Differentiable Calibration of Inexact Stochastic Simulation Models via Kernel Score Minimization

Ziwei Su
Northwestern University

Diego Klabjan
Northwestern University

Abstract

Stochastic simulation models are generative models that mimic complex systems to help with decision-making. The reliability of these models heavily depends on well-calibrated input model parameters. However, in many practical scenarios, only output-level data are available to learn the input model parameters, which is challenging due to the often intractable likelihood of the stochastic simulation model. Moreover, stochastic simulation models are frequently inexact, with discrepancies between the model and the target system. No existing methods can effectively learn and quantify the uncertainties of input parameters using only output-level data. In this paper, we propose to learn differentiable input parameters of stochastic simulation models using output-level data via kernel score minimization with stochastic gradient descent. We quantify the uncertainties of the learned input parameters using a frequentist confidence set procedure based on a new asymptotic normality result that accounts for model inexactness. The proposed method is evaluated on exact and inexact G/G/1 queueing models as well as a stochastic volatility model.

1 INTRODUCTION

Stochastic simulation models play a crucial role in diverse domains, including manufacturing, supply chain management, cloud computing, and epidemiology. They serve as powerful tools for mimicking complex systems that are costly or impractical to study directly. Examples of stochastic simulation models include queueing models (Pan et al., 2021) (Section A.1.1 of the Supplementary Material introduces queues), digital twins (Biller et al., 2022), stochastic inventory models (Zhang et al., 2021), and compartmental models (Frazier et al., 2022). These generative models generate random samples to help with decision-making. For instance, queueing models for the waiting lines of cloud computing systems help develop strategies to reduce customer waiting times.

The output distribution of a stochastic simulation model is determined by a set of unknown input values. We term this set of input values the *simulation parameter*. Calibrating the simulation parameter to align the simulation withth the target system is key to the reliability of the simulation model. Typically, this calibration is performed using input-level data or domain knowledge of the simulation parameter, termed input modeling (Barton et al., 2022). In practical scenarios where only output-level data are available, the process of learning the input simulation parameter is termed *stochastic simulation calibration* (Kleijnen, 1995).

Existing methods for stochastic simulation calibration, such as likelihood-free inference techniques (Cranmer et al., 2020), often assume that the simulation model is exact, i.e., there exists a simulation parameter that perfectly aligns the model output with the target system. However, stochastic simulation models are frequently inexact in reality due to the complexity of real-world target systems. No existing frequentist methods can effectively learn and quantify the uncertainties of input parameters using only output-level data under model inexactness. We address this challenge by exhibiting a frequentist learning and confidence set procedure based on kernel score (Dawid, 2007) minimization.

Notation and Setting Stochastic simulation calibration leverages a set of multi-dimensional output-level data $X_1, \ldots, X_m$ from the target system, presumed drawn independently from the target system output $P_\star$, i.e. $X_1, \ldots, X_m \overset{\text{i.i.d.}}{\sim} P_\star$. The simulation parameter θ in some parameter space $\Theta \subset \mathbb{R}^p$ determines the output distribution of the stochastic simulation model denoted P_θ. When the probability density

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

function f_θ exists, the maximum likelihood estimation (MLE) is the standard approach to estimate θ: $\widehat{\theta}_m^{\mathrm{MLE}} \in \arg\max_{\theta\in\Theta} \frac{1}{m}\sum_{i=1}^m \log f_\theta(X_i)$. MLE is an instance of the *optimum score estimation* (Gneiting and Raftery, 2007): $\widehat{\theta}_m \in \arg\min_{\theta\in\Theta} L_m(\theta) = \frac{1}{m}\sum_{i=1}^m S\left(P_\theta, X_i\right)$, where $S(\cdot,\cdot)$ is a *strictly proper scoring rule* (Gneiting and Raftery, 2007), $L_m(\theta)$ is the *optimum score*, and $\widehat{\theta}_m$ is the *optimum score estimator*. The negative log-likelihood $-\log f_\theta(x)$ is an example of a strictly proper scoring rule (*logarithmic score* (Good, 1952)) that corresponds to MLE. Minimizing the optimum score $L_m(\theta)$ amounts to minimizing a statistical distance induced by S between P_θ and the empirical measure $\widehat{P}_\star^m = \frac{1}{m}\sum_{i=1}^m \delta_{X_i}$, where δ denotes the Dirac delta function. Under model inexactness, optimum score minimization aims to find the best possible θ that minimizes the statistical distance between P_θ and $\widehat{P}_\star^m$. The optimum score estimator is hence also referred to as the minimum distance estimator (Basu et al., 2011).

Despite the asymptotic properties of optimum score estimators, such as MLE, the likelihood-free nature of stochastic simulation models hinders their direct application for simulation parameter estimation. While it is feasible to simulate realizations from P_θ, evaluating P_θ is usually impractical due to the complex probabilistic mapping between θ and P_θ. We denote the simulated sample from P_θ by $Y_1(\theta), \ldots, Y_n(\theta) \overset{\text{i.i.d.}}{\sim} P_\theta$. Consider a queueing model where the service time distributions for servers are parameterized by unknown service rates, forming the simulation parameter θ. It is feasible to simulate from the queueing model to obtain a random sample for any θ, such as generating a sample of the average waiting times for customers. However, deriving an expression for P_θ is analytically challenging. The likelihood of the model output P_θ is hence inaccessible. Stochastic simulation models can thus be viewed as intractable generative models analogous to deep generative models such as generative adversarial networks (GANs) (Goodfellow et al., 2014), diffusion models (Sohl-Dickstein et al., 2015), and variational autoencoders (VAEs) (Kingma and Welling, 2014). However, unlike deep generative models, stochastic simulation models are often considered *inexact*, with an inherent model discrepancy where the simulation does not match the target system under any parameter. No existing frequentist methods are well-equipped for stochastic simulation calibration under model inexactness. We further discuss the key differences between deep generative models and stochastic simulation models in Section A.1.4 of the Supplementary Material.

We address two distinct types of model discrepancy in this work: model contamination (correct model with contaminated data) and model inexactness (uncontaminated data with misspecified model). Our primary focus is on the latter case, where there exists an inherent structural mismatch between the simulation model and the target system. This is distinct from contamination models where data is noisy but the model is well-specified. Our experiments in Section 3 with varying shape parameters for service time distributions provide a clear measure of model inexactness—smaller values indicate greater structural discrepancy between the simulation and target system. Section A.1.3 presents a detailed comparison between model contamination and inexactness.

Contribution and Outline In this paper, we propose the first frequentist method to learn and quantify the uncertainties of a differentiable simulation parameter with output-level data under model inexactness. Our method, termed the *kernel optimum score estimation*, leverages kernel score minimization to estimate the simulation parameter. Kernel score is a strictly proper scoring rule that induces the maximum mean discrepancy (MMD) (Gretton et al., 2012). Notably, kernel score gradients can be unbiasedly estimated with U-statistic approximations using simulated samples, allowing stochastic gradient descent (SGD) optimization. Our proposed learning and confidence set procedure hence applies to learning and quantifying uncertainties of a differentiable simulation parameter. We validate the conditions outlined in Bińkowski et al. (2018) to ensure the unbiasedness of U-statistic gradient estimates in the stochastic simulation context, focusing on the G/G/1 queueing model. We then present the first asymptotic normality result for the kernel optimum score estimator under model inexactness. Based on this result, we propose a novel confidence set procedure for the simulation parameter. We evaluate the proposed method on both exact and inexact G/G/1 queueing models and a stochastic volatility model with higher-dimensional simulation parameter, demonstrating its effectiveness in learning and quantifying uncertainties of a simulation parameter.

Our work differs significantly from both Bayesian approaches (Pacchiardi et al., 2024) and previous frequentist analyses under model exactness (Briol et al., 2019). Compared to Bayesian approaches, our asymptotic normality result guarantees asymptotically correct coverage of our confidence sets, while posterior consistency results cannot guarantee proper frequentist coverage even under ideal conditions. Our theoretical developments also extend beyond Briol et al. (2019) in fundamental ways. We establish new asymptotic theory for model inexactness and prove asymptotic normality under different regularity conditions, leveraging different statistical tools such as the strong law of large numbers for generalized k-sample U-statistics (Theorem A.20). Section A.2.4 of the Supplementary

Material presents a detailed comparison with related Bayesian and frequentist approaches.

The remainder of the paper is organized as follows. Section 1.1 concludes the introduction with a literature review. Section 2 introduces the notions of scoring rules and kernel optimum score estimation, validates the unbiasedness of the gradient estimate, demonstrates the asymptotic normality of kernel optimum score estimators, and introduces a confidence set procedure for the simulation parameter. Section 3 studies the performance of the proposed method with extensive simulation experiments. Section 4 concludes the paper.

1.1 Literature Review

Our proposed kernel optimum score estimator is closely related to several works in the frequentist and Bayesian literature that leverage MMD or scoring rules. In the frequentist literature, it is akin to the minimum MMD estimator for generative models (Briol et al., 2019; Chérief-Abdellatif and Alquier, 2022), such as MMD-GANs (Dziugaite et al., 2015). Dziugaite et al. (2015) propose training generative neural networks with MMD as the discriminator and utilize SGD for network training. Bińkowski et al. (2018) provide the regularity conditions for the unbiasedness of the gradient estimate in the context of deep generative models. Briol et al. (2019) comprehensively investigate the statistical properties of minimum MMD estimators under model exactness. Chérief-Abdellatif and Alquier (2022) further investigate minimum MMD estimators under data dependence and outliers. Oates et al. (2022) offers the most general conditions for the strong consistency of minimum kernel discrepancy estimators, with the minimum MMD estimator as a special case. In the Bayesian literature, several works leverage MMD (Chérief-Abdellatif and Alquier, 2020) and scoring rules (Pacchiardi and Dutta, 2021; Pacchiardi et al., 2024) for generalized Bayesian inference. In particular, Pacchiardi et al. (2024) provides a posterior consistency result similar to our frequentist asymptotic normality result.

