optimally pooled estimator $\widetilde{\mu}_i$ is equal to the total ESS of the estimators $\widetilde{\mu}_{ij}, j = 1, \ldots, M$.

Once $\{N_j^\star\}$ and $\{\gamma_{ij}^\star\}$ are found, we run inner replications at $\{\boldsymbol{\theta}_j\}$ as prescribed by $\{N_j^\star\}$. For each (i, j) pair such that $\gamma_{ij}^\star > 0$, we compute $\widetilde{\mu}_{ij}$ defined in (2). The optimally pooled conditional mean estimators are computed as $\widetilde{\mu}_i^\star = \sum_{j=1}^M \gamma_{ij}^\star \widetilde{\mu}_{ij}$, for $i = 1, \ldots, M$. Then, the Type-(i) nested statistic can be estimated by $\widetilde{\zeta} = \sum_{i=1}^M \zeta(\widetilde{\mu}_i^\star)/M$. The α-quantile of μ_i is estimated by the empirical quantile, $\widetilde{\mu}^\star_{(\lceil M\alpha \rceil)}$, computed from $\widetilde{\mu}_1, \ldots, \widetilde{\mu}_M$.

5. Asymptotic Analysis

In this section, we treat $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M$ as random vectors, not realizations, and establish asymptotic properties of the estimated nested statistics, $\widetilde{\zeta}$ and $\widetilde{\mu}^\star_{(\lceil M\alpha \rceil)}$, computed from the optimized design. These analyses provide a guidance for selecting N as a function of M. Prior to presenting the main results, we first establish convergence results for $\widetilde{\mu}_i^\star$.

Since feasibility of (11) is always guaranteed, $\widetilde{\mu}_i^\star$ is well-defined for any N and M. The following theorem states a strong consistency result for $\widetilde{\mu}_i^\star$.

THEOREM 1. *Suppose Assumption 1 holds. Given* $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M$, $\widetilde{\mu}_i^\star \xrightarrow{a.s.} \mu_i$ *as* $N \to \infty$ *for* $i = 1, \ldots, M$.

Proof. Recall that from Proposition 1, $\{\gamma_{ij}^\star\}$ do not depend on N and are constants once $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M$ are given. From the definition of $\widetilde{\mu}_i^\star$,

$$\lim_{N\to\infty} \widetilde{\mu}_i^\star = \lim_{N\to\infty} \sum\nolimits_{j=1}^M \gamma_{ij}^\star \widetilde{\mu}_{ij} = \sum\nolimits_{j=1}^M \gamma_{ij}^\star \lim_{N_j\to\infty} \widetilde{\mu}_{ij} \xrightarrow{a.s.} \sum\nolimits_{j=1}^M \gamma_{ij}^\star \mu_i = \mu_i \sum\nolimits_{j=1}^M \gamma_{ij}^\star = \mu_i,$$

where the second equality holds because $N_j^\star \propto N$ from Proposition 1 and $\gamma_{ij}^\star = 0$ whenever $N_j^\star = 0$. The almost sure convergence holds from Lemma 1. □

Next, we examine the MSE convergence rate of $\widetilde{\mu}_i^\star$. Recall that Lemma 1 shows both bias and variance of $\widetilde{\mu}_{ij}$ are $\mathcal{O}(N_j^{-1})$ under Assumption OS.1. Because $N_j^\star \propto N$, to show $\mathrm{MSE}[\widetilde{\mu}_i^\star] = \mathcal{O}(N^{-1})$, it suffices to have the result of Lemma 1 hold for each (i, j) pair such that $\gamma_{ij}^\star > 0$. Assumption 2 below formally states the sufficient condition.

ASSUMPTION 2. *For each* $\boldsymbol{\theta}_i \in \boldsymbol{\Theta}$, *let* $\boldsymbol{\Theta}_i \equiv \{\boldsymbol{\theta}_j \in \boldsymbol{\Theta} | \mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2] < \infty\}$. *Then, (i) there exists* $C > 0$ *such that for any* $N_j > C, \boldsymbol{\theta}_i \in \boldsymbol{\Theta}$ *and* $\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_i$, $N_j |\mathrm{E}_{\boldsymbol{\theta}_j}[\widetilde{\mu}_{ij}] - \mu_i| < |\mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2(g(\mathbf{X}) - \mu_i)]| + 1$ *and* $N_j \mathrm{V}_{\boldsymbol{\theta}_j}[\widetilde{\mu}_{ij}] < \mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2(g(\mathbf{X}) - \mu_i)^2] + 1$*; and (ii)* $\sup_{\boldsymbol{\theta}_i \in \boldsymbol{\Theta}} \sup_{\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_i} \mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2(g(\mathbf{X}) - \mu_i)] < \infty$ *and* $\sup_{\boldsymbol{\theta}_i \in \boldsymbol{\Theta}} \sup_{\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_i} \left|\mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2(g(\mathbf{X}) - \mu_i)^2]\right| < \infty$.

Assumption 2.(i) ensures the conclusions of Lemma 1 hold for any $\boldsymbol{\theta}_i, \boldsymbol{\theta}_j \in \boldsymbol{\Theta}$, while (ii) assures the moments in the constants to be uniformly bounded in $\boldsymbol{\Theta}$.

Additionally, we make a minor modification to the definition of $\{N_j^\star\}$:

$$N_j^\star = \begin{cases} \delta N, & \text{if } 0 < c_j^\star < \delta, \\ c_j^\star N, & \text{otherwise,} \end{cases} \tag{13}$$

where δ is a small positive constant. In words, (13) guarantees that if any replications are made at $\boldsymbol{\theta}_j$, then $N_j^\star$ is to be at least δ fraction of N. The number of outer scenarios we sample at does not increase because $N_j^\star = 0$ if $c_j^\star = 0$. We emphasize that (13) has no practical impact as δ can be chosen to be arbitrarily small in the experiments. The following theorem establishes that for any sample of M outer scenarios, $\mathrm{MSE}[\widetilde{\mu}_i^\star] = \mathcal{O}(N^{-1})$.

THEOREM 2. *Suppose Assumptions 1 and 2 hold. Then, for any finite M*

$$\sup_{\{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M\}\subset\boldsymbol{\Theta}} |\mathrm{E}[\widetilde{\mu}_i^\star|\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M] - \mu_i| = \mathcal{O}(N^{-1}) \text{ and}$$
$$\sup_{\{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M\}\subset\boldsymbol{\Theta}} \mathrm{V}[\widetilde{\mu}_i^\star|\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M] = \mathcal{O}(N^{-1})$$

as $N \to \infty$. Moreover, the same statement holds when $M \to \infty$.

Proof. By construction, $\widetilde{\mu}_i^\star$ only pools replications at $\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_i$. For sufficiently large N,

$$|\mathrm{E}[\widetilde{\mu}_i^\star|\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M] - \mu_i| \le \sum_{j=1}^M \gamma_{ij}^\star \left|\mathrm{E}_{\boldsymbol{\theta}_j}[\widetilde{\mu}_{ij}] - \mu_i\right| \overset{(*)}{<} \sum_{j=1,N_j^\star>0}^M \frac{\gamma_{ij}^\star}{N_j^\star}\left(|\mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2(g(\mathbf{X})-\mu_i)]|+1\right)$$
$$\overset{(**)}{\le} \sum_{j=1}^M \frac{\gamma_{ij}^\star}{\delta N} \sup_{\boldsymbol{\theta}_j\in\boldsymbol{\Theta}_i}\left(|\mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2(g(\mathbf{X})-\mu_i)]|+1\right) \overset{(***)}{=} \frac{1}{\delta N}\sup_{\boldsymbol{\theta}_j\in\boldsymbol{\Theta}_i}\left(|\mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2(g(\mathbf{X})-\mu_i)]|+1\right),$$

where $(*)$, $(**)$, and $(***)$ follow from Assumption 2, (13), and $\sum_{j=1,N_j^\star>0}^M \gamma_{ij}^\star = 1$ respectively. Because $\sup_{\boldsymbol{\theta}_i\in\boldsymbol{\Theta}}\sup_{\boldsymbol{\theta}_j\in\boldsymbol{\Theta}_i}\left(|\mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2(g(\mathbf{X})-\mu_i)]|+1\right)$ is bounded from Assumption 2, we conclude $\sup_{\{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M\}\in\boldsymbol{\Theta}} |\mathrm{E}[\widetilde{\mu}_i^\star|\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M] - \mu_i| = \mathcal{O}(N^{-1})$ for finite M as well as when $M \to \infty$. For the variance, as all inner replications are independently simulated,

$$\mathrm{V}[\widetilde{\mu}_i^\star|\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M] = \sum_{j=1}^M (\gamma_{ij}^\star)^2 \mathrm{V}_{\boldsymbol{\theta}_j}[\widetilde{\mu}_{ij}] \overset{(*)}{<} \sum_{j=1,N_j^\star>0}^M \frac{(\gamma_{ij}^\star)^2}{N_j^\star}\left(\mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2(g(\mathbf{X})-\mu_i)^2]+1\right)$$
$$\overset{(**)}{\le} \sum_{j=1}^M \frac{\gamma_{ij}^\star}{\delta N}\sup_{\boldsymbol{\theta}_j\in\boldsymbol{\Theta}_i}\left(\mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2(g(\mathbf{X})-\mu_i)^2]+1\right) = \frac{1}{\delta N}\sup_{\boldsymbol{\theta}_j\in\boldsymbol{\Theta}_i}\left(\mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2(g(\mathbf{X})-\mu_i)^2]+1\right)$$

for sufficiently large N, where $(*)$ follows from Assumption 2 and (13) and $(**)$ holds since $0 \le \gamma_{ij}^\star \le 1$. Because $\sup_{\boldsymbol{\theta}_i \in \boldsymbol{\Theta}} \sup_{\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_i} \left(\mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2 (g(\mathbf{X}) - \mu_i)^2] + 1 \right) < \infty$ from Assumption 2, we conclude $\sup_{\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\} \in \boldsymbol{\Theta}} \mathrm{V}[\widetilde{\mu}_i^\star | \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M] = \mathcal{O}(N^{-1})$ for finite M and when $M \to \infty$. □

In Sections 5.1–5.4, we analyze asymptotic properties of Type-(i) and Type-(ii) nested statistics. For $\widetilde{\zeta} = \sum_{i=1}^M \zeta(\widetilde{\mu}_i^\star)/M$, we show that its bias and variance converge in $\mathcal{O}(N^{-1})$ and $\mathcal{O}(N^{-1}) + \mathcal{O}(M^{-1})$, respectively, when ζ is an indicator, hockey stick, or smooth function with bounded second derivative. Sections 5.1–5.3 present different assumptions and proofs for each choice of ζ. In Section 5.4, we show that the empirical quantile $\widetilde{\mu}^\star_{(\lceil M\alpha \rceil)}$ converges to the true α-quantile of $\mu(\boldsymbol{\theta})$ in $\mathcal{O}_p(M^{-1/2}) + \mathcal{O}_p(N^{-1/2})$.