Another primary approach to likelihood-free inference for simulation models is the approximate Bayesian computation (ABC) (Beaumont et al., 2002). However, ABC is not robust to model inexactness. Several works address this issue by revising the original ABC (Frazier et al., 2020b,a; Fujisawa et al., 2021), using synthetic likelihood (Frazier et al., 2021; Frazier and Drovandi, 2021), posterior bootstrap and minimum MMD estimators (Dellaporta et al., 2022), or neural network approximations of the likelihood (Kelly et al., 2023) or posterior (Ward et al., 2022; Wehenkel et al., 2024). These works focus on contamination models where the data are noisy but the model is well-specified, while our approach considers the inexact case where the data is not noisy but the model is misspecified. Moreover, these works do not assess whether their methods can produce credible sets with valid coverage. In fact, Frazier et al. (2020b) demonstrates that under model inexactness, ABC posteriors can produce credible sets with arbitrary coverage levels. Another key difference is that existing works do not specifically consider inexact queueing models, which are a focus of our study. By addressing model inexactness in the context of queueing systems, our work fills an important gap in the literature and provides a frequentist approach to quantify the uncertainty of the simulation parameter.

Recent works in simulation-based inference explore alternative approaches to uncertainty quantification for simulation parameters. WALDO (Masserano et al., 2022) constructs confidence regions for simulation parameters by leveraging Wald statistics, offering a non-asymptotic approach to uncertainty quantification. Similarly, conformal prediction methods (LeRoy and Shafer, 2021) provide distribution-free coverage guarantees, although their application to stochastic simulators remains relatively unexplored. For handling model inexactness specifically, Huang et al. (2023) develop robust statistics using neural networks. In the realm of amortized inference, researchers propose efficient methods to address model inexactness through cost estimation (Gao et al., 2023) and data-driven calibration approaches (Wehenkel et al., 2023, 2024).

Our work also relates to existing research in stochastic operations research. The discrepancy between simulation model outputs and the target system is traditionally handled via iterative model *validation* and *calibration* (Kleijnen, 1995; Sargent, 2010). Validation confirms the simulation's accuracy by comparing output data from the target system and the simulation, often using statistical tests such as the Schruben-Turing test (Schruben, 1980) and the two-sample mean-difference test (Balci and Sargent, 1982). Calibration aligns the simulation with the target system by adjusting parameters using data. Existing calibration works investigate computing bounds on input simulation parameters (Bai and Lam, 2020), bounds on input-dependent quantities (Goeva et al., 2019), or learning model discrepancy (Plumlee and Lam, 2017). Similar to our approach, Bai and Lam (2020) construct confidence sets for the simulation parameter via empirical matching between the simulation and target system output using the Kolmogorov-Smirnov statistic. However, their method assumes model exactness, potentially leading to empty confidence sets in inexact model cases.

Lastly, extensive literature addresses calibration of deterministic computer models from both Bayesian (Kennedy and O'Hagan, 2001; Storlie et al., 2015; Plum-

lee, 2017) and frequentist (Tuo and Wu, 2015, 2016; Tuo, 2019; Plumlee, 2019) perspectives. However, these deterministic calibration approaches are not directly applicable to stochastic simulation calibration, where both target system and simulation model outputs are represented by probabilistic measures.

2 METHODOLOGY

2.1 Setting and Formal Model

Scoring Rules and Kernel Score A scoring rule is a function $S(P, x) : \mathcal{P} \times \mathbb{R}^d \to \mathbb{R}$ that evaluates a model based on its distribution $P \in \mathcal{P}$ and a realization $x \in \mathbb{R}^d$ of the d-dimensional random variable X (Gneiting and Raftery, 2007). Set $\mathcal{P}$ is the domain of the score function such that $S(P, x) < \infty$ for all $P \in \mathcal{P}$ and $x \in \mathbb{R}^d$. Denoting the distribution of X by $Q \in \mathcal{P}$, the *expected score* of P under Q is denoted $\mathbb{E}_{X\sim Q} S(P, X)$. It is assumed throughout this paper that $\mathbb{E}_{X\sim Q} S(P, X) < \infty$ for all $P, Q \in \mathcal{P}$. A scoring rule is strictly proper relative to $\mathcal{P}$ if and only if $\mathbb{E}_{X\sim Q} S(Q, X) = \mathbb{E}_{X\sim Q} S(P, X)$ implies $P = Q$. The statistical distance that the strictly proper scoring rule induces is the difference between expected scores, $d_S(P, Q) = \mathbb{E}_{X\sim Q} [S(P, X) - S(Q, X)]$, termed the *divergence function* associated with S. When a scoring rule is strictly proper, $d_S(P, Q) > 0$ for all $P \neq Q$.

Examples of strictly proper scoring rules include the logarithmic score, the *energy score* (ES) (Székely and Rizzo, 2013), and the *kernel score* (KS). The energy score is defined as: $\mathrm{ES}(P, x) = \mathbb{E}_Y \|Y - x\|^\beta - \frac{1}{2}\mathbb{E}_{Y,Y'} \|Y - Y'\|^\beta$, where $\|\cdot\|$ is the Euclidean norm, $\beta \in (0, 2)$, and Y, Y' are identically distributed, independent random variables with distribution P i.e. $Y, Y' \overset{\text{i.i.d.}}{\sim} P$. The energy score is a special case of the kernel score $\mathrm{KS}(P, x) = \mathbb{E}_{Y,Y'}[k(Y, Y')] - 2\mathbb{E}_Y [k(x, Y)]$, where $k(\cdot, \cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a measurable kernel on $\mathbb{R}^d$, and $Y, Y' \overset{\text{i.i.d.}}{\sim} P$. The energy score is the kernel score with the Riesz kernel $k(x, y) = -\frac{1}{2}\|x - y\|^\beta$. The kernel score is strictly proper when the kernel k is *characteristic*. Examples of characteristic kernels include the Gaussian kernel $k(x, y) = \exp\left(-\frac{\|x-y\|_2^2}{2\sigma}\right)$, where $\sigma > 0$ is the bandwidth parameter, and $\|\cdot\|_2$ is the Euclidean (L_2) distance on $\mathbb{R}^d$, and the Laplacian kernel $k(x, y) = \exp\left(-\frac{\|x-y\|_1}{\sigma}\right)$, where $\sigma > 0$ and $\|\cdot\|_1$ is the L_1 distance on $\mathbb{R}^d$. While the Riesz kernel is not characteristic, the fractional Brownian motion kernel (Sejdinovic et al., 2013) is a characteristic kernel whose corresponding squared MMD is the energy distance (see Section A.1.2 of the Supplementary Material): $k(x, y) = \frac{1}{2}\left(\|x\|^\beta + \|y\|^\beta - \|x - y\|^\beta\right)$.

Kernel Optimum Score Estimation Given the output distribution of the stochastic simulation model P_θ and the output data $X_1, \ldots, X_m \overset{\text{i.i.d.}}{\sim} P_\star$, recall that the optimum score estimator is defined as: $\widehat{\theta}_m \in \arg\min_{\theta\in\Theta} L_m(\theta) = \frac{1}{m}\sum_{i=1}^m S(P_\theta, X_i)$, where $S(\cdot, \cdot)$ is a strictly proper scoring rule, and $L_m(\theta)$ is the optimum score. Optimum score estimation is challenging since the output distribution of the stochastic simulation model P_θ is not directly observable. Instead, the simulation model output P_θ can only be assessed by means of the simulated sample $Y_1(\theta), \ldots, Y_n(\theta) \overset{\text{i.i.d.}}{\sim} P_\theta$. The idea is to pick a good strictly proper scoring rule that allows for an empirical unbiased approximation of its optimum score with the simulated sample. To this end, we propose to use kernel scores:

$$\widehat{\theta}_m^{\mathrm{KS}} \in \arg\min_{\theta\in\Theta} L_m^{\mathrm{KS}}(\theta) = \mathbb{E}_{Y(\theta),Y'(\theta)}[k(Y(\theta), Y'(\theta))] - \frac{2}{m}\sum_{i=1}^m \mathbb{E}_{Y(\theta)}\left[k(Y(\theta), X_i)\right],$$

where $Y(\theta), Y'(\theta) \overset{\text{i.i.d.}}{\sim} P_\theta$. We call $\widehat{\theta}_m^{\mathrm{KS}}$ *kernel optimum score estimator*. We term the U-statistic approximation of the kernel optimum score with the simulated sample the *kernel simulated score*:

$$\widehat{L}_{m,n}^{\mathrm{KS}}(\theta) = \frac{1}{n(n-1)}\sum_{1\le i\neq j\le n} k(Y_i(\theta), Y_j(\theta)) - \frac{2}{mn}\sum_{i=1}^m\sum_{j=1}^n k(Y_j(\theta), X_i),$$

and we call the minimizer $\widehat{\theta}_{m,n}^{\mathrm{KS}}$ the *kernel simulated score estimator*. The gradient of the kernel simulated score (Dziugaite et al., 2015; Briol et al., 2019) is:

$$\nabla_\theta \widehat{L}_{m,n}^{\mathrm{KS}}(\theta) = \frac{1}{n(n-1)}\sum_{1\le i\neq j\le n} \nabla_\theta k(Y_i(\theta), Y_j(\theta)) - \frac{2}{mn}\sum_{i=1}^m\sum_{j=1}^n \nabla_\theta k(Y_j(\theta), X_i),$$

where $\nabla_\theta(\cdot)$ denotes the gradient with respect to θ.

A key requirement of our approach is the differentiability of the simulation parameter θ, which is essential for both the gradient-based optimization of the kernel score and the asymptotic properties that underpin our confidence set procedure. Section 2.2 delves into the asymptotic properties of the kernel optimum score estimator. Briol et al. (2019) presents a generalization bound for the kernel optimum score estimator when the kernel k is bounded (Theorem A.18 in Section A.2.2 of the Supplementary Material). The generalization bound for unbounded kernels such as the Riesz kernel is still an open problem.

Unbiased Gradient Estimation The gradient of the kernel simulated score is an unbiased estimate of the gradient of the kernel optimum score under certain regularity conditions. Bińkowski et al. (2018) first establishes these conditions for deep generative models, and we adapt them in the rest of Section 2.1 to stochastic simulation models. We further validate these conditions for G/G/1 queueing models in Section A.3.2 of the Supplementary Material. This unbiased gradient estimation relies on the existence of pathwise derivatives. Section A.3.3 of the Supplementary Material discusses connections to infinitesimal perturbation analysis and alternative approaches for simulation parameters that lack pathwise differentiability.

We require an alternative representation of the stochastic simulation model output P_θ, referred to as the 'pushforward' representation (Bińkowski et al., 2018; Briol et al., 2019), to facilitate our discussion. Let P be a Borel probability measure on a measurable space $\mathcal{Z}$, typically a subspace of $\mathbb{R}^d$ or $\mathbb{R}^d$ itself. The simulation model output can be expressed as $P_\theta = G_\theta^\# P$, where $G_\theta : \mathcal{Z} \to \mathbb{R}^d$ is a measurable parametric map, and $G_\theta^\# P$ is the *pushforward* of P through G_θ.

In stochastic simulation models, G_θ can be interpreted as the input model or the deterministic part of the input-output map, with θ as its (input) simulation parameter. The pushforward entails simulating n i.i.d. samples $Y_1(\theta), \ldots, Y_n(\theta) \overset{\text{i.i.d.}}{\sim} P_\theta$ by first simulating n i.i.d. samples $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} P$ and then setting $Y_j(\theta) = G_\theta(Z_j)$ for $j = 1, \ldots, n$. Function G_θ is usually a composition of multiple nonparametric component functions $\phi_1, \ldots, \phi_i, \ldots$ and parametric component functions $\psi_{\theta,1}, \ldots, \psi_{\theta,j}, \ldots$. For instance, consider simulating a random sample Y from an exponential distribution with rate θ denoted $\text{Exp}(\theta)$ using inverse transform sampling. We first simulate pseudo-random number $Z \sim P = U(0,1)$, then set $Y = G_\theta(Z) = \psi_{\theta,1}(\phi_1(Z)) = \frac{\log(1-Z)}{\theta}$, where $\phi_1(z) = \log(1-z)$ and $\psi_{\theta,1}(z) = \frac{z}{\theta}$. The reference measure P, as in inverse transform sampling, is typically the uniform measure on $(0,1)^d$, though it can be specified differently based on requirements. Different choice of P yields different G_θ. The selection of P hence plays a key role in ensuring unbiasedness. The unbiased gradient estimation relies on the following regularity conditions.

Assumption 2.1. 1. $\mathbb{E}_X\|X\|^\alpha, \mathbb{E}_Z\|Z\|^\alpha < \infty$ for some $\alpha \geq 1$. 2. **Bounded kernel:** There exist constants C_0, C_1 such that $|k(x,y)| \leq C_0\left(\left(\|x\|^2 + \|y\|^2\right)^{\alpha/2} + 1\right)$, $\|\nabla_{x,y} k(x,y)\| \leq C_1\left(\left(\|x\|^2 + \|y\|^2\right)^{(\alpha-1)/2} + 1\right)$, where α is the same α in Assumption 1. 3. **Lipschitzness:** Each non-parametric component of G_θ, ϕ_i, is M-Lipschitz. 4. **Piecewise analytic:** For each ϕ_i, there are K_i functions ϕ_i^k, $k = 1, \ldots, K_i$, each real analytic on its input space, and these functions agree with ϕ_i on the closure of a set $\mathcal{D}^k$, and the sets $\mathcal{D}_i^k$ cover the whole input space. 5. **Differentiability:** Each parametric component of G_θ, $\psi_{\theta,j}$, is almost everywhere differentiable on the parameter space Θ.

Assumption 1 in Assumption 2.1 naturally hold. Assumption 2 (Assumption E in Bińkowski et al. (2018)) holds for a wide range of kernels, including the Gaussian kernel and the Riesz kernel for $1 \leq \beta < 2$. Assumptions 3 and 4 (Assumptions C and D in Bińkowski et al. (2018)) hold for the vast majority of the activation functions used in deep neural networks, such as softmax and ReLU. However, Assumptions 3, 4, and 5 deal with components of G_θ, they are hence context-dependent. We leverage the Lindley equation (Lindley, 1952) to discuss these conditions for G/G/1 queueing models in Section A.3.2 of the Supplementary Material.

Proposition 2.2 (Unbiased Gradient Estimate)**.** *If Assumption 2.1 holds, then* $\mathbb{E}_{Y_1(\theta),\cdots,Y_n(\theta)} \nabla_\theta \widehat{L}_{m,n}^{\text{KS}}(\theta) = \nabla_\theta L_m^{\text{KS}}(\theta)$.

The proof follows from Theorem 5 in Bińkowski et al. (2018) and the chain rule. Proposition 2.2 enables the design of an SGD algorithm in the model inexactness setting as outlined in Algorithm 1. The loop is the proposed SGD algorithm and the remaining steps are about the confidence set procedure; see Section 2.3 for details. Note that Briol et al. (2019) offers a similar SGD algorithm under model exactness.

Time Complexity of Kernel Optimum Score Estimation Applying SGD in Algorithm 1 can be computationally costly. Following from Briol et al. (2019), the time complexity for SGD in Algorithm 1 in each iteration is $O((n^2 + mn)dp)$, where d is the dimension of data and p is the dimension of the parameter space Θ. The quadratic time complexity in the simulated sample size n renders the use of large n impractical within limited computational resources. To determine the most computationally efficient n, we need to find the optimal ratio $\lambda = \frac{n}{m+n}$ corresponding to the highest statistical efficiency i.e. minimizing the asymptotic variance-covariance for the kernel simulated score estimator. Note that simulated score estimator is not used in practice as it needs to resample from P_θ to obtain $Y_1(\theta), \ldots, Y_n(\theta)$ in each SGD iteration. Theorem 2.4 provides the asymptotic normality of both the kernel simulated score estimator and the kernel optimum score estimator. Section A.3.1 of the Supplementary Material discusses the optimal ratio λ in detail.

Optimal Simulation Parameter Minimizing the optimum score asymptotically amounts to min-

imizing the statistical distance $d_S(P_\theta, P_\star)$ between P_θ and $P_\star$. The parameter that minimizes $d_S(P_\theta, P_\star)$ is termed the *optimal simulation parameter*: $\theta_\star \in \arg\min_{\theta\in\Theta} d_S(P_\theta, P_\star) = \mathbb{E}_{X\sim P_\star}\left[S(P_\theta, X) - S(P_\star, X)\right]$. The choice of the kernel influences both the optimal simulation parameter and the statistical properties of the kernel optimum score estimator, discussed in Section 2.3 and Section A.3.1 of the Supplementary Material. The optimal simulation parameter may not be unique due to the complex nature of the input-output probabilistic map. This issue is known as model non-identifiability, a common challenge in inverse problems where the input is calibrated with the output (Tarantola, 2005). This challenge is exacerbated when the scoring rule used for estimation lacks the strict propriety. An example is the Dawid-Sebastiani score (Dawid and Sebastiani, 1999), which is proper but not strictly proper, relying solely on the first two moments. Multiple $\theta \in \Theta$ may exist such that P_θ aligns with the first two moments of $P_\star$. Using strictly proper scoring rules for estimation helps alleviate model non-identifiability problems. In the case of an exact $(P_\star \in \{P_\theta : \theta \in \Theta\})$ and an identifiable stochastic simulation model, the optimal simulation parameter $\theta_\star$ is the true simulation parameter θ_0 (Proposition A.21).

2.2 New Results on Asymptotic Normality under Model Inexactness

We denote the optimal simulation parameter with the kernel score by $\theta_\star^{\mathrm{KS}}$. The strong consistency property of the kernel optimum score estimator, i.e., $\widehat{\theta}_m^{\mathrm{KS}} \overset{\text{a.s.}}{\to} \theta_\star^{\mathrm{KS}}$, is given in the Supplementary Material (Proposition A.15). Here, we provide the speed of convergence in terms of asymptotic variance-covariance with asymptotic normality of the kernel optimum score estimator. We also exhibit asymptotic normality of the kernel simulated score estimator to facilitate the discussion of the choice of the ratio $\lambda = \frac{n}{m+n}$. The following assumption gives the regularity conditions for the asymptotic normality of the kernel optimum score estimator and kernel simulated score estimator.