5.1. Indicator function of the conditional mean

Suppose $\zeta(\mu_i) = I(\mu_i \le \xi)$ for some $\xi \in \mathbb{R}$. Let Φ be the cumulative distribution function (cdf) of μ_i. By definition, $\mathrm{E}[\zeta(\mu_i)] = \Phi(\xi)$. Thus, we denote the corresponding estimator $\widetilde{\zeta} \equiv M^{-1} \sum_{i=1}^M I(\widetilde{\mu}_i^\star \le \xi)$ by $\Phi_{M,N}(\xi)$, where $\Phi_{M,N}(\cdot)$ is the empirical cdf (ecdf) constructed from $\widetilde{\mu}_1^\star, \ldots, \widetilde{\mu}_M^\star$.

For ease of exposition, let $\epsilon_i \equiv \sqrt{N}(\widetilde{\mu}_i^\star - \mu_i)$, which is the scaled estimation error of $\widetilde{\mu}_i^\star$ so that its limiting distribution is not degenerate as $N \to \infty$. From Theorem 2, $\mathrm{E}[\epsilon_i | \boldsymbol{\theta}_i] = \mathrm{E}[\mathrm{E}[\epsilon_i | \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M] | \boldsymbol{\theta}_i] = \mathcal{O}(N^{-1/2})$ uniformly for all $\boldsymbol{\theta}_i \in \boldsymbol{\Theta}$. Similarly, $\mathrm{V}[\epsilon_i | \boldsymbol{\theta}_i] = \mathcal{O}(1)$ uniformly for all $\boldsymbol{\theta}_i \in \boldsymbol{\Theta}$. In the following, we denote the joint distribution of μ_i and ϵ_i by $f_i(\mu, \epsilon)$, where $\boldsymbol{\theta}_i$ is an arbitrary scenario in $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$. Similarly, $f_{ij}(\mu_i, \mu_j, \epsilon_i, \epsilon_j)$ refers to the joint distribution of μ_i, μ_j, ϵ_i, and ϵ_j for some arbitrary $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ among $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$. We make Assumption 3 below to facilitate Theorem 3 that stipulates the MSE convergence rate of $\Phi_{M,N}(\xi)$.

ASSUMPTION 3. *The cdf of μ_i, Φ, is absolutely continuous with continuous probability density function (pdf) ϕ. For any $\boldsymbol{\theta}_i \in \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$, $f_i(\mu, \epsilon)$, is differentiable with respect to μ for each M and N. Moreover, there exist $p_{s,M,N}(\epsilon), s = 0, 1$, such that $f_i(\mu, \epsilon) \le p_{0,M,N}(\epsilon)$ and $\left| \frac{\partial f_i(\mu,\epsilon)}{\partial \mu} \right| \le p_{1,M,N}(\epsilon)$ for all μ and for each M and N, and $\sup_M \sup_N \int_{-\infty}^{\infty} |\epsilon|^k p_{s,M,N}(\epsilon) d\epsilon < \infty$ for $s = 0, 1$ and $0 \le k \le 2$. Also, for any $\boldsymbol{\theta}_i \ne \boldsymbol{\theta}_j$, $f_{ij}(\mu_i, \mu_j, \epsilon_i, \epsilon_j)$, is differentiable with respect to μ_i and μ_j for each M and N. There exist $p_{s,M,N}(\epsilon_i, \epsilon_j), s = 0, 1$, such that $f_{i,j}(\mu_i, \mu_j, \epsilon_i, \epsilon_j) \le p_{0,M,N}(\epsilon_i, \epsilon_j)$ and $\left| \frac{\partial f_{i,j}(\mu_i,\mu_j,\epsilon_i,\epsilon_j)}{\partial \mu} \right| \le p_{1,M,N}(\epsilon_i, \epsilon_j)$ for all $\mu_i, \mu_j, i \ne j$ and for each M and N, and $\sup_M \sup_N \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |\epsilon_i|^{k_i} |\epsilon_j|^{k_j} p_{s,M,N}(\epsilon_i, \epsilon_j) d\epsilon_i d\epsilon_j < \infty$ for $s = 0, 1$, and $0 \le k_i, k_j \le 2$ such that $k_i + k_j \le 3$.*

Assumption 3 is likely to hold when M and N are large and $\rho(\mu(\boldsymbol{\theta})) = \mathrm{E}[\zeta(\mu(\boldsymbol{\theta}))]$ is well-estimated via SNS. Indeed, Assumption 3 is identical to Assumption 1 in Gordy and Juneja (2010) made to show the MSE convergence rate for the SNS estimator of $\rho(\mu(\boldsymbol{\theta}))$ except that i) we define ϵ_i as the scaled estimation error of the pooled LR estimator, $\tilde{\mu}_i^\star$, not the standard MC estimator, $\bar{\mu}_i$; and ii) we impose the boundedness and moment conditions for the joint distribution of $(\mu_i, \mu_j, \epsilon_i, \epsilon_j)$, not just for that of (μ_i, ϵ_i). The latter is because $\tilde{\mu}_i^\star$ pools simulation outputs from $\boldsymbol{\theta}_j \ne \boldsymbol{\theta}_i$. Recall that $\tilde{\mu}_i^\star$ is the weighted average of all $\tilde{\mu}_{ij}$'s such that $N_j^\star > 0$ and $\gamma_{ij}^\star > 0$, where each $\tilde{\mu}_{ij}$ is a sample average of $N_j^\star$ i.i.d. observations of $W_{ij} g(\mathbf{X})$.

Therefore, for sufficiently large N, $\epsilon_i = \sqrt{N}(\tilde{\mu}_i^\star - \mu_i)$ behaves like a normal random variable as in SNS. Note that if $\tilde{\mu}_i^\star$ is replaced by a generic pooled LR estimator $\tilde{\mu}_i$ in (6) (e.g., constructed with suboptimal weights γ_{ij}), then ϵ_i does not converge to a normal random variable in general since some $\tilde{\mu}_{ij}$ may have unbounded variance. For optimized $\tilde{\mu}_i^\star$, the variance result in Theorem 2 guarantees that a central limit theorem can be applied to ϵ_i. On the other hand, if SNS does not estimate $\rho(\mu(\boldsymbol{\theta}))$ well, then there is little hope that Assumption 3 holds.

THEOREM 3. *Under Assumptions 1–3,* $\mathrm{E}[\Phi_{M,N}(\xi)] - \Phi(\xi) = \mathcal{O}(N^{-1})$ *and* $\mathrm{V}[\Phi_{M,N}(\xi)] = \mathcal{O}(N^{-1}) + \mathcal{O}(M^{-1})$.

Proof. Let us define Φ_M as the ecdf constructed from $\mu_1, \ldots, \mu_M$ given the same outer scenarios as $\Phi_{M,N}$. Because Φ_M is an unbiased estimator of Φ, $\mathrm{E}[\Phi_{M,N}(\xi)] - \Phi(\xi) = \mathrm{E}[\Phi_{M,N}(\xi) - \Phi_M(\xi)]$. From definitions, $\Phi_{M,N}(\xi) - \Phi_M(\xi) = M^{-1}\sum_{i=1}^M (I(\widetilde{\mu}_i^\star \le \xi) - I(\mu_i \le \xi))$. For each i,

$$\begin{aligned}\mathrm{E}[I(\widetilde{\mu}_i^\star \le \xi) - I(\mu_i \le \xi)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\xi-\frac{\epsilon}{\sqrt{N}}} f_i(\mu,\epsilon)d\mu d\epsilon - \int_{-\infty}^{\infty}\int_{-\infty}^{\xi} f_i(\mu,\epsilon)d\mu d\epsilon \\ &= -\int_{-\infty}^{\infty}\int_{\xi-\epsilon/\sqrt{N}}^{\xi} f_i(\mu,\epsilon)d\mu d\epsilon. \end{aligned} \tag{14}$$

The first-order Taylor series expansion of $f_i(\mu,\epsilon)$ at $\mu \in [\xi - \epsilon/\sqrt{N}, \xi]$ is

$$f_i(\mu,\epsilon) = f_i(\xi,\epsilon) + \tfrac{\partial f_i(\breve{\mu},\epsilon)}{\partial \mu}(\mu - \xi), \tag{15}$$

where $\breve{\mu} \in (\mu, \xi)$. From Assumption 3, for all M and N,

$$\frac{\epsilon}{\sqrt{N}} f_i(\xi,\epsilon) - \frac{\epsilon^2}{2N} p_{1,M,N}(\epsilon) \le \int_{\xi-\epsilon/\sqrt{N}}^{\xi} f_i(\mu,\epsilon)d\mu \le \frac{\epsilon}{\sqrt{N}} f_i(\xi,\epsilon) + \frac{\epsilon^2}{2N} p_{1,M,N}(\epsilon). \tag{16}$$

Thus, from (14), integrating all three sides of (16) with respect to $\epsilon \in (-\infty, \infty)$ gives upper and lower bounds to $\mathrm{E}[I(\widetilde{\mu}_i^\star \le \xi) - I(\mu_i \le \xi)]$. Note that $f_i(\xi,\epsilon) = f_i(\epsilon|\xi)\phi(\xi)$, where $f_i(\epsilon|\xi)$ is the conditional pdf of ϵ_i given $\mu_i = \xi$. Thus, $\int_{-\infty}^{\infty} f_i(\xi,\epsilon)\frac{\epsilon}{\sqrt{N}}d\epsilon = \phi(\xi)\mathrm{E}\left[\frac{\epsilon_i}{\sqrt{N}}\middle|\mu_i = \xi\right] = \phi(\xi)\mathrm{E}[\widetilde{\mu}_i^\star - \mu_i|\mu_i = \xi]$, where $\mathrm{E}[\widetilde{\mu}_i^\star - \mu_i|\mu_i] = \mathcal{O}(N^{-1})$ uniformly for all μ_i from Theorem 2. Also, Assumption 3 guarantees that $\int_{-\infty}^{\infty}\epsilon^2 p_{1,M,N}(\epsilon)d\epsilon$ is bounded. Therefore, $\mathrm{E}[I(\widetilde{\mu}_i^\star \le \xi) - I(\mu_i \le \xi)] = \mathcal{O}(N^{-1})$ for each i, which in turn implies $\mathrm{E}[\Phi_{M,N}(\xi) - \Phi_M(\xi)] = \mathcal{O}(N^{-1})$. Next, noticing $\Phi_{M,N}(\xi) = \Phi_{M,N}(\xi) - \Phi_M(\xi) + \Phi_M(\xi)$, $\mathrm{V}[\Phi_{M,N}(\xi)]$ can be written as

$$\begin{aligned}\mathrm{V}[\Phi_{M,N}(\xi)] &= \mathrm{V}[\Phi_{M,N}(\xi) - \Phi_M(\xi)] + \mathrm{V}[\Phi_M(\xi)] + 2\mathrm{Cov}[\Phi_{M,N}(\xi) - \Phi_M(\xi), \Phi_M(\xi)] \\ &\le 2\mathrm{V}[\Phi_{M,N}(\xi) - \Phi_M(\xi)] + 2\mathrm{V}[\Phi_M(\xi)] \end{aligned} \tag{17}$$