Assumption 2.3. 1. There exists an open and convex neighborhood of $\theta_\star^{\mathrm{KS}}$ denoted $\mathcal{N}$ such that $\mathbb{E}_{Z,X}\sup_{\theta\in\mathcal{N}}\|\nabla_\theta k(G_\theta(Z), X)\| < \infty$ and $\mathbb{E}_{Z,Z'}\sup_{\theta\in\mathcal{N}}\|\nabla_\theta k(G_\theta(Z), k(G_\theta(Z'))\| < \infty$ for all $\theta \in \mathcal{N}$. 2. $\mathbb{E}_{Z,X}\left\|\nabla_\theta k(G_\theta(Z), X)\right\|^2 |_{\theta=\theta_\star^{\mathrm{KS}}} < \infty$ and $\mathbb{E}_{Z,Z'}\left\|\nabla_\theta k(G_\theta(Z), k(G_\theta(Z'))\right\|^2 |_{\theta=\theta_\star^{\mathrm{KS}}} < \infty$. 3. For all $r, s \in \{1, \ldots, p\}$ and $\theta = (\theta_1, \ldots, \theta_p) \in \mathbb{R}^p$:

$$\mathbb{E}_{Z,X}\left[\left|\nabla_{\theta_r\theta_s} k(G_\theta(Z), X) |_{\theta=\theta_\star^{\mathrm{KS}}}\right| \log^+\left(\left|\nabla_{\theta_r\theta_s} k(G_\theta(Z), X) |_{\theta=\theta_\star^{\mathrm{KS}}}\right|\right)\right],$$
$$\mathbb{E}_{Z,Z'}\left[\left|\nabla_{\theta_r\theta_s} k(G_\theta(Z), G_\theta(Z')) |_{\theta=\theta_\star^{\mathrm{KS}}}\right|\right] < \infty.$$

4. $\nabla_\theta k(G_\theta(z), x)$, $\nabla_\theta k(G_\theta(z), G_\theta(z'))$ are differentiable on $\mathcal{N}$ for P-almost all $z, z' \in Z$ and $P_\star$-almost all $x \in \mathbb{R}^d$. 5. $\nabla_{\theta\theta} k(G_\theta(z), x)$, $\nabla_{\theta\theta} k(G_\theta(z), G_\theta(z'))$ are uniformly continuous at $\theta_\star^{\mathrm{KS}}$ for P-almost all $z, z' \in Z$ and $P_\star$-almost all $x \in \mathbb{R}^d$, where $\nabla_{\theta\theta}(\cdot)$ denotes the Hessian with respect to θ. 6. The Hessian $H = \mathbb{E}_{Z,Z'}\left[\nabla_{\theta\theta} k(G_\theta(Z), G_\theta(Z')) |_{\theta=\theta_\star^{\mathrm{KS}}}\right] - 2\mathbb{E}_{Z,X}\left[\nabla_{\theta\theta} k(G_\theta(Z), X) |_{\theta=\theta_\star^{\mathrm{KS}}}\right]$ is positive definite.

While technical in nature, these regularity conditions are standard assumptions for central limit theorems in this context. Section A.3.4 of the Supplementary Material discusses practical verification of these assumptions for specific model classes and identifies scenarios where they are likely to hold.

Theorem 2.4 (Asymptotic Normality)**.** *Let us suppose* $\widehat{\theta}_m^{\mathrm{KS}} \overset{\text{a.s.}}{\to} \theta_\star^{\mathrm{KS}}$, $\widehat{\theta}_{m,n}^{\mathrm{KS}} \overset{\text{a.s.}}{\to} \theta_\star^{\mathrm{KS}}$ *as* $m, n \to \infty$, $\lim_{m,n\to\infty} \frac{n}{m+n} = \lambda \in (0, 1)$, *and Assumption 2.3 holds. Then as* $m \to \infty$,

$$\sqrt{m}(\widehat{\theta}_m^{\mathrm{KS}} - \theta_\star^{\mathrm{KS}}) \overset{\text{d}}{\to} N(0, C),$$

where $\overset{\text{d}}{\to}$ *denotes convergence in distribution. The variance-covariance matrix is the Godambe matrix* $C = H^{-1}\Sigma H^{-1}$, *where* $\Sigma = 4\mathbb{C}_X\left[\mathbb{E}_Z\left[\nabla_\theta k(G_\theta(Z), X) |_{\theta=\theta_\star^{\mathrm{KS}}}\right]\right]$, $\mathbb{C}[\cdot]$ *denotes the variance-covariance matrix. Furthermore,*

$$\sqrt{m+n}(\widehat{\theta}_{m,n}^{\mathrm{KS}} - \theta_\star^{\mathrm{KS}}) \overset{\text{d}}{\to} N(0, C_\lambda),$$

where $C_\lambda = H^{-1}\Sigma_\lambda H^{-1}$, $\Sigma_\lambda = \frac{\mathbb{C}_Z[2h(Z) - g_{0,1}(Z)]}{\lambda} + \frac{\mathbb{C}_X[g_{1,0}(X)]}{1-\lambda}$, $h(z) = \mathbb{E}_{Z'}\nabla_\theta k(G_\theta(z), G_\theta(Z')) |_{\theta=\theta_\star^{\mathrm{KS}}}$, $g_{0,1}(z) = 2\mathbb{E}_X\nabla_\theta k(G_\theta(z), X) |_{\theta=\theta_\star^{\mathrm{KS}}}$, *and* $g_{1,0}(x) = 2\mathbb{E}_Z\nabla_\theta k(G_\theta(Z), x) |_{\theta=\theta_\star^{\mathrm{KS}}}$.

See Section A.2.5 of the Supplementary Material for the proof. Note that $\Sigma = \mathbb{C}_X\left[g_{1,0}(X)\right]$. When the model is exact and identifiable, Theorem 2.4 reduces to the asymptotic normality under model exactness as in Theorem 2 of Briol et al. (2019), where the asymptotic variance-covariance of $\widehat{\theta}_{m,n}^{\mathrm{KS}}$ is $\frac{C}{\lambda(1-\lambda)}$. The asymptotic variance-covariance is hence minimized at $\lambda = \frac{1}{2}$ i.e. the simulated sample size n equals the output data size from the target system m, meaning that using n much larger than m is not computationally efficient.

However, under model inexactness, the determination of an optimal λ becomes complex due to the instance-dependent nature of the asymptotic variance-covariance C_λ, influenced by various factors, such as the kernel k, the reference measure of Z, P, and the input model G_θ, rendering it computationally intractable. A nuanced trade-off between the variance-covariance terms from the simulated sample Z and the data X further complicates the selection of λ.

Despite the complexity of C_λ, we can derive analytical insights for determining the optimal λ. For one-dimensional parameters ($p = 1$), asymptotic variance C_λ is minimized at $\lambda = \frac{\sqrt{c_1}}{\sqrt{c_1}+\sqrt{c_2}}$, where $c_1 = C_Z[2h(Z) - g_{0,1}(Z)]$ and $c_2 = C_X[g_{1,0}(X)]$. When $c_1 = c_2$, including the model exactness case, $\lambda = \frac{1}{2}$. Heuristically, when the distributions of X and $G_\theta(Z)|_{\theta=\theta^{\mathrm{KS}}_\star}$ are similar (small model discrepancy), λ approaches toward $\frac{1}{2}$, suggesting that using n much larger than m is inefficient. Conversely, when the variance of $Y(\theta^{\mathrm{KS}}_\star)$ significantly exceeds that of X, increasing the simulation sample size might be beneficial for variance reduction.

We can also use C and C_λ for kernel selection (Briol et al., 2019). While both C_λ and C are computationally intractable as the unknown parameter $\theta^{\mathrm{KS}}_\star$, consistent estimations can be obtained (See Section 2.3). Section A.3.1 of the Supplementary Material further discusses λ and kernel selection.

2.3 New Confidence Set Procedure and Analyses

Theorem 2.4 allows the construction of an asymptotically valid confidence set for the optimal simulation parameter $\theta^{\mathrm{KS}}_\star$. Such a confidence procedure requires a consistent estimator for the asymptotic variance-covariance matrix $C = H^{-1}\Sigma H^{-1}$. Both H and Σ have expectation and covariance terms concerning X and Z, and depend on $\theta^{\mathrm{KS}}_\star$. To obtain consistent estimators for H and Σ, we can estimate expectation and covariance terms with empirical estimates, and replace $\theta^{\mathrm{KS}}_\star$ with the obtained kernel optimum score estimator $\widehat{\theta}^{\mathrm{KS}}_m$. We can use a larger simulation sample size n_c to estimate expectation and covariance terms. Let $\widehat{H}_{m,n_c}$, $\widehat{\Sigma}_{m,n_c}$ denote the consistent estimators to H and Σ, respectively. Then

$$\widehat{H}_{m,n_c} = \frac{1}{n_c(n_c-1)} \sum_{1\le i\ne j\le n_c} \nabla_{\theta\theta} k(G_\theta(Z_i), G_\theta(Z_j)) \Big|_{\theta=\widehat{\theta}^{\mathrm{KS}}_m} - \frac{2}{mn_c} \sum_{i=1}^{m}\sum_{j=1}^{n_c} \nabla_{\theta\theta} k(G_\theta(Z_j), X_i) \Big|_{\theta=\widehat{\theta}^{\mathrm{KS}}_m}. \tag{1}$$

Let $\widehat{\mu}_{m,n_c} = \frac{1}{mn_c}\sum_{i=1}^{m}\sum_{j=1}^{n_c} \nabla_\theta k\left(G_\theta(Z_j), X_i\right)|_{\theta=\widehat{\theta}^{\mathrm{KS}}_m}$, $\widehat{\mu}_i = \frac{1}{n_c}\sum_{j=1}^{n_c} \nabla_\theta k\left(G_\theta(Z_j), X_i\right)|_{\theta=\widehat{\theta}^{\mathrm{KS}}_m}$. Then we have

$$\widehat{\Sigma}_{m,n_c} = \frac{4}{m-1}\sum_{i=1}^{m}(\widehat{\mu}_i - \widehat{\mu}_{m,n_c})^T(\widehat{\mu}_i - \widehat{\mu}_{m,n_c}). \tag{2}$$

The following assumption is regarding the consistency of $\widehat{\Sigma}_{m,n_c}$.