Clearly, $\mathrm{V}[\Phi_M(\xi)] = \mathcal{O}(M^{-1})$. In the following, we show $\mathrm{V}[\Phi_{M,N}(\xi) - \Phi_M(\xi)] = \mathcal{O}(M^{-1}) + \mathcal{O}(N^{-1})$ by first expanding it as

$$\frac{1}{M^2}\sum_{i=1}^{M}\mathrm{V}[I(\widetilde{\mu}_i^\star \le \xi) - I(\mu_i \le \xi)] + \frac{1}{M^2}\sum_{i=1}^{M}\sum_{\substack{j=1\\ j\neq i}}^{M}\mathrm{Cov}[I(\widetilde{\mu}_i^\star \le \xi) - I(\mu_i \le \xi), I(\widetilde{\mu}_j^\star \le \xi) - I(\mu_j \le \xi)] \tag{18}$$

and showing that the leading order of (18) is $\mathcal{O}(N^{-1})$. First, $\mathrm{V}[I(\widetilde{\mu}_i^\star \le \xi) - I(\mu_i \le \xi)] = \mathrm{E}[(I(\widetilde{\mu}_i^\star \le \xi) - I(\mu_i \le \xi))^2] - \mathrm{E}[I(\widetilde{\mu}_i^\star \le \xi) - I(\mu_i \le \xi)]^2$, where the latter term is $\mathcal{O}(N^{-2})$ from the bias derivation above. Because $(I(\widetilde{\mu}_i^\star \le \xi) - I(\mu_i \le \xi))^2 = 1$, if and only if, $\widetilde{\mu}_i^\star \le \xi$ and $\mu_i > \xi$, or $\widetilde{\mu}_i^\star > \xi$ and $\mu_i \le \xi$,

$$\mathrm{E}[(I(\widetilde{\mu}_i^\star \le \xi) - I(\mu_i \le \xi))^2] = \int_{-\infty}^{0}\int_{\xi}^{\xi-\epsilon/\sqrt{N}} f_i(\mu,\epsilon)d\mu d\epsilon + \int_{0}^{\infty}\int_{\xi-\epsilon/\sqrt{N}}^{\xi} f_i(\mu,\epsilon)d\mu d\epsilon.$$

Following the same logic in (16), both inner integrals are bounded from above by $\frac{\epsilon}{\sqrt{N}}f_i(\xi,\epsilon) + \frac{\epsilon^2}{2N}p_{1,M,N}(\epsilon)$. Hence,

$$\mathrm{E}[(I(\widetilde{\mu}_i^\star \le \xi) - I(\mu_i \le \xi))^2] \le \int_{-\infty}^{\infty}\left\{\frac{\epsilon}{\sqrt{N}}f_i(\xi,\epsilon) + \frac{\epsilon^2}{2N}p_{1,M,N}(\epsilon)\right\}d\epsilon = \mathcal{O}(N^{-1}).$$

This concludes that the first summation of (18) is $\mathcal{O}((NM)^{-1})$.

The covariance term of (18) can be rewritten as

$$\mathrm{E}\left[(I(\widetilde{\mu}_i^\star \le \xi) - I(\mu_i \le \xi))\ \left(I(\widetilde{\mu}_j^\star \le \xi) - I(\mu_j \le \xi)\right)\right] - \mathrm{E}\left[I(\widetilde{\mu}_i^\star \le \xi) - I(\mu_i \le \xi)\right]\mathrm{E}\left[I(\widetilde{\mu}_j^\star \le \xi) - I(\mu_j \le \xi)\right]. \tag{19}$$

The first expectation of (19) is equal to

$$\begin{aligned}&\Pr\{\widetilde{\mu}_i^\star \le \xi, \widetilde{\mu}_j^\star \le \xi\} - \Pr\{\mu_i \le \xi, \widetilde{\mu}_j^\star \le \xi\} + \Pr\{\mu_i \le \xi, \mu_j \le \xi\} - \Pr\{\widetilde{\mu}_i^\star \le \xi, \mu_j \le \xi\}\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{\xi-\epsilon_i/\sqrt{N}}^{\xi}\int_{\xi-\epsilon_j/\sqrt{N}}^{\xi} f_{ij}(\mu_i,\mu_j,\epsilon_i,\epsilon_j)d\mu_j d\mu_i d\epsilon_j d\epsilon_i.\end{aligned}$$

Applying the first-order Taylor series expansion to $f_{ij}(\mu_i,\mu_j,\epsilon_i,\epsilon_j)$ gives

$$f_{ij}(\mu_i,\mu_j,\epsilon_i,\epsilon_j) = f_{ij}(\xi,\xi,\epsilon_i,\epsilon_j) + \frac{\partial f_{ij}(\check{\mu}_i,\check{\mu}_j,\epsilon_i,\epsilon_j)}{\partial \mu_i}(\mu_i - \xi) + \frac{\partial f_{ij}(\check{\mu}_i,\check{\mu}_j,\epsilon_i,\epsilon_j)}{\partial \mu_j}(\mu_j - \xi) \tag{20}$$

for some $\check{\mu}_i \in (\mu_i,\xi)$ and $\check{\mu}_j \in (\mu_j,\xi)$. Under Assumption 3, the integral of (20) with respect to $\mu_i \in [\xi - \epsilon_i/\sqrt{N},\xi]$ and $\mu_j \in [\xi - \epsilon_j/\sqrt{N},\xi]$ is lower/upper-bounded by

$$\frac{\epsilon_i\epsilon_j}{N}f_{ij}(\xi,\xi,\epsilon_i,\epsilon_j) \mp \frac{|\epsilon_i^2\epsilon_j + \epsilon_j^2\epsilon_i|}{2N^{3/2}}p_{1,M,N}(\epsilon_i,\epsilon_j), \tag{21}$$

Integrating (21) once again with respect to $\epsilon_i, \epsilon_j \in (-\infty, \infty)$, we have $\mathrm{E}[(I(\widetilde{\mu}_i^\star \le \xi) - I(\mu_i \le \xi))\left(I(\widetilde{\mu}_j^\star \le \xi) - I(\mu_j \le \xi)\right)] = \mathcal{O}(N^{-1})$. Since it is already shown that $\mathrm{E}\big[I(\widetilde{\mu}_i^\star \le \xi) - I(\mu_i \le \xi)\big] = \mathcal{O}(N^{-1})$, we conclude $\mathrm{Cov}[I(\widetilde{\mu}_i^\star \le \xi) - I(\mu_i \le \xi), I(\widetilde{\mu}_j^\star \le \xi) - I(\mu_j \le \xi)] = \mathcal{O}(N^{-1})$ from (19), which in turn implies $\mathrm{V}[\Phi_{M,N}(\xi) - \Phi_M(\xi)] = \mathcal{O}(M^{-1}) + \mathcal{O}(N^{-1})$ from (17). □

5.2. Hockey stick function of the conditional mean

Next, we analyze the case when ζ is a hockey stick function, i.e., $\zeta(\mu_i) = \max\{\mu - \xi, 0\} = (\mu - \xi)I(\mu_i > \xi)$ for some $\xi \in \mathbb{R}$, which requires the following additional moment conditions.

ASSUMPTION 4. *For $p_{1,N,M}(\epsilon)$ in Assumption 3, $\sup_M \sup_N \int_{-\infty}^{\infty} |\epsilon|^3 p_{1,M,N}(\epsilon) d\epsilon < \infty$. Similarly, for $p_{s,N,M}(\epsilon_i, \epsilon_j)$ defined in Assumption 3,*

$$\sup_M \sup_N \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |\epsilon_i|^{k_i} |\epsilon_j|^{k_j} p_{s,M,N}(\epsilon_i, \epsilon_j) d\epsilon_i d\epsilon_j < \infty$$

for $s = 0, 1$, $0 \le k_i, k_j \le 3$ and $k_i + k_j \le 5$. Also, for each M and N, there exist functions $q_{s,M,N}(\mu_i, \epsilon_i, \epsilon_j), s = 0, 1$, such that $f_{ij}(\mu_i, \mu_j, \epsilon_i, \epsilon_j) \le q_{0,M,N}(\mu_i, \epsilon_i, \epsilon_j)$ and $\left|\frac{\partial f_{ij}(\mu_i, \mu_j, \epsilon_i, \epsilon_j)}{\partial \mu_j}\right| \le q_{1,M,N}(\mu_i, \epsilon_i, \epsilon_j)$ for all μ_i, μ_j, ϵ_i, and ϵ_j. Lastly, for $s = 0, 1$,

$$\sup_M \sup_N \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{\xi}^{\infty} |\mu_i|^{k_m} |\epsilon_i|^{k_i} |\epsilon_j|^{k_j} q_{s,M,N}(\mu_i, \epsilon_i, \epsilon_j) d\mu_i d\epsilon_i d\epsilon_j < \infty$$

for any $\xi \in \mathbb{R}, 0 \le k_m \le 1$, and $0 \le k_i, k_j \le 3$ such that $k_i + k_j \le 4$.

Similar to Assumption 3, Assumption 4 is an extension of Assumption 1 in Gordy and Juneja (2010) to analyze the hockey stick function, ζ, and is likely to hold when SNS estimates $\mathrm{E}[\zeta(\mu(\theta))]$ well. Theorem 4 below shows that the hockey stick function results in the same bias and variance convergence rates as the those of the indicator function. Due to the space limit, we present the proof of Theorem 4 in Section OS.5.

THEOREM 4. *Suppose $\zeta(\mu_i) = (\mu_i - \xi)I(\mu_i > \xi)$ for some constant $\xi \in \mathbb{R}$ and $\widetilde{\zeta} = \sum_{i=1}^{M} \zeta(\widetilde{\mu}_i^\star)/M$. Under Assumptions 1–4, $\mathrm{E}[\widetilde{\zeta} - \zeta(\mu_i)] = \mathcal{O}(N^{-1})$ and $\mathrm{V}[\widetilde{\zeta}] = \mathcal{O}(M^{-1}) + \mathcal{O}(N^{-1})$.*

5.3. Smooth function of the conditional mean

To analyze the case when ζ is a smooth function of μ_i, we make the following assumption.

ASSUMPTION 5. *The continuous function, $\zeta : \mathbb{R} \to \mathbb{R}$, is twice differentiable everywhere with bounded second derivative ζ''. Also, $\mathrm{E}[(\zeta(\mu_i))^2], \mathrm{E}[(\zeta'(\mu_i)\epsilon_i)^2]$ and $\mathrm{E}[\epsilon_i^4]$ are bounded.*

Similar to the indicator and the hockey stick functions, the following theorem shows that the MSE convergence rate of $\widetilde{\zeta}$ is $\mathcal{O}(M^{-1}) + \mathcal{O}(N^{-1})$ for ζ that satisfies Assumption 5.