Assumption 2.5. $\nabla_\theta k(G_\theta(z), x)$ is uniformly continuous and uniformly bounded at $\theta^{\mathrm{KS}}_\star$ for P-almost all $z, z' \in Z$ and $P_\star$-almost all $x \in \mathbb{R}^d$.

Algorithm 1 Kernel Optimum Score Estimation and Confidence Set Estimation

Input: Initial $\widehat{\theta}^{(0)}_{m,n} \in \Theta$, output data $X_1, \ldots, X_m \overset{\text{i.i.d.}}{\sim} P_\star$, step sizes $\{\eta_t\}_{t\ge 0}$, simulation sample size n per iteration, simulation sample size n_c for confidence set estimation, significance level α. Initialize $t = 0$.

repeat Sample $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} P$ and set $Y_j(\widehat{\theta}^{(t)}_{m,n}) = G_{\widehat{\theta}^{(t)}_{m,n}}(Z_j)$ for $j = 1, \ldots, n$. Compute $\widehat{\theta}^{(t+1)}_{m,n} = \Pi_\Theta\left[\widehat{\theta}^{(t)}_{m,n} - \eta_t \nabla_\theta \widehat{L}^{\mathrm{KS}}_{m,n}(\widehat{\theta}^{(t)}_{m,n})\right]$. Set $t = t + 1$.

until convergence criterion is met.

Let $\widehat{\theta}^{\mathrm{KS}}_m = \widehat{\theta}^{(t)}_{m,n}$. Sample $Z_1, \ldots, Z_{n_c} \overset{\text{i.i.d.}}{\sim} P$. Compute $\widehat{\Sigma}_{m,n_c}$ with (2) and $\widehat{H}_{m,n_c}$ with (1). Compute $\mathrm{CS}^\alpha_{m,n_c}$ with (3).

Output: $\widehat{\theta}^{\mathrm{KS}}_m$ and $\mathrm{CS}^\alpha_{m,n_c}$.

We denote the consistent estimator for C by $\widehat{C}_{m,n_c} = \widehat{H}^{-1}_{m,n_c}\widehat{\Sigma}_{m,n_c}\widehat{H}^{-1}_{m,n_c}$.

Theorem 2.6. *If Assumptions 2 and 3 in Assumption 2.3, and Assumption 2.5 hold, then* $\widehat{H}_{m,n_c} \xrightarrow{\mathrm{p}} H$, $\widehat{\Sigma}_{m,n_c} \xrightarrow{\mathrm{p}} \Sigma$, *as* $m, n_c \to \infty$, *and* $\widehat{C}_{m,n_c} = \widehat{H}^{-1}_{m,n_c}\widehat{\Sigma}_{m,n_c}\widehat{H}^{-1}_{m,n_c} \xrightarrow{\mathrm{p}} H^{-1}\Sigma H^{-1}$ *as* $m, n_c \to \infty$.

See Section A.2.6 of the Supplementary Material for the proof. Theorem 2.6 guarantees that the ellipsoid

$$\left\{\theta \in \Theta : \left\|\sqrt{m}\widehat{\Sigma}^{-\frac{1}{2}}_{m,n_c}\widehat{H}_{m,n_c}\left(\theta - \widehat{\theta}^{\mathrm{KS}}_m\right)\right\|^2 \le \chi^2_{1-\alpha}(p)\right\} \tag{3}$$

denoted $\mathrm{CS}^\alpha_{m,n_c}$ is an asymptotically valid $100(1-\alpha)\%$ confidence set for $\theta^{\mathrm{KS}}_\star$, where $\chi^2_{1-\alpha}(p)$ is the $(1-\alpha)$-quantile of the chi-squared distribution with p degrees of freedom where p is the dimension of the parameter space Θ. The procedure to compute $\mathrm{CS}^\alpha_{m,n_c}$ is summarized in Algorithm 1. The time complexity for computing $\widehat{H}_{m,n_c}$ and $\widehat{\Sigma}_{m,n_c}$ in total is $O((n_c^2 + mn_c)pd)$ if using automatic differentiation for Hessian estimation, and $O((n_c^2 + mn_c)p^2d)$ if using finite difference for Hessian estimation. The time complexity for checking whether an arbitrary point θ is in $\widehat{C}^\alpha_{m,n_c}$ is $O(p^3)$.

3 NUMERICAL EXPERIMENTS

This section studies the performance of Algorithm 1 on exact and inexact stochastic simulation models of G/G/1 queueing models. Throughout the experiments, unless otherwise specified, we use the Riesz kernel with $\beta = 1$ (KOSE-Riesz) and the Gaussian kernel with the bandwidth determined by the median heuristic (Gretton et al., 2012) (KOSE-Gaussian). We consider the case where a sample of average waiting times $X_1, \ldots, X_m$ from the target G/G/1 system output $P_\star$

is available. The corresponding G/G/1 queueing model output is $P_\theta = G_\theta(Z)$, where θ is the input simulation parameter. Table 1 gives the details of the settings for Experiments 1 to 4. We use Experiments 2 and 4 to explore how the performance changes with increasing model inexactness. Specifically, we set the shape parameter of the service time distribution to be $a = 1, 0.8, 0.6, 0.4, 0.2$, where $a = 1$ is the exact model case, and smaller values of a indicate higher model inexactness. Further experiment details are reported in Section A.4 of the Supplementary Material. Our SGD-based approach requires the differentiability of θ and the unbiasedness of the gradient estimate. We discuss and validate these requirements for G/G/1 queueing models and for other models in Sections A.3.2 and A.3.4 of the Supplementary Material, respectively.

Table 1: Experiment settings. Arr., Serv. give the distributions of the inter-arrival time and service time in the target G/G/1 queueing system and the G/G/1 simulation model. Gam. means Gamma.

No.	Dist.	Target	Model	θ
1	Arr.	Exp(1)	Exp(1)	μ
	Serv.	Exp(1.2)	Exp(μ)	
2	Arr.	Exp(1)	Exp(1)	μ
	Serv.	Gam.(a, 1.2)	Exp(μ)	
3	Arr.	Gam.(0.5, 1)	Gam.(0.5, λ)	(μ, λ)
	Serv.	Exp(1)	Exp(μ)	
4	Arr.	Gam.(0.5, 1)	Gam.(0.5, λ)	(μ, λ)
	Serv.	Gam.(a, 2.5)	Exp(μ)	

We compare our method with three baselines: accept-reject ABC (ABC-AR) (Beaumont et al., 2002) with MMD, MMD posterior bootstrap (NPL-MMD) (Dellaporta et al., 2022), and eligibility set (ESet) (Bai and Lam, 2020). For each baseline method and our proposed method, we record performance metrics such as mean-squared error (MSE), Monte Carlo coverage of credible or confidence sets, width of credible or confidence sets, and average run time. These results are presented in Tables 4, 2, 5, 3 for Experiments 1, 2, 3, 4, respectively. We also provide the plots for the confidence set produced in Experiment 4 for KOSE-Riesz (Figures 1 and 2) and KOSE-Gaussian (Figures 3 and 4) in $a = 1, 0.6$.

We conclude from Tables 2 and 3 that KOSE-Riesz consistently produces valid confidence sets (despite a slight undercoverage issue for $a = 1$ in Table 2), except for $a = 0.4, 0.2$ in Experiment 4, both more extreme model inexactness scenarios. We use $R = 1,000$ in most cases except when the run time is larger we use $R = 100$. In all model inexactness scenarios, KOSE-Riesz produces stable optimum score estimates with constantly small MSE. KOSE-Gaussian suffers from small undercoverage issues for $a = 0.8, 0.2$ in Experiment 2, and the produced optimum score estimates appear less stable than those of KOSE-Riesz, while all the benchmarks do not produce valid credible or confidence intervals under model inexactness. We also observe from Tables 4 and 5 of the Supplementary Material that our method consistently produces valid confidence sets, for both KOSE-Riesz and KOSE-Gaussian under model exactness. And both KOSE-Riesz and KOSE-Gaussian yield better MSE for the optimum score estimates and smaller confidence set width as m increases, as expected. ABC-AR and ESet methods also produce valid credible or confidence sets for larger m under model exactness, while NPL-MMD does not produce valid credible sets even under model exactness.

We observe that our confidence sets consistently achieve coverage rates that exceed the nominal 95% level. This conservative behavior is preferable to undercoverage, which would invalidate uncertainty quantification. In statistical practice, confidence sets that contain the true parameter with probability higher than the nominal level ensure reliability at the cost of potentially wider intervals. This trade-off is particularly valuable in our context where model inexactness and optimization error introduce additional uncertainty to simulation parameter estimation. Future work could explore methods such as regularized asymptotic variance estimation (Oates et al., 2022) to potentially reduce confidence set width while maintaining valid coverage.

Additional experiments on different choices of β in KOSE-Riesz, sensitivity of the simulation sample size n, bias in uncertainty evaluation, contamination models, and a stochastic volatility model are reported in Sections A.4.1, A.4.2, A.4.3, A.4.4, and A.4.5 of the Supplementary Material, respectively.

4 CONCLUSION

In this paper, we introduce the first frequentist method to learn and quantify the uncertainties of differentiable simulation parameters with output-level data under model inexactness. Our approach is based on kernel score minimization and a new asymptotic normality result. We demonstrate the superior numerical performance of our approach for both exact and inexact G/G/1 queueing models. Our approach produces accurate estimators with low mean-squared error (MSE) and more valid confidence sets, outperforming both Bayesian (Beaumont et al., 2002; Dellaporta et al., 2022) and frequentist (Bai and Lam, 2020) methods.

While our method shows promising results, we acknowledge several important considerations. The scalability to higher parameter dimensions remains an open challenge, as practical difficulties arise in Hessian estima-

Table 2: Experiment 2 results; Value a is the shape parameter of the service time distribution in the target system; R is the number of independent runs; $m = n = 500$. MSE is the average mean-squared error, Cov. is the coverage of the 95% confidence or credible set, Width is the average length of the confidence or credible set, and T is the average time per run in seconds. Empty is the number of empty confidence sets. Parameter $\theta_\star^{\mathrm{KS}}$ is the optimal simulation parameter obtained using KOSE-Gaussian.

	KOSE-Riesz, $R = 1,000$				
a	MSE	Cov.	Width	T	$\theta_\star^{\mathrm{KS}}$
1	$4.0 \cdot 10^{-4}$	0.920	0.07	3	1.2
0.8	$1.6 \cdot 10^{-4}$	**0.996**	0.07	3	1.5
0.6	$2.6 \cdot 10^{-4}$	**0.999**	0.10	3	1.9
0.4	$6.7 \cdot 10^{-4}$	**0.999**	0.15	3	2.5
0.2	$3.0 \cdot 10^{-3}$	**1.000**	0.34	3	4.1

	KOSE-Gaussian, $R = 1,000$				
a	MSE	Cov.	Width	T	$\theta_\star^{\mathrm{KS}}$
1	$1.8 \cdot 10^{-4}$	**0.990**	0.06	4	1.2
0.8	$7.4 \cdot 10^{-4}$	0.842	0.08	4	1.5
0.6	$8.0 \cdot 10^{-4}$	**0.942**	0.10	3	1.9
0.4	$1.0 \cdot 10^{-3}$	**0.997**	0.17	4	2.5
0.2	$1.9 \cdot 10^{-2}$	0.833	0.37	4	4.1

	ABC-AR, $R = 1,000$			
a	MSE	Cov.	Width	T
1	$4.4 \cdot 10^{-5}$	**0.994**	0.04	3
0.8	$9.0 \cdot 10^{-2}$	0.000	0.06	3
0.6	0.83	0.000	0.13	3
0.4	0.93	0.000	0.08	3
0.2	0.81	0.000	0.24	3

	NPL-MMD, $R = 100$			ESet, $R = 1,000$		
a	Cov.	Width	T	Cov.	Empty	T
1	0.63	0.09	81	**0.989**	0	3
0.8	0.00	0.10	77	0.000	0	3
0.6	0.00	0.12	79	0.000	0	3
0.4	0.00	0.15	76	NA	1,000	3
0.2	0.00	0.22	78	NA	1,000	3

Table 3: Experiment 4 results; Here, $m = n = 1,000$. MSE, μ and MSE, λ are the average mean-squared errors for the service and arrival rates.

	KOSE-Riesz, $R = 1,000$				
a	MSE, μ	MSE, λ	Cov.	T	$\theta_\star^{\mathrm{KS}}$
1	$1.3 \cdot 10^{-2}$	$4.7 \cdot 10^{-2}$	**1**	10	$(2.5, 1.0)$
0.8	$1.2 \cdot 10^{-2}$	$3.1 \cdot 10^{-4}$	**1**	9	$(6.0, 2.9)$
0.6	$4.2 \cdot 10^{-3}$	$4.7 \cdot 10^{-2}$	**1**	9	$(7.9, 3.1)$
0.4	$2.3 \cdot 10^{-4}$	$7.6 \cdot 10^{-4}$	0	9	$(10, 2.9)$
0.2	$1.2 \cdot 10^{-5}$	$4.3 \cdot 10^{-3}$	0	9	$(9.5, 1.0)$

	KOSE-Gaussian, $R = 1,000$				
a	MSE, μ	MSE, λ	Cov.	T	$\theta_\star^{\mathrm{KS}}$
1	$1.1 \cdot 10^{-2}$	$4.0 \cdot 10^{-3}$	**1**	9	$(2.5, 1.0)$
0.8	$4.6 \cdot 10^{-2}$	$2.0 \cdot 10^{-2}$	**1**	9	$(6.0, 2.9)$
0.6	$2.0 \cdot 10^{-2}$	$1.1 \cdot 10^{-2}$	**1**	9	$(7.9, 3.1)$
0.4	$4.6 \cdot 10^{-2}$	$8.4 \cdot 10^{-3}$	0	9	$(10, 2.9)$
0.2	$6.7 \cdot 10^{-5}$	$3.9 \cdot 10^{-1}$	0	9	$(9.5, 1.0)$

	ABC-AR, $R = 100$			
a	MSE	Cov.	Width, μ, λ	T
1	$8.9 \cdot 10^{-3}$	0.41	$0.54, 0.69$	88
0.8	$1.7 \cdot 10^{-3}$	0.00	$0.24, 0.25$	88
0.6	$3.0 \cdot 10^{-4}$	0.00	$0.26, 0.24$	88
0.4	$7.9 \cdot 10^{-5}$	0.00	$0.32, 0.24$	88
0.2	$2.3 \cdot 10^{-5}$	0.00	$0.23, 0.35$	87

	NPL-MMD, $R = 100$			ESet, $R = 1,000$		
a	Cov.	Width, μ, λ	T	Cov.	Empty	T
1	0.59	$0.74, 0.62$	66	0	3	2
0.8	0.59	$1.99, 1.44$	64	NA	1,000	3
0.6	0.32	$3.11, 1.65$	63	NA	1,000	3
0.4	0.66	$8.00, 4.67$	58	NA	1,000	3
0.2	0.27	$0.06, 4.91$	55	NA	1,000	3

tion and memory requirements despite our dimension-independent asymptotic theory. Our empirical validation focuses primarily on G/G/1 queueing models, though our preliminary work with stochastic volatility models suggests broader applicability. The connections between our approach and neural network models reveal a key distinction: while neural networks offer unlimited approximation capacity, stochastic simulation models have inherent structural constraints, making model inexactness unavoidable. We further discuss scalability, applicability, and the connection between our approach and neural network-based generative models in Section A.3.6, Section A.3.4 and Section A.1.4 of the Supplementary Material, respectively.

In future work, we plan to explore methods to reduce optimization cost, such as the weighted MMD approach (Bharti et al., 2023) for MMD estimation or using natural gradient descent (Briol et al., 2019). We also aim to explore avenues for simplifying our assumptions, particularly in the i.i.d. assumption and Assumption 2.3. Extending our results to non-i.i.d. scenarios presents a significant challenge, primarily due to the absence of a law of large numbers for generalized U-statistics in such settings. Moreover, many of the stringent conditions in Assumption 2.3 stem from Theorem A.20. One potential approach to simplify these assumptions is establishing asymptotic normality through the Hilbert space version of the central limit theorem (Oates et al., 2022).

Acknowledgements

The authors would like to thank the reviewers for their constructive, detailed feedback and in-depth discussions that greatly helped improve the paper, as well as the area chair for their dedication and hard work.

The authors would also like to thank Dr. Matthew Plumlee, who contributed greatly to the ideas in this paper. Ziwei Su also thanks Professor Barry Nelson, Professor Moses Chan, Taejong Joo, and Professor Imon Banerjee for insightful discussions on various aspects of the paper.

References

Asmussen, S. and Glynn, P. W. (2007). *Stochastic simulation: algorithms and analysis*, volume 57. Springer.

Bai, Y., Balch, T., Chen, H., Dervovic, D., Lam, H., and Vyetrenko, S. (2021). Calibrating over-parametrized simulation models: A framework via eligibility set. *arXiv preprint arXiv:2105.12893*.

Bai, Y. and Lam, H. (2020). Calibrating input parameters via eligibility sets. In *2020 Winter Simulation Conference (WSC)*, pages 2114–2125. IEEE.

Balci, O. and Sargent, R. G. (1982). Some examples of simulation model validation using hypothesis testing. Technical report, Institute of Electrical and Electronics Engineers (IEEE).

Baringhaus, L. and Franz, C. (2004). On a new multivariate two-sample test. *Journal of multivariate analysis*, 88(1):190–206.

Barton, R. R., Lam, H., and Song, E. (2022). Input Uncertainty in Stochastic Simulation. In Salhi, S. and Boylan, J., editors, *The Palgrave Handbook of Operations Research*, pages 573–620. Springer International Publishing, Cham.

Basu, A., Shioya, H., and Park, C. (2011). *Statistical inference: the minimum distance approach*. CRC press.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.

Bharti, A., Naslidnyk, M., Key, O., Kaski, S., and Briol, F.-X. (2023). Optimally-weighted estimators of the maximum mean discrepancy for likelihood-free inference. In *International Conference on Machine Learning*, pages 2289–2312. PMLR.