THEOREM 5. *Suppose ζ satisfies Assumption 5 and $\widetilde{\zeta} = \sum_{i=1}^{M} \zeta(\widetilde{\mu}_i^\star)/M$. Under Assumptions 1 and 2,* $\mathrm{E}[\widetilde{\zeta} - \zeta(\mu_i)] = \mathcal{O}(N^{-1})$ *and* $\mathrm{V}[\widetilde{\zeta}] = \mathcal{O}(M^{-1}) + \mathcal{O}(N^{-1})$.

The proof of Theorem 5 can be found in Section OS.6.

5.4. Quantile of the conditional mean

Let $\widetilde{q}_\alpha = \widetilde{\mu}^\star_{(\lceil M\alpha \rceil)}$. In this section, we prove weak consistency of $\widetilde{q}_\alpha$ as M and N increase. Recall that in SNS, q_α is estimated by the $\lceil M\alpha \rceil$th order statistic of M independent conditional mean estimators. In our design, $\widetilde{q}_\alpha$ is an order statistic of the correlated estimators, $\widetilde{\mu}_1^\star, \ldots, \widetilde{\mu}_M^\star$. Consistency of an empirical quantile estimator constructed from dependent outputs has been studied (Sen 1972, Heidelberger and Lewis 1984) under the assumption that the output sequence has a strong mixing property, which ensures that pairwise correlation between distant outputs in the sequence vanishes. Our pooled LR estimators do not have this property. Instead, their pairwise correlation decreases as N increases.

To show weak consistency of $\widetilde{q}_\alpha$, we need the following intermediate result. Lemma 2 states that $\Phi_{M,N}(\cdot)$ is uniformly weakly consistent to $\Phi(\cdot)$; see Section OS.3 for its proof. Theorem 6 is the main result of this section.

LEMMA 2. *Under Assumptions 1–3,* $\sup_{\xi\in\mathbb{R}} |\Phi_{M,N}(\xi) - \Phi(\xi)| = \mathcal{O}_p(M^{-1/2}) + \mathcal{O}_p(N^{-1/2})$.

THEOREM 6. *Suppose Assumptions 1–3 hold and $\phi(q_\alpha) > 0$ for given $0 < \alpha < 1$. Then,* $|\widetilde{q}_\alpha - q_\alpha| = \mathcal{O}_p(M^{-1/2}) + \mathcal{O}_p(N^{-1/2})$.

Proof. For each M, $|\Phi_{M,N}(\widetilde{q}_\alpha) - \alpha| \le 1/M$. Also, Lemma 2 implies $|\Phi_{M,N}(\widetilde{q}_\alpha) - \Phi(\widetilde{q}_\alpha)| \le \sup_{\xi\in\mathbb{R}} |\Phi_{M,N}(\xi) - \Phi(\xi)| = \mathcal{O}_p(M^{-1/2}) + \mathcal{O}_p(N^{-1/2})$. Therefore, $|\Phi(\widetilde{q}_\alpha) - \alpha| = \mathcal{O}_p(M^{-1/2}) + \mathcal{O}_p(N^{-1/2})$. Then, for sufficiently large M and N, there exists $U^\star \in (\Phi(\widetilde{q}_\alpha), \alpha)$ such that $\phi(\Phi^{-1}(U^\star)) > 0$ and $\Phi^{-1}(\Phi(\widetilde{q}_\alpha)) = \Phi^{-1}(\alpha) + \frac{1}{\phi(\Phi^{-1}(U^\star))}(\Phi(\widetilde{q}_\alpha) - \alpha)$ with probability arbitrarily close to 1. Because ϕ is bounded in a neighborhood of q_α, $|\Phi^{-1}(\Phi(\widetilde{q}_\alpha)) - \Phi^{-1}(\alpha)| = |\widetilde{q}_\alpha - q_\alpha| = \mathcal{O}_p(M^{-1/2}) + \mathcal{O}_p(N^{-1/2})$. □

6. Growth Rate of Simulation Budget in the Optimal Design

Section 5 suggests that selecting $N = \Theta(M)$ leads to the most efficient experiment design. In general, it is difficult to derive an analytical expression for the minimized simulation budget as a function of M and N. In this section, we study how quickly the optimal budget grows in M asymptotically when we set $N = M$ and the input distribution of the inner simulation belongs to an exponential family.

DEFINITION 1. A probability distribution function, $h(\mathbf{x};\boldsymbol{\theta})$, is said to belong to an *exponential family* in the canonical form, if it can be written as

$$h(\mathbf{x};\boldsymbol{\theta}) = B(\mathbf{x}) \exp(\boldsymbol{\theta}^\top T(\mathbf{x}) - A(\boldsymbol{\theta})), \tag{22}$$

where $B(\mathbf{x})$, $T(\mathbf{x})$, $\boldsymbol{\theta}$, and $A(\boldsymbol{\theta}) \equiv \ln\left(\int_{\mathcal{X}} B(\mathbf{x}) \exp(\boldsymbol{\theta}^\top T(\mathbf{x})) d\mathbf{x}\right)$ are referred to as the base measure, sufficient statistic, natural parameter, and log-partition function, respectively. The natural parameter space is

defined as $\bar{\boldsymbol{\Theta}} \equiv \left\{\boldsymbol{\theta} : \int_{\mathcal{X}} B(\mathbf{x}) \exp(\boldsymbol{\theta}^\top T(\mathbf{x})) d\mathbf{x} < \infty\right\}$. An exponential family is called *regular* if $\bar{\boldsymbol{\Theta}}$ is a non-empty open set in $\mathbb{R}^p$.

LEMMA 3. *Suppose $h(\mathbf{x};\boldsymbol{\theta})$ belongs to an exponential family with natural parameter space $\bar{\boldsymbol{\Theta}}$. Then, for any $\boldsymbol{\theta}_i, \boldsymbol{\theta}_j \in \bar{\boldsymbol{\Theta}}$, $\mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2] = \exp(A(\boldsymbol{\theta}_j) - 2A(\boldsymbol{\theta}_i) + A(2\boldsymbol{\theta}_i - \boldsymbol{\theta}_j))$ if $2\boldsymbol{\theta}_i - \boldsymbol{\theta}_j \in \bar{\boldsymbol{\Theta}}$ or ∞ otherwise.*

Proof. Substituting (22) into the definition of likelihood ratio W_{ij} we have

$$\begin{aligned}\mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2] &= \int_{\mathcal{X}} \left(\frac{h(\mathbf{x};\boldsymbol{\theta}_i)}{h(\mathbf{x};\boldsymbol{\theta}_j)}\right)^2 h(\mathbf{x};\boldsymbol{\theta}_j) d\mathbf{x} = \exp(A(\boldsymbol{\theta}_j) - 2A(\boldsymbol{\theta}_i)) \int_{\mathcal{X}} B(\mathbf{x}) \exp((2\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)^\top T(\mathbf{x})) d\mathbf{x} \\ &= \exp(A(\boldsymbol{\theta}_j) - 2A(\boldsymbol{\theta}_i) + A(2\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)) \int_{\mathcal{X}} \underbrace{B(x) \exp((2\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)^\top T(\mathbf{x}) - A(2\boldsymbol{\theta}_i - \boldsymbol{\theta}_j))}_{=h(\mathbf{x};2\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)} d\mathbf{x}.\end{aligned}$$

Note that $\int_{\mathcal{X}} h(\mathbf{x}; 2\boldsymbol{\theta}_i - \boldsymbol{\theta}_j) d\mathbf{x}$ equals one, if $2\boldsymbol{\theta}_i - \boldsymbol{\theta}_j \in \bar{\boldsymbol{\Theta}}$, and diverges, otherwise. □

Lemma 3 implies that $(\mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2])^{-1} = 0$, if $2\boldsymbol{\theta}_i - \boldsymbol{\theta}_j \notin \bar{\boldsymbol{\Theta}}$. In this case, the optimal design does not pool replications from $h(\mathbf{x};\boldsymbol{\theta}_j)$ to estimate μ_i as $n_{ij}^e = 0$.

ASSUMPTION 6. *The following holds for $\boldsymbol{\Theta}$ and $h(\mathbf{x};\boldsymbol{\theta})$:*

(i) $h(\mathbf{x};\boldsymbol{\theta})$ belongs to a regular exponential family with natural parameter space $\bar{\boldsymbol{\Theta}}$.

(ii) $\boldsymbol{\Theta}$ is a closed and bounded (or, equivalently, compact) subset of $\bar{\boldsymbol{\Theta}}$.

Assumption 6.(i) includes many common choices of input models such as exponential, normal, and log-normal distributions. Assumption 6.(ii) is not restrictive as the bounds for $\boldsymbol{\Theta}$ can be chosen arbitrarily large. Under Assumption 6, we first show the following lemma before stating the main theorem of this section.

LEMMA 4. *Suppose Assumption 6 holds. Then, there exists finite set $\boldsymbol{\Theta}_{M^*} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{M^*}\}$ such that for each $\theta_i \in \boldsymbol{\Theta}$, $n_{ij}^e > 1/2$ for some $\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_{M^*}$.*

Proof. From (22), the moment generating function of the sufficient statistic, $T(\mathbf{x})$, of $h(\mathbf{x};\boldsymbol{\theta})$ can be derived as $M_T(\boldsymbol{u}) = \mathrm{E}[e^{\boldsymbol{u}^\top T(\mathbf{x})}] = e^{A(\boldsymbol{\theta}+\boldsymbol{u}) - A(\boldsymbol{\theta})}$ for $\boldsymbol{u} \in \mathbb{R}^p$. Because $\bar{\boldsymbol{\Theta}}$ is a non-empty open set, $M_T(\mathbf{u})$ is finite for any $\boldsymbol{\theta} \in \bar{\boldsymbol{\Theta}}$ and sufficiently small $\boldsymbol{u}$. This implies that $\mathrm{E}_{\boldsymbol{\theta}}[T(\mathbf{x})] < \infty$ for any $\boldsymbol{\theta} \in \bar{\boldsymbol{\Theta}}$. From Definition 1, we have

$$\nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta}) = \frac{\nabla_{\boldsymbol{\theta}} \left(\int_{\mathcal{X}} B(\mathbf{x}) \exp(\boldsymbol{\theta}^\top T(\mathbf{x})) d\mathbf{x}\right)}{\int_{\mathcal{X}} B(\mathbf{x}) \exp(\boldsymbol{\theta}^\top T(\mathbf{x})) d\mathbf{x}} = \int_{\mathcal{X}} T(\mathbf{x}) B(\mathbf{x}) \exp(\boldsymbol{\theta}^\top T(\mathbf{x}) - A(\boldsymbol{\theta})) d\mathbf{x} = \mathrm{E}_{\boldsymbol{\theta}}[T(\mathbf{x})] < \infty.$$

Therefore, $A(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta} \in \bar{\boldsymbol{\Theta}}$ as its gradient exists everywhere in $\bar{\boldsymbol{\Theta}}$.