Biller, B., Jiang, X., Yi, J., Venditti, P., and Biller, S. (2022). Simulation: The critical technology in digital twin development. In *2022 Winter Simulation Conference (WSC)*, pages 1340–1355. IEEE.

Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. (2018). Demystifying MMD GANs. In *International Conference on Learning Representations*.

Briol, F.-X., Barp, A., Duncan, A. B., and Girolami, M. (2019). Statistical inference for generative models with maximum mean discrepancy. *arXiv preprint arXiv:1906.05944*.

Chambers, M. and Mount-Campbell, C. (2002). Process optimization via neural network metamodeling. *International Journal of Production Economics*, 79(2):93–100.

Chérief-Abdellatif, B.-E. and Alquier, P. (2020). Mmd-bayes: Robust bayesian estimation via maximum mean discrepancy. In *Symposium on Advances in Approximate Bayesian Inference*, pages 1–21. PMLR.

Chérief-Abdellatif, B.-E. and Alquier, P. (2022). Finite sample properties of parametric mmd estimation: robustness to misspecification and dependence. *Bernoulli*, 28(1):181–213.

Cranmer, K., Brehmer, J., and Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062.

Dawid, A. P. (2007). The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59:77–93.

Dawid, A. P. and Sebastiani, P. (1999). Coherent dispersion criteria for optimal experimental design. *Annals of Statistics*, pages 65–81.

Dellaporta, C., Knoblauch, J., Damoulas, T., and Briol, F.-X. (2022). Robust bayesian inference for simulator-based models via the mmd posterior bootstrap. In *International Conference on Artificial Intelligence and Statistics*, pages 943–970. PMLR.

Dellaporta, C., Knoblauch, J., Damoulas, T., and Briol, F.-X. (Accessed 04-17-2024). npl_mmd_project. `https://github.com/haritadell/npl_mmd_project`.

Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. (2015). Training generative neural networks via maximum mean discrepancy optimization. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 258–267.

Frazier, D. T. and Drovandi, C. (2021). Robust approximate bayesian inference with synthetic likelihood. *Journal of Computational and Graphical Statistics*, 30(4):958–976.

Frazier, D. T., Drovandi, C., and Loaiza-Maya, R. (2020a). Robust approximate bayesian computation: An adjustment approach. *arXiv preprint arXiv:2008.04099*.

Frazier, D. T., Drovandi, C., and Nott, D. J. (2021). Synthetic likelihood in misspecified models: Consequences and corrections. *arXiv preprint arXiv:2104.03436*.

Frazier, D. T., Robert, C. P., and Rousseau, J. (2020b). Model misspecification in approximate bayesian computation: consequences and diagnostics. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(2):421–444.

Frazier, P. I., Cashore, J. M., Duan, N., Henderson, S. G., Janmohamed, A., Liu, B., Shmoys, D. B., Wan, J., and Zhang, Y. (2022). Modeling for covid-19 college reopening decisions: Cornell, a case study. *Proceedings of the National Academy of Sciences*, 119(2):e2112532119.

Fu, M. C. (2015). *Stochastic gradient estimation*. Springer.

Fujisawa, M., Teshima, T., Sato, I., and Sugiyama, M. (2021). γ-abc: Outlier-robust approximate bayesian computation based on a robust divergence estimator. In *International Conference on Artificial Intelligence and Statistics*, pages 1783–1791. PMLR.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2007). Kernel measures of conditional dependence. *Advances in Neural Information Processing Systems*, 20.

Gao, R., Deistler, M., and Macke, J. H. (2023). Generalized bayesian inference for scientific simulators via amortized cost estimation. *Advances in Neural Information Processing Systems*, 36:80191–80219.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.

Goeva, A., Lam, H., Qian, H., and Zhang, B. (2019). Optimization-based calibration of simulation input models. *Operations Research*, 67(5):1362–1382.

Good, I. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 107–114.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.

Huang, D., Bharti, A., Souza, A., Acerbi, L., and Kaski, S. (2023). Learning robust statistics for simulation-based inference under model misspecification. *Advances in Neural Information Processing Systems*, 36:7289–7310.

Huber, P. J. (1992). Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer.

Kelly, R. P., Nott, D. J., Frazier, D. T., Warne, D. J., and Drovandi, C. (2023). Misspecification-robust sequential neural likelihood. *arXiv preprint arXiv:2301.13368*.

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464.

Kim, S., Shephard, N., and Chib, S. (1998). Stochastic volatility: likelihood inference and comparison with arch models. *The review of economic studies*, 65(3):361–393.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Kingma, D. P. and Welling, M. (2014). Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*.

Kleijnen, J. P. (1995). Verification and validation of simulation models. *European Journal of Operational Research*, 82(1):145–162.

LeRoy, B. and Shafer, C. (2021). Conformal prediction for simulation models. In *Proceedings of the 2021 ICML Workshop on Distribution-Free Uncertainty Quantification, Online*, volume 24.

Li, Y., Swersky, K., and Zemel, R. (2015). Generative moment matching networks. In *International Conference on Machine Learning*, pages 1718–1727. PMLR.

Lindley, D. V. (1952). The theory of queues with a single server. In *Mathematical proceedings of the Cambridge philosophical society*, volume 48, pages 277–289. Cambridge University Press.

Liu, F., Xu, W., Lu, J., Zhang, G., Gretton, A., and Sutherland, D. J. (2020). Learning deep kernels for non-parametric two-sample tests. In *International Conference on Machine Learning*, pages 6316–6326. PMLR.

Masserano, L., Dorigo, T., Izbicki, R., Kuusela, M., and Lee, A. B. (2022). Simulator-based inference with waldo: Confidence regions by leveraging prediction algorithms and posterior estimators for inverse problems. *arXiv preprint arXiv:2205.15680*.

Mohamed, S., Rosca, M., Figurnov, M., and Mnih, A. (2020). Monte carlo gradient estimation in machine learning. *Journal of Machine Learning Research*, 21(132):1–62.

Nelson, B. and Pei, L. (2013). *Foundations and methods of stochastic simulation*, volume 187. Springer.

Oates, C. et al. (2022). Minimum kernel discrepancy estimators. *arXiv preprint arXiv:2210.16357*.

Ojeda, C., Cvejoski, K., Georgiev, B., Bauckhage, C., Schuecker, J., and Sánchez, R. J. (2021). Learning deep generative models for queuing systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9214–9222.

Pacchiardi, L. and Dutta, R. (2021). Generalized bayesian likelihood-free inference using scoring rules estimators. *arXiv preprint arXiv:2104.03889*, 2(8).

Pacchiardi, L., Khoo, S., and Dutta, R. (2024). Generalized bayesian likelihood-free inference. *Electronic Journal of Statistics*, 18(2):3628–3686.

Pan, Y., Xu, Z., Guang, J., Chen, X., Dai, J., Wang, C., Zhang, X., Sun, J., Shi, P., Ding, Y., Wu, S., Yang, K., and Pan, H. (2021). A high-fidelity, machine-learning enhanced queueing network simulation model for hospital ultrasound operations. In *2021 Winter Simulation Conference (WSC)*, pages 1–12.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.

Plumlee, M. (2017). Bayesian calibration of inexact computer models. *Journal of the American Statistical Association*, 112(519):1274–1285.

Plumlee, M. (2019). Computer model calibration with confidence and consistency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(3):519–545.

Plumlee, M. and Lam, H. (2017). An uncertainty quantification method for inexact simulation models. *arXiv preprint arXiv:1707.06544*.

Polyak, B. T. and Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855.

Sargent, R. G. (2010). Verification and validation of simulation models. In *Proceedings of the 2010 Winter Simulation Conference*, pages 166–183. IEEE.

Schruben, L. W. (1980). Establishing the credibility of simulations. *Simulation*, 34(3):101–105.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013). Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The annals of statistics*, pages 2263–2291.

Sen, P. K. (1977). Almost sure convergence of generalized u-statistics. *The Annals of Probability*, pages 287–290.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007). A hilbert space embedding for distributions. In *International conference on algorithmic learning theory*, pages 13–31. Springer.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR.

Steinwart, I. and Christmann, A. (2008). *Support vector machines*. Springer Science & Business Media.

Storlie, C. B., Lane, W. A., Ryan, E. M., Gattiker, J. R., and Higdon, D. M. (2015). Calibration of computational models with categorical parameters and correlated outputs via bayesian smoothing spline anova. *Journal of the American Statistical Association*, 110(509):68–82.

Sutherland, D. J., Tung, H.-Y., Strathmann, H., De, S., Ramdas, A., Smola, A., and Gretton, A. (2017). Generative models and model criticism via optimized maximum mean discrepancy. In *International Conference on Learning Representations*.

Székely, G. J. and Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272.

Tarantola, A. (2005). *Inverse problem theory and methods for model parameter estimation*. SIAM.

Tuo, R. (2019). Adjustments to computer models via projected kernel calibration. *SIAM/ASA Journal on Uncertainty Quantification*, 7(2):553–578.

Tuo, R. and Wu, C. J. (2015). Efficient calibration for imperfect computer models. *The Annals of Statistics*, 43(6):2331–2352.

Tuo, R. and Wu, C. J. (2016). A theoretical framework for calibration in computer models: Parametrization, estimation and convergence properties. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):767–795.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Ward, D., Cannon, P., Beaumont, M., Fasiolo, M., and Schmon, S. (2022). Robust neural posterior estimation and statistical model criticism. *Advances in Neural Information Processing Systems*, 35:33845–33859.