Consider any $\boldsymbol{\theta}_i, \boldsymbol{\theta}_j \in \boldsymbol{\Theta}$. By Lemma 3 we have $\mathrm{E}_{\boldsymbol{\theta}_j}[W_{jj}^2] = 1$ and that $\mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2]$ is continuous in $\boldsymbol{\theta}_i$ because $A(\cdot)$ is continuous. Then, there exists open neighborhood $\mathcal{N}(\boldsymbol{\theta}_j)$ of $\boldsymbol{\theta}_j$ such that $\mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2] < 2$ for all $\boldsymbol{\theta}_i \in \mathcal{N}(\boldsymbol{\theta}_j)$. Equivalently, $n_{ij}^e = 1/\mathrm{E}_{\boldsymbol{\theta}_j}[W_{ij}^2] > 1/2$ for all $\boldsymbol{\theta}_i \in \mathcal{N}(\boldsymbol{\theta}_j)$. We can construct $\mathcal{N}(\boldsymbol{\theta}_j)$ for each $\boldsymbol{\theta}_j \in \boldsymbol{\Theta}$ and the collection of them is an open cover of $\boldsymbol{\Theta}$. Then, there exists finite subcover $\boldsymbol{\Theta}_{M^*} =$

$\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{M^*}\}$ of the open cover by the compactness of $\boldsymbol{\Theta}$. Consequently, for any $\theta_i \in \boldsymbol{\Theta}$, there exists $\theta_j \in \boldsymbol{\Theta}_{M^*}$ such that $\boldsymbol{\theta}_i \in \mathcal{N}(\boldsymbol{\theta}_j)$ and thus, the conclusion follows. □

Lemma 4 implies that by simulating $2M$ inner replications at all outer scenarios in $\boldsymbol{\Theta}_{M^*}$, we can guarantee that all $\boldsymbol{\theta}_i \in \boldsymbol{\Theta}$ have ESS no less than M. Building upon this insight, Theorem 7 shows the main result of this section.

THEOREM 7. *Suppose Assumption 6 holds. Furthermore, for $M > M^*$, suppose we choose the first M^* outer scenarios to be those in $\boldsymbol{\Theta}_{M^*}$ and let $N = M$. Let Γ_M be the optimal objective function value of* (11) *with $M = N$. Then, $\Gamma_M = \mathcal{O}(M)$.*

Proof. As discussed, by Lemma 4, $N_j = 2M$, $j = 1, \ldots, M^*$, is a feasible solution to (11) with the resulting objective function value of $2MM^*$. Therefore, $\Gamma_M \leq 2MM^*$ for all $M > M^*$, which implies $\Gamma_M = \mathcal{O}(M)$. □

Recall that the MSE convergence rates of the nested statistic estimators studied in Sections 5.1–5.3 are $\mathcal{O}(M^{-1})$. Therefore, Theorem 7 stipulates that the MSEs converge in $\mathcal{O}(\Gamma^{-1})$ given simulation budget Γ when Assumption 6 holds.

The effect of fixing the first M^* outer scenarios as described in Theorem 7 is negligible as M increases. In Section 7.3, we empirically observe that the minimized simulation budget indeed grows linearly in M when $N = M$ even if all outer scenarios are chosen at random.

In the proof for Theorem 7, we postulate the existence of $\boldsymbol{\Theta}_{M^*}$. In Section OS.7, we provide two examples to illustrate how $\boldsymbol{\Theta}_{M^*}$ can be constructed.

7. Numerical Studies

In this section, we present three numerical examples to demonstrate the performance of the proposed optimal design. The following experiment designs are compared:

- *Optimal design*: Given the outer scenarios, the optimal sampling and pooling decisions are obtained by solving (11) with $N = M$. Let Γ denote the minimized simulation budget in our optimal design. We emphasize again that N is only a benchmark number of inner replications for setting the ESS constraints in (11). The optimal design does not run N inner replications in each scenario.
- *SNS*$^+$: Standard nested simulation design with the same M outer scenarios as the optimal design and $N = M$, which results in the total budget of $MN = M^2$. SNS$^+$ sets a benchmark for the precision requirement for the optimal design as the latter shows the same MSE convergence rate as the standard nested simulation when N is set to M.
- *SNS*: Standard nested simulation design that consumes the same simulation budget Γ as the optimal design. Following the recommendations in Gordy and Juneja (2010), we set $M = \lceil \Gamma^{2/3} \rceil$ and $N = \lceil \Gamma^{1/3} \rceil$, respectively. The SNS design's simulation budget may be slightly higher than the optimal design's due to the ceiling function.

- *Regression*: Pooling based on a linear regression model; as recommended by Broadie et al. (2015), we sample Γ initial design points and run one replication at each to fit the model. We chose different basis functions to suit each example. The fitted model is evaluated at the same M outer scenarios used in the optimal design to compute the performance measures.

The first two examples are ERM problems whose objectives are to evaluate the risk of the future portfolio values due to the price fluctuation of the underlying assets. The first example has two options written on one underlying asset while the second example considers a portfolio with 300 options written on 10 assets. The third example demonstrates Bayesian input model uncertainty quantification applied to a multi-product newsvendor problem. The experiment designs are evaluated by the coverage probabilities and widths of the credible intervals (CrIs) of the expected profit.

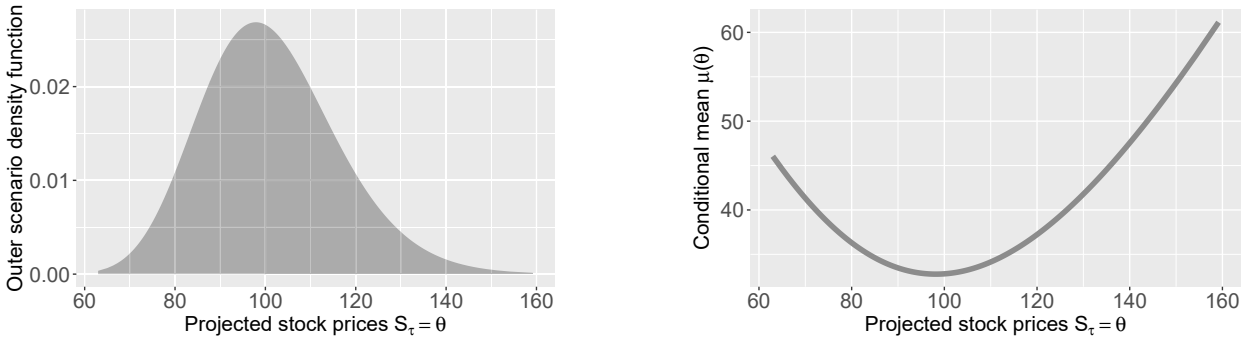
The expression for $\mu(\boldsymbol{\theta})$ is known in all three examples, which facilitates performance evaluation of the experiment designs in comparison. All source codes are available at `https://github.com/INFORMSJoC/2022.0392/` (Feng and Song 2024).

7.1. Single-Asset Enterprise Risk Management (ERM) Problem

In this section, we consider a straddle option portfolio that consists of a call option and a put option with the same underlying stock, strike, and maturity. Let the underlying stock price at $t = 0$ be $S_0 = \$100$. We assume that the stock is non-dividend-paying and follows the Black-Scholes model with $\eta = 5\%$ annualized expected rate of return and $\sigma = 30\%$ annualized volatility. The annualized risk-free rate is $r = 2\%$. The common maturity of both options is $T = 2$ years and the common strike price is $K = \$110$.

We are interested in the value of the portfolio in $\tau = 1/4$ year from now. The future portfolio value can be evaluated via nested simulation by first simulating the stock price at τ, S_τ, then computing the expected payoff of the stock at the maturity given S_τ by running inner replications. Here, the outer scenario is one-dimensional $\theta = S_\tau$, and $\mu(\theta)$ corresponds to the conditional expected payoff given S_τ. From the Black-Scholes model, the outer scenarios are simulated under the real-world measure, i.e., $S_\tau | S_0 = S_0 e^{Z_\tau}$, where the rate of return Z_τ, is distributed as $\mathcal{N}((\eta - \frac{1}{2}\sigma^2)\tau, \sigma^2\tau)$. Thus, $\theta = S_\tau$ has a log-normal distribution whose density function is shown in Figure 1a. In this example, we choose M equally-spaced quantiles of the log-normal distribution as the outer scenarios.

Given any scenario $\theta = S_\tau$, the input random variable for the inner replication is the stock price at maturity, $X = S_T$. From the Black-Scholes model, the inner simulation is conducted under the risk-neutral measure, i.e., $S_T | S_\tau = S_\tau e^{Z_T}$, where $Z_T \sim \mathcal{N}((r - \frac{1}{2}\sigma^2)(T - \tau), \sigma^2(T - \tau))$. The simulation model g computes the discounted payoff of the straddle option from X; $g(X) = e^{-r(T-t)}[(K - X)_+ + (X - K)_+]$, where $(x)_+ = \max\{x, 0\}$. Thus, $\mu(\theta) = \mathrm{E}_\theta[g(X)] = \mathrm{E}[g(X)|S_\tau]$. The analytical expression for $\mu(\theta)$ can be derived from the Black-Scholes model without simulation. Figure 1b depicts $\mu(\theta)$, which shows that the portfolio value is high when S_τ takes extreme values. We consider a financial institution that offers the



(a) pdf of outer scenarios $S_\tau = \theta$.

(b) Conditional expectation $\mu(\theta)$.

Figure 1 Problem settings 1D ERM example: (a) shows the pdf of θ and (b) plots $\mu(\theta)$ against θ.

straddle strategies to investors, i.e., a short position. So the company suffers large losses when $\mu(\theta)$ is large, or when S_τ takes extreme values. For the regression approach, the weighted Laguerre polynomials up to order 2 are adopted as the basis functions, which is a common choice in pricing American options (Longstaff and Schwartz 2001).

Firstly, we consider a fixed number of outer scenarios, i.e., $M = 1{,}000$ equally-spaced quantiles from the outer scenario's log-normal distribution, and compare the estimation accuracies for different nested simulation designs. We set $N = M = 1{,}000$ when solving (11) for the optimal design. The resulting simulation budget in the optimal design is $\Gamma = 2{,}148$ replications. The optimal sampling decision, $\{c_j^\star\}$, allocates 29%, 21%, 21%, and 29% of the simulation budget Γ to only 4 scenarios, i.e., $\theta = 70.63, 71.01, 141.18,$ and 141.94. These points are near the two tail ends of distribution of θ, which can be explained by the ESS formula. In this example, the inner simulation random variable $X = S_T|S_\tau$ for any outer scenario $\theta = S_\tau$ follows a log-normal distribution with the common variance $\sigma^2(T-\tau)$ with mean $m_\theta = \ln\theta + (r - \frac{1}{2}\sigma^2)(T-\tau)$. Thus, for any θ_i and θ_j, the LR is $W_{ij}(X) = \exp\left(\frac{(\ln X - m_{\theta_j})^2 - (\ln X - m_{\theta_i})^2}{2\sigma^2(T-\tau)}\right)$ and its second moment is $\mathrm{E}_{\theta_j}[W_{ij}^2] = \exp\left(\frac{(\ln\theta_i - \ln\theta_j)^2}{\sigma^2}\right)$. Consequently, the ESS of using one sample from θ_j to estimate the conditional mean for scenario θ_i is inversely proportional to $\exp((\ln\theta_i - \ln\theta_j)^2)$, which indicates that the ESS falls off quickly when $\theta_i \neq \theta_j$. Therefore, the θs on the tails benefit the most by pooling from nearby θs, whereas the θs in the middle can achieve the desired ESS by pooling from both tails.

For each of the $1{,}000$ θ's, we run $10{,}000$ macro replications for all three simulation designs to estimate the corresponding $\mu(\theta)$. For each simulation design, the 2.5% and 97.5% quantiles of the estimated $\mu(\theta)$ from these $10{,}000$ macro runs at all θ's are connected to form a 95% confidence band. Three resulting confidence bands, one for each simulation design, are depicted in Figure 2. Note that SNS is omitted from Figure 2 as its confidence band is too wide to be compared in the same plot. Figure 2b is a zoomed in version of Figure 2a near the mode of the outer scenario distribution. The black curve in these figures shows the true option portfolio value $\mu(\theta)$ calculated from closed-form option pricing formulas in the Black-Scholes

model. We see in Figure 2a that the optimal design's confidence band is almost indistinguishable from that of the SNS$^+$ design. This demonstrates that the precision requirement (7) in our optimization formulation is effective despite approximation (4). The small approximation error (5) shows up in the zoomed Figure 2b, where we see the optimal design's confidence band is slightly wider than the SNS$^+$ design's. We note that the optimal design achieves similar precision as the SNS$^+$ design with significant smaller simulation budget: The optimal design's simulation budget $2{,}148$ is 465 times smaller than that of the SNS$^+$ design ($MN = 10^6$). We also see in Figure 2a that the regression-based design's confidence band is significantly wider than the optimal design's and the SNS$^+$ design's when $\theta = S_\tau$ takes extreme values. In general, users do not know the relationship between $\mu(\theta)$ and θ prior to simulation. So, particularly low precision in any region could significantly impact tail risk measure estimation thus is undesirable.

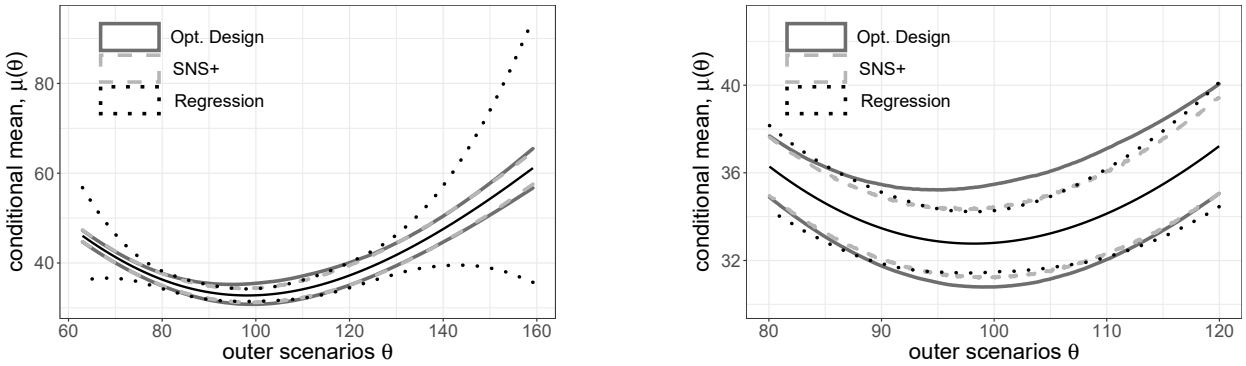


Figure 2 (a) shows the 95% confidence bands for different experiment designs based on 10,000 macro runs; (b) is a zoomed-in version of (a) around the median of θ.

Next, we consider different numbers of outer scenarios M and compare the MSEs of the portfolio risk measures computed from all four designs. We set confidence level $\alpha = 0.99$ to emulate tail risk estimation. The four nested statistics we consider are: (i) the α-quantile of $\mu(\theta)$, (ii) indicator risk measure $\mathrm{E}[I(\mu(\theta) > 49)]$, (iii) hockey stick risk measure $\mathrm{E}[(\mu(\theta) - 49)_+]$, and (iv) squared tail risk measure $\mathrm{E}[(\mu(\theta) - 49)_+^2]$. The latter three measures emulate probability of large losses, expected excess loss, and expected squared excess loss (similar to semivariance, but measures variability in the tail), respectively. All are common in practical ERM problems. These statistics cannot be calculated analytically, thus were estimated via MC simulation by sampling 10^8 θs and computing $\mu(\theta)$ at each θ from its analytical expression, which are then used to compute the four nested statistics of interest. Based on these 10^8 conditional means, the 99%-quantile is $\mu_\alpha \approx 48.916$, which is why 49 is chosen as the threshold in the other nested statistics.

Table 1 presents the MSE of the nested statistics computed from the four experiment designs by running $10{,}000$ macro runs. Excluding when $M = 128$, the optimal design's MSEs for all nested statistics are lower

Table 1 Single-asset ERM example: MSEs of the nested simulation statistics computed from $1{,}000$ macro runs of the four experiment designs. All methods use the same simulation budget except for SNS$^+$.

M	Quantile				Indicator Function $\zeta(\cdot)$			
	Opt. Design	SNS	SNS$^+$	Regression	Opt. Design	SNS	SNS$^+$	Regression
128	23	259	6.52	102	8.38E-04	8.77E-03	4.43E-05	5.66E-04
512	5.45	146	1.31	37.6	1.21E-04	3.58E-03	4.84E-06	1.27E-04
1024	2.52	109	0.503	20	2.54E-05	2.06E-03	1.71E-06	6.72E-05
2048	1.16	81.2	0.195	10.2	9.01E-06	1.29E-03	6.38E-07	3.64E-05
4096	0.517	52.9	0.059	4.93	2.40E-06	6.66E-04	2.46E-07	1.88E-05
M	Hockey Stick Function $\zeta(\cdot)$				Square Function $\zeta(\cdot)$			
	Opt. Design	SNS	SNS$^+$	Regression	Opt. Design	SNS	SNS$^+$	Regression
128	5.34E-02	8.89E-01	1.38E-03	9.01E-02	22	541	0.229	126
512	2.29E-03	2.28E-01	2.80E-04	1.53E-02	0.185	81	0.108	10.6
1024	4.81E-04	1.09E-01	1.06E-04	6.64E-03	0.0875	30.3	0.0585	3.29
2048	2.26E-04	5.97E-02	3.87E-05	3.15E-03	0.0456	14.1	0.0281	1.25
4096	8.51E-05	2.73E-02	1.41E-05	1.58E-03	0.0216	5.5	0.0125	0.543

than those of SNS and the regression approach. Moreover, the optimal design's MSEs are within the same orders of magnitude as those of SNS+ even though the latter requires a much larger simulation budget; the optimal design's simulation budget is about 247 and $1{,}760$ times smaller than that of SNS$^+$ when $M = 512$ and $M = 4{,}096$, respectively. Further, for $M \geq 512$, the MSE of each risk measure computed from the optimal design also shrinks by approximately a half when M increases by a factor of two, which is consistent with our asymptotic results in Section 5.

However, when $M = 128$, we observe small-sample performance degradation of the optimal design; the MSE of the probability (indicator ζ) is larger for the optimal design than the regression method. Also, when M shrinks to 128 from 512, the MSE of the estimated nested statistics increase by more than 4 times; notably, 119 times in the square function case. Since the performance guarantee for the optimal design (and all other benchmarks) is shown asymptotically, such degradation for small M is not surprising. Nevertheless, because the optimal design is formulated to minimize the simulation budget, when M is small and thus the support of the outer scenario is not covered well by the size-M sample, then the optimal design may not be as effective as when M is large.

7.2. Multi-Assets Enterprise Risk Management (ERM) Problem

In this section, we consider a more realistic ERM problem for an option portfolio consists of 10 underlying assets and 300 options. The underlying assets are 10 non-dividend-paying stocks that follow the multidimensional Black-Scholes model and their respective expected annualized returns are $5\%, 6\%, \ldots, 14\%$, and the respective expected annualized volatilities are $30\%, 32\%, \ldots, 48\%$. The correlations among stock prices are randomly generated using the R library, `randcorr`. All stocks have an initial price $S_0 = \$100$ at time 0 and the risk-free rate is 2%. The following 30 options are written on each of the 10 stocks:

- 10 European options: 5 puts with strike prices $\{85, 90, 95, 100, 105\}$ and 5 calls with strike prices $\{110, 115, 120, 125, 130\}$.

- 10 fixed-strike geometric Asian options: 5 puts with strike prices $\{85, 90, 95, 100, 105\}$ and 5 calls with strike prices $\{110, 115, 120, 125, 130\}$.
- 10 down-and-out put barrier options with strike prices $\{85, 90, \ldots, 130\}$ and lower barriers $\{75, 80, \ldots, 120\}$. The same type of barrier options are studied in Broadie et al. (2015) to showcase the performances of the regression-based experiment design.

All options have maturity $T = 2$; all time units are in years. Stock prices are simulated at discrete time steps, i.e., $h = 1/52$ (approx. weekly), and are used to calculate the average prices in Asian options' payoffs. Also, conditioning on the stock prices at two consecutive times, the minimum price between the two times is simulated via the Brownian bridge (Glasserman 2003, Chapter 6.4). Nested simulation is applied to estimate the portfolio's values at time $\tau = 4/52$ (approx. one month) under different scenarios. The Asian and barrier options are partial-time options, i.e., the Asian options' average prices are calculated between τ and T and the barrier options can be knocked out only between τ and T. These settings are similar to those in the numerical examples in Broadie et al. (2015).

The portfolio value can be decomposed into 10 groups of 3 options written on the same stock. Then, one can estimate the value of each group using only the simulated prices of the corresponding underlying stock. This decomposition scheme is inspired by Section 5 in Hong et al. (2017) and is implemented for all experiment designs we compare.

Compared to the single-asset example in Section 7.1, this multi-asset example has more stocks, options, and time steps in both outer- and inner-level simulations. The proposed optimal nested simulation design can be adapted to these additional complexities:

1. Exploiting the decomposition of the portfolio, the optimal design can be applied for each underlying stock separately, allowing different simulation budget for each.
2. As shown in Section OS.8, the LRs are calculated using the one-step transition densities $S_{\tau+h}|S_\tau$ for different scenarios S_τ. The ESS is also calculated from them.
3. The asymptotic properties shown in Section 5 are not affected by the portfolio decomposition. That is, if they apply to each group of options written on the same stock, they apply to the portfolio value.

We run $K = 1{,}000$ macro runs to assess the performances of different experiment designs. For SNS$^+$, we simulate $M = 1{,}000$ outer scenarios and $N = 1{,}000$ inner replications at each scenario. Unlike in Section 7.1, the outer scenarios in different macro runs vary, but the same scenarios are used in the optimal design and SNS$^+$ design in each macro run. For SNS, we set M and N to be $\lceil \Gamma^{2/3} \rceil$ and $\lceil \Gamma^{1/3} \rceil$, respectively, where Γ is the average optimal simulation budget for the 10 stocks. For the regression-based design, for each stock, we set the number of outer scenarios to be the average of the optimal design's simulation budgets for the 10 stocks, then simulate one inner replication per scenario. We fit the portfolio value to the weighted Laguerre polynomials up to degree 3 of the 10 underlying stocks' prices, which has 31 explanatory variables including the intercept. The fitted regression model is used to estimate $\mu(\theta)$ for the same $1{,}000$ outer scenarios as those in the optimal design and SNS$^+$ design.

Table 2 Multi-asset ERM example: Average number of outer scenarios, simulation budgets, and CPU time (in seconds) per macro run. The third row shows the number of outer scenarios at which inner replications are made. Speedup factors relative to SNS$^+$ are shown in parentheses.

	SNS$^+$	Optimal Design	SNS	Regression
Simulation Budget	10^6	8200.8 (122x↓)	8525.7 (117x↓)	8200.8 (122x↓)
CPU time (seconds)	997	25.5 (39.16x↓)	28.2 (35.42x↓)	296 (3.37x↓)
# Outer Scenarios	10^3	21.9 (46x↓)	407.3 (2.5x↓)	8200.8 (8.2x↑)

Table 3 Multi-asset ERM example: MSEs of nested simulation statistics computed from 1,000 macro runs of the four experiment designs. Numbers in parentheses are the MSE inflation factors compared to the SNS$^+$ design.

	SNS$^+$	Opt. Design	SNS	Regression
AMSE(μ)	763	740 (0.97x)	36628 (48x)	3052 (4.1x)
MSE.Quantile	622	764 (1.2x)	100637 (162x)	11298 (18x)
MSE.Indicator $\zeta(\cdot)$	7.23E-06	8.51E-06 (1.18x)	9.22E-03(1275x)	1.44E-04 (20x)
MSE.Hockey Stick $\zeta(\cdot)$	4.81E-02	5.36E-02 (1.12x)	144 (3004x)	2.4 (50x)
MSE.Square $\zeta(\cdot)$	1933	1993 (1.03x)	7845487 (4060x)	231,413 (120x)

Experiments were run on Intel Xeon Gold 6334 8-core 3.6 GHz (Ice Lake) CPUs. Table 2 shows the average numbers of outer scenarios, simulation budgets, CPU time, and the corresponding speedup relative to SNS$^+$. Notice that the optimal design's simulation budget is over 120 times smaller than SNS$^+$'s. The SNS has slightly larger simulation budget due to rounding. The CPU time for the optimal design is nearly $1/40$ of the SNS's, which includes the time for solving the LP, LR calculation, and simulation. For this example, the optimal design's CPU time is evenly split between solving the LP (13.5 seconds), and simulation and LR calculation (12 seconds). The SNS's CPU time is exclusively for simulation and more than double the simulation and LR calculation time for the optimal design.

To compare the estimation error of all experiment designs, we compute the average MSE (AMSE) of the conditional means in addition to the MSEs of the nested statistics.

$$\text{AMSE}(\mu) \equiv \frac{1}{K}\sum_{k=1}^{K}\frac{1}{M}\sum_{m=1}^{M}(\hat{\mu}_{m,k}-\mu_{m,k})^2,$$

where $\hat{\mu}_{m,k}$ is the estimated conditional mean in the mth scenario of the kth macro run and $\mu_{m,k}$ is the corresponding true conditional mean.

Table 3 summarizes the AMSE and MSEs for different experiment designs. Notice that the optimal design's AMSE and MSEs closely match those of SNS$^+$. This demonstrates that the ESS constraints in (7) work as intended. Considering the computational cost comparison in Table 2, the optimal design is extremely efficient. While SNS has a similar CPU time as the optimal design, its AMSE and MSEs are orders of magnitudes larger than those of the optimal design. The regression has higher AMSE and MSEs than both SNS$^+$ and the optimal design; although the AMSE is only mildly (4.1x) higher, the MSEs of the

nested statistics are orders of magnitude higher. This implies that the regression-based method has reasonable precision within the central region of the outer scenarios but performs poorly at outer scenarios with extreme conditional means. This observation is consistent with that made in Section 7.1.

In summary, the multi-asset ERM example demonstrates the applicability of the optimal design in practice and that it outperforms not only SNS but also the regression approach.

7.3. Input Uncertainty Quantification for Multi-Product Newsvendor Problem

In this section, we consider a single-stage newsvendor problem with ten products. We assume that the ℓth product's demand, X_ℓ, follows a Poisson distribution. All product demands are independent. For the ℓth product, let c_ℓ and p_ℓ be the unit cost and sale price, respectively. The stocking policy, $\{k_1, \ldots, k_{10}\}$, is a vector representing the stocking levels of the ten products. We chose $p_\ell = 7 + 3\ell$, $c_\ell = 2$, and $k_\ell = 9 + \ell$ for all ℓ for the experiment. Given these inputs, the simulator computes the total profit, $g(\boldsymbol{X}) = \sum_{\ell=1}^{10} \{p_\ell \min(X_\ell, k_\ell) - c_\ell k_\ell\}$.

Unknown to us, the mean demand of the ℓth product is $\vartheta_\ell^c = 5 + \ell$. We have $50 + 5\ell$ i.i.d. realizations from Poisson(ϑ_ℓ^c) to estimate ϑ_ℓ^c for the simulation study. Taking the Bayesian view, we model the unknown parameter ϑ_ℓ as a random variable with a prior distribution and update it with the observations from Poisson(ϑ_ℓ^c). To exploit conjugacy, the Gamma prior with rate 0.001 and shape 0.001 is adopted for each ϑ_ℓ. Then, the posterior distribution of ϑ_ℓ is still Gamma with rate $0.001 + 50 + 5\ell$ and shape 0.001 plus the sum of observed demands of the ℓth product. Let $\boldsymbol{\theta} = \{\vartheta_1, \ldots, \vartheta_{10}\}$ be a parameter vector sampled from the joint posterior distribution. The expected profit given $\boldsymbol{\theta}$, $\mu(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}[g(\boldsymbol{X})]$, is a random variable whose distribution is induced by the posterior of $\boldsymbol{\theta}$.

To measure uncertainty in the model, we construct a $1 - \alpha$ credible interval (CrI) for $\mu(\boldsymbol{\theta})$ via nested simulation. The analytical expression for $\mu(\boldsymbol{\theta})$ can be derived easily using the Poisson distribution function. Thus, a CrI can be constructed by sampling $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M$ from the posterior of $\boldsymbol{\theta}$ and computing the empirical $\alpha/2$ and $1 - \alpha/2$ quantiles from $\mu(\boldsymbol{\theta}_1), \mu(\boldsymbol{\theta}_2), \ldots, \mu(\boldsymbol{\theta}_M)$; this interval is referred to as the *oracle* CrI in the following and used as a benchmark to compare the performances of the algorithms.

Table 4 compares the CrIs constructed by the four experiment designs as well as the oracle CrI from $1{,}000$ macro-runs. Three different target coverage probabilities, $1 - \alpha = 0.9, 0.95$, and 0.99, are tested. For each macro-run, a new set of real-world demands are sampled from the true demand distributions and the joint posterior of $\boldsymbol{\theta}$ is updated conditional on the data. The oracle CrI is constructed from $M = 1{,}000$ $\boldsymbol{\theta}$s sampled from its posterior. The optimal design and the regression use the same $1{,}000$ $\boldsymbol{\theta}$s as outer scenarios to construct CrIs. The average of the simulation budget used by the optimal design across $1{,}000$ macro-runs is $1{,}471$ (with standard error 1.1), which is significantly less than $MN = 10^6$ for SNS$^+$. For the regression, polynomial basis functions up to order 2 were adopted without cross-terms reflecting that all product demands are independent.

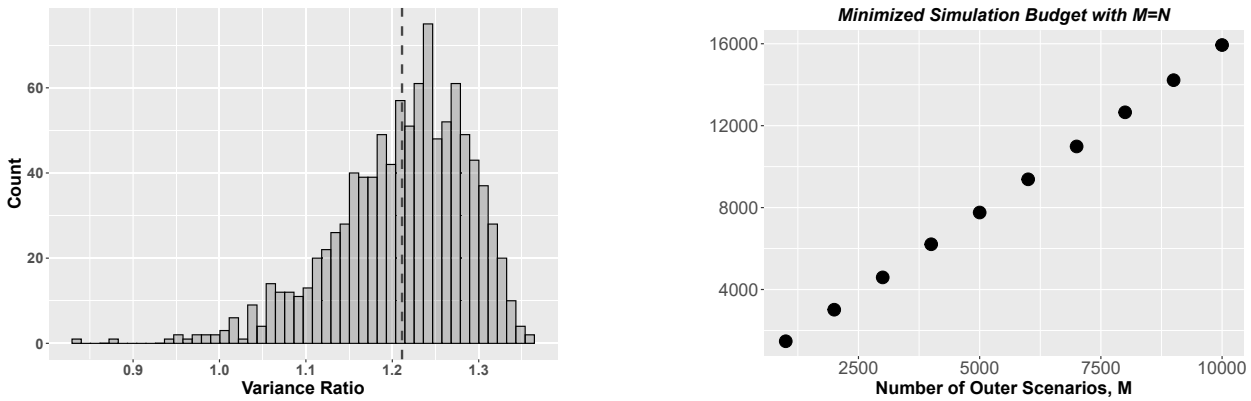
The empirical coverage probabilities and the widths of CrIs in Table 4 are averaged over $1{,}000$ macro runs. The former is computed via Monte Carlo simulation: a million $\boldsymbol{\theta}$s were drawn from the posterior distribution of $\boldsymbol{\theta}$ and $\mu(\boldsymbol{\theta})$s were computed using the analytical expression from which the coverage probability is estimated. For all algorithms, the same Monte Carlo sample was used.

Table 4 The estimated coverage probabilities and the widths of CrIs constructed by the oracle, optimal design, SNS, SNS$^+$, and regression from $1{,}000$ macro-runs. All methods use the same simulation budget except for SNS$^+$. The standard errors are in parentheses.

	Empirical coverage			Width		
Target $1-\alpha$	0.9	0.95	0.99	0.9	0.95	0.99
Oracle	0.898(3E-04)	0.948(2E-04)	0.988(1E-04)	81.20(2E-03)	96.55(3E-03)	125.71(5E-03)
Opt. Design	0.886(7E-04)	0.940(5E-04)	0.985(3E-04)	81.01(4E-03)	96.37(5E-03)	125.40(7E-03)
SNS	1.000(2E-06)	1.000(3E-07)	1.000(1E-08)	235.38(2E-02)	277.18(2E-02)	348.41(3E-02)
SNS$^+$	0.912(3E-04)	0.958(2E-04)	0.991(8E-05)	84.80(2E-03)	100.87(3E-03)	131.37(5E-03)
Regression	0.972(8E-04)	0.991(4E-04)	0.999(1E-04)	120.28(2E-02)	146.35(2E-02)	200.55(3E-02)

Table 4 shows that the CrIs constructed by the oracle and the optimal design are very close in both coverage and width across for all choices for α, although the latter shows a slight undercoverage compared to the former. The undercoverage is caused by that the optimal design interpolates the simulation outputs run at the sampling outer scenarios, however, it is alleviated as M grows. SNS clearly exhibits overcoverage and wide CrIs. This is because the small inner sample size, N, makes $\mathrm{V}[\bar{\mu}_i]$ large, which inflates the CrI. Notice that SNS$^+$ still overcovers and slightly wider CrIs than the oracle indicating that the inflation of CrI from MC error of $\bar{\mu}_i$ persists even with $N = 1{,}000$. The optimal design and SNS$^+$ show comparable performances across all αs, however, the former costs only $1/670$ of the simulation replications of the latter on average. The regression method shows clear overcoverage across all αs compared to the optimal design. In particular, the difference between the CrI widths from the two methods is wider for smaller α, which coincides with the observations from the ERM examples; the regression tends to work poorly at predicting $\mu(\boldsymbol{\theta})$ for extreme $\boldsymbol{\theta}$s. On the other hand, the optimal nested simulation design does not suffer from this by allocating more replications to the extreme outer scenarios so that they achieve the same target ESS.

We also examine how well our optimal design achieves the target variance of the conditional means, as required by the constraint (7), which illustrates effectiveness of the ESS expression in approximating the variance. Figure 3a shows the histogram of the ratios between the variance of the MC estimator, $\mathrm{V}[\bar{\mu}(\boldsymbol{\theta}_i)]$, and the variance of the pooled self-normalized LR estimator, $\mathrm{V}[\widetilde{\mu}(\boldsymbol{\theta}_i)]$, of the conditional means for a fixed set of 1,000 scenarios. Each MC estimator has $N = 1{,}000$ inner replications and the optimal design has the same target $N = 1{,}000$. The variance estimates are obtained from $1{,}000$ macro runs of both designs given the same outer scenarios. Figure 3a shows that the variance ratios are close to 1, with average 1.21 and maximum 1.38. This shows the ESS expression slightly underestimates the variance of the self-normalized



(a) (b)

Figure 3 **(a) is a histogram of ratios between the variances of the $\mu(\theta)$ estimates by the optimal design and by SNS. (b) plots the number of outer scenarios M against minimized simulation budget, assuming a constant target inner simulation $N = 1000$.**

LR estimator, which is a price we pay to approximate the variances without having to simulate the model at all.

Lastly, Figure 3b shows the growth of optimal simulation budget as M increases. In all designs, $N = M$ is chosen for M ranging from $1{,}000$ to $10{,}000$ and the plotted budgets are averaged over 100 macro runs. Figure 3b depicts that the minimized simulation budget increases approximately linearly in M. In comparison, the standard nested simulation's budget would be MN, so it would increase with M^2 when $N = M$. This comparison once again demonstrates efficiency of our experiment design.

8. Conclusions

We study an efficient nested simulation experiment design when the set of outer scenarios is given to the experimenter. Using the LR method, our design makes both sampling and pooling decision for the nested simulation. We minimize the total simulation budget while ensuring the (approximate) variance of the conditional mean estimator at each outer scenario is of the same level as the MC estimator's from the SNS design with user-provided N. Our asymptotic analysis shows that the estimation errors of the studied nested statistics achieve the fastest convergence rate when $M = \Theta(N)$. When the inner simulation's input distribution belongs in an exponential family, we show that the simulation budget of the optimal design grows in $O(M)$ by setting $N = M$. Compared to a state-of-the-art regression approach with the same simulation budget, the empirical analyses show that our optimal design delivers significantly smaller estimation error.

In this paper, we focused on the setting where the vector of all inputs generated within each inner replication has a known and fixed dimension across outer scenarios. In future research, we will extend the proposed scheme to a more general class of stochastic simulation models. One example is when the dimension is unknown and random, which is a typical characteristic of a queuing-type simulation model. For these cases,

computing the second moment of the LR without running the simulation model is a challenge, since the joint density of the inputs is unknown a priori. A two-stage, or sequential experiment design may be considered.

Acknowledgments

We would like to express our sincere gratitude to the anonymous reviewers who provided us constructive feedback.

References

Barton RR, Lam H, Song E (2022) Input uncertainty in stochastic simulation. Salhi S, Boylan J, eds., *The Palgrave Handbook of Operations Research*, 573–620 (Cham, Switzerland: Springer International Publishing).

Broadie M, Du Y, Moallemi CC (2011) Efficient risk estimation via nested sequential simulation. *Management Science* 57(6):1172–1194.

Broadie M, Du Y, Moallemi CC (2015) Risk estimation via regression. *Operations Research* 63(5):1077–1097.

Dang O, Feng M, Hardy MR (2022) Dynamic importance allocated nested simulation for variable annuity risk measurement. *Annals of Actuarial Science* 1–30.

Dong J, Feng M, Nelson BL (2018) Unbiased metamodeling via likelihood ratios. *Proceedings of the 2018 Winter Simulation Conference*, 1778–1789 (IEEE).

Elvira V, Martino L, Robert CP (2018) Rethinking the effective sample size. *arXiv preprint* Https://arxiv.org/abs/1809.04129, Last accessed on 6/20/2020.

Feng BM, Song E (2024) Efficient nested simulation experiment design via the likelihood ratio method. URL `http://dx.doi.org/10.1287/ijoc.2022.0392.cd`, available for download at https://github.com/INFORMSJoC/2022.0392/.

Feng M, Song E (2019) Efficient input uncertainty quantification via green simulation using sample path likelihood ratios. *Proceedings of the 2019 Winter Simulation Conference*, 3693–3704 (IEEE).

Feng M, Staum J (2015) Green simulation designs for repeated experiments. *Proceedings of the 2015 Winter Simulation Conference*, 403–413 (IEEE).

Feng M, Staum J (2017) Green simulation: Reusing the output of repeated experiments. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 27(4):1–28.

Fu MC (2015) *Handbook of Simulation Optimization* (New York, New York: Springer).

Giles MB, Haji-Ali AL (2019) Multilevel nested simulation for efficient risk estimation. *SIAM/ASA Journal on Uncertainty Quantification* 7(2):497–525.

Glasserman P (2003) *Monte Carlo Methods in Financial Engineering* (Springer).

Glasserman P, Xu X (2014) Robust risk measurement and model risk. *Quantitative Finance* 14(1):29–58.

Gordy MB, Juneja S (2010) Nested simulation in portfolio risk measurement. *Management Science* 56(10):1833–1848.

Ha H, Bauer D (2022) A least-squares monte carlo approach to the estimation of enterprise risk. *Finance and Stochastics* 26(3):417–459.

Heidelberger P, Lewis PAW (1984) Quantile estimation in dependent sequences. *Operations Research* 32(1):185–209.

Hesterberg TC (1988) *Advances in Importance Sampling*. Ph.D. thesis, Stanford University.

Hong LJ, Juneja S, Liu G (2017) Kernel smoothing for nested estimation with application to portfolio risk measurement. *Operations Research* 65(3):657–673.

Kleijnen JP, Rubinstein RY (1996) Optimization and sensitivity analysis of computer simulation models by the score function method. *European Journal of Operational Research* 88(3):413–427.

Kong A (1992) A note on importance sampling using standardized weights. university of chicago. Technical Report 348, Department of Statistics, University of Chicago.

L'Ecuyer P (1990) A unified view of the IPA, SF, and LR gradient estimation techniques. *Management Science* 36(11):1364–1383.

L'Ecuyer P (1993) Two approaches for estimating the gradient in functional form. *Proceedings of the 1993 Winter Simulation Conference*, 338–346 (IEEE).

Lee S (1998) *Monte Carlo Computation of Conditional Expectation Quantiles*. Ph.D. thesis, Stanford University Department of Engineering-Economic Systems and Operations Research.

Li P, Feng R (2021) Nested monte carlo simulation in financial reporting: a review and a new hybrid approach. *Scandinavian Actuarial Journal* 2021(9):744–778.

Liu JS (1996) Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and Computing* 6(2):113–119.

Liu M, Staum J (2010) Stochastic kriging for efficient nested simulation of expected shortfall. *Journal of Risk* 12(3):3–27.

Longstaff FA, Schwartz ES (2001) Valuing american options by simulation: a simple least-squares approach. *The review of financial studies* 14(1):113–147.

Maggiar A, Waechter A, Dolinskaya IS, Staum J (2018) A derivative-free trust-region algorithm for the optimization of functions smoothed via Gaussian convolution using adaptive multiple importance sampling. *SIAM Journal on Optimization* 28(2):1478–1507.

Martino L, Elvira V, Louzada F (2017) Effective sample size for importance sampling based on discrepancy measures. *Signal Processing* 131:386–401.

Owen AB (2013) *Monte Carlo theory, methods and examples* (https://statweb.stanford.edu/ owen/mc/, Last accessed on 6/20/2020).

Risk J, Ludkovski M (2018) Sequential design and spatial modeling for portfolio tail risk measurement. *SIAM Journal on Financial Mathematics* 9(4):1137–1174.

Rockafellar RT, Uryasev S (2002) Conditional value-at-risk for general loss distributions. *Journal of banking & finance* 26(7):1443–1471.

Rubinstein RY, Shapiro A (1993) *Discrete event systems: Sensitivity analysis and stochastic optimization by the score function method* (John Wiley & Sons Inc).

Sen PK (1972) On the bahadur representation of sample quantiles for sequences of ϕ-mixing random variables. *Journal of Multivariate Analysis* 2(1):77 – 95.

Song E, Wu-Smith P, Nelson BL (2020) Uncertainty quantification in vehicle content optimization for general motors. *INFORMS Journal on Applied Analytics* 50(4):225–238.

Sun Y, Apley DW, Staum J (2011) Efficient nested simulation for estimating the variance of a conditional expectation. *Operations Research* 59(4):998–1007.

Wang W, Wang Y, Zhang X (2022) Smooth nested simulation: Bridging cubic and square root convergence rates in high dimensions. *arXiv preprint arXiv:2201.02958* .

Zhang K, Feng BM, Liu G, Wang S (2022) Sample recycling for nested simulation with application in portfolio risk measurement. *arXiv preprint arXiv:2203.15929* .

Zhou E, Liu T (2019) Online quantification of input model uncertainty by two-layer importance sampling, https://arXiv:1912.11172, Last accessed on 6/20/2020.