Wehenkel, A., Behrmann, J., Miller, A. C., Sapiro, G., Sener, O., Cuturi, M., and Jacobsen, J.-H. (2023). Simulation-based inference for cardiovascular models. *arXiv preprint arXiv:2307.13918*.

Wehenkel, A., Gamella, J. L., Sener, O., Behrmann, J., Sapiro, G., Cuturi, M., and Jacobsen, J.-H. (2024). Addressing misspecification in simulation-based inference through data-driven calibration. *arXiv preprint arXiv:2405.08719*.

Zhang, Y., Lu, H., Zhou, Z., Yang, Z., and Xu, S. (2021). Analysis and optimisation of perishable inventory with stocks-sensitive stochastic demand and two-stage pricing: A discrete-event simulation study. *Journal of Simulation*, 15(4):326–337.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No] We will release all the code and data once the paper is accepted.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [No] We provide instructions to reproduce the results. We will release all the code and data once the paper is accepted.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable] We did not find any license information available.
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A SUPPLEMENTARY MATERIAL

A.1 Background and Preliminaries

A.1.1 G/G/1 Queueing System

We provide a brief introduction to G/G/1 queues for interested readers. In probability theory, the G/G/1 queue is a single-server queue with the first-in-first-out (FIFO) discipline, meaning that the customer who first enters the server is also the first to exit the server. In G/G/1 queues, the interarrival times of customers follow a general distribution, and service times follow a different general distribution. The dynamics of a G/G/1 queue are described by the Lindley equation (see Section A.3.2 for details).

The M/M/1 queue and M/G/1 queue are two special cases of the G/G/1 queue. In M/M/1 queues, both the interarrival times and the service times are exponentially distributed. In M/G/1 queues, the interarrival times are exponentially distributed, while the service times follow a general distribution.

A.1.2 Energy Distance and Maximum Mean Discrepancy (MMD)

The associated divergence function of the energy score is the energy distance (Baringhaus and Franz, 2004; Székely and Rizzo, 2013)

$$\mathcal{E}^{(\beta)}(P,Q) = \mathbb{E}_{X,Y}\|X-Y\|^{\beta} - \frac{1}{2}\mathbb{E}_{X,X'}\|X-X'\|^{\beta} - \frac{1}{2}\mathbb{E}_{Y,Y'}\|Y-Y'\|^{\beta},$$

where $X, X' \overset{\text{i.i.d.}}{\sim} Q$.

The associated divergence function of the kernel score is the squared maximum mean discrepancy (Gretton et al., 2012):

$$\mathrm{MMD}_k^2(P,Q) = \mathbb{E}_{X,X'}\left[k\left(X,X'\right)\right] - 2\mathbb{E}_{X,Y}[k(X,Y)] + \mathbb{E}_{Y,Y'}\left[k\left(Y,Y'\right)\right].$$

Sejdinovic et al. (2013) establishes the equivalence of the energy distance and the MMD. The kernel score is strictly proper, and the MMD is a metric on the space of Borel probability measures such that $\mathbb{E}_Y\sqrt{k(Y,Y)} < \infty$, denoted $\mathcal{P}_k$, if the kernel k is *characteristic* (Fukumizu et al., 2007).

A positive-definite, measurable kernel k is said to be characteristic if the *kernel mean embedding* $\mu_k : \mathcal{P}_k \to \mathcal{H}_k$ (Smola et al., 2007) is injective, where $\mathcal{H}_k$ is the reproducing kernel Hilbert space (RKHS) associated with k and equipped with inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}_k}$ and norm $\|\cdot\|_{\mathcal{H}_k}$.

The kernel mean embedding μ_k is a continuous map that embeds a probability measure P into an element $\mu_k^P = \mathbb{E}_Y\left[k(\cdot,Y)\right]$ of the RKHS $\mathcal{H}_k$. The MMD can be expressed as the distance between kernel mean embeddings in $\mathcal{H}_k$: $\mathrm{MMD}_k(P,Q) = \|\mu_k^P - \mu_k^Q\|_{\mathcal{H}_k}$ (Gretton et al., 2012). As a result, when k is characteristic (i.e. μ_k is injective), $\|\mu_k^P - \mu_k^Q\|_{\mathcal{H}_k} = 0$ implies $P = Q$, the MMD is a metric on $\mathcal{P}_k$ and the kernel score is strictly proper.

A.1.3 Model Inexactness versus Model Contamination

This section expands on the distinction between model contamination and model inexactness introduced in the main paper, providing additional technical context for these two fundamental types of model discrepancy.

Model Contamination In model contamination scenarios, the underlying model specification is correct, but the observed data contain noise, outliers, or other forms of contamination. Formally, the classical Huber's contamination model (Huber, 1992) is given as follows:

$$X_1,\ldots,X_m \sim (1-\epsilon)P_\star + \epsilon Q, \epsilon \in (0,1),$$

where Q represents the distribution of the contamination, and $P_\star$ represents the true data-generating distribution that belongs to the model class $\{P_\theta : \theta \in \Theta\}$, $P_\star = P_{\theta_0}, \theta_0 \in \Theta$. The statistical challenge is estimating θ_0 while being robust to the presence of contaminated observations.

Model contamination has been extensively studied in robust statistics. In the context of simulation-based inference, several Bayesian approaches like γ-ABC (Fujisawa et al., 2021) and MMD posterior bootstrap (Dellaporta

et al., 2022) have been developed to address contamination. Section A.4.4 presents additional experiments on contamination models with our confidence set procedure.

We also note that queueing and contamination models are distinct concepts. Queueing models are a specific class of stochastic simulation models, while any class of models can be contamination models if the observed data are contaminated.

Model Inexactness In contrast, model inexactness (or model misspecification) occurs when the observed data is uncontaminated, but the model class $\{P_\theta : \theta \in \Theta\}$ is incapable of perfectly representing the true data-generating distribution $P_\star$. Formally, $P_\star \notin \{P_\theta : \theta \in \Theta\}$. This means that there exists no parameter $\theta \in \Theta$ such that $P_\theta = P_\star$. The challenge becomes to find the "best possible" parameter $\theta_\star$ that minimizes some statistical distance between P_θ and $P_\star$.

Model inexactness is inherent in many scientific domains where simplifying assumptions is necessary for computational feasibility. For instance, queueing models often assume specific parametric families for service time distributions that may not perfectly capture real-world behavior.

Experimental Designs to Study Different Types of Discrepancy In our experimental evaluation, we designed distinct approaches to study these two types of discrepancy:

1. Model Inexactness: In Experiments 2 and 4, we systematically vary the shape parameter a of the service time distribution in the target system while maintaining a fixed parametric family in the simulation model. As a decreases from 1.0 to 0.2, the structural mismatch between the target system and simulation model increases, allowing us to evaluate performance under increasing model inexactness.

2. Model Contamination: In Section A.4.4 of the Supplementary Material, we conduct an additional experiment where we add white noise from $N(0, 0.01)$ to varying percentages ($\epsilon = 0.01, 0.05, 0.1, 0.2, 0.5$) of the output data, simulating different levels of data contamination.

These experimental designs allow us to separately evaluate the performance of our method under each type of discrepancy. While our primary focus is on model inexactness, understanding performance under contamination provides a more comprehensive assessment of robustness. The distinction between these two types of discrepancy is crucial for proper method selection and evaluation. Methods designed for contamination may not perform well under model inexactness, and vice versa, as they address fundamentally different statistical challenges.

A.1.4 Stochastic Simulation Models versus Deep Generative Models

The key difference between deep generative models and stochastic simulation models lies in the deterministic structure of the latter, determined by target system mechanics. For instance, if the target system is an email server queue governed by a Poisson process for arrivals and unknown service time distributions, the stochastic simulation model would be an M/G/1 queueing model. Unlike deep generative models, which can be considered as well-specified (justified by universal approximation theorems), altering the structure of stochastic simulation models, such as increasing network depth, is not applicable to enhance their approximation capabilities.

Consider the previous instance of modeling an email server queue. Given the unknown service time distribution of the target system, the model service time distribution likely does not align with that of the target system for any simulation parameter θ. As a result, the stochastic simulation model is inherently inexact. One way to enhance approximation capabilities of stochastic simulation models is through meta-modeling, where service and arrival time distributions are modeled using deep neural networks rather than fixed distribution families typically used in queueing networks. We further discuss meta-modeling in the last paragraph of Section A.3.2.

A.2 Technical Framework and Properties

A.2.1 Useful Assumptions

The following assumptions are those summarized in Assumption 2.1.

Assumption A.1 (Finite moments). $\mathbb{E}_X\|X\|^\alpha, \mathbb{E}_Z\|Z\|^\alpha < \infty$ for some $\alpha \geq 1$.

Assumption A.1 corresponds to Assumption A in Bińkowski et al. (2018).

Assumption A.2. There exists constants C_0, C_1 such that

$$|k(x,y)| \leq C_0 \left(\left(\|x\|^2 + \|y\|^2 \right)^{\alpha/2} + 1 \right)$$
$$\|\nabla_{x,y} k(x,y)\| \leq C_1 \left(\left(\|x\|^2 + \|y\|^2 \right)^{(\alpha-1)/2} + 1 \right),$$

where α is the same α in Assumption A.1.

The following two conditions (Assumption C and D in Bińkowski et al. (2018)) are imposed on the non-parametric components of G_θ.

Assumption A.3 (Lipschitzness)**.** Each non-parametric component of G_θ, ϕ_i, is M-Lipschitz.

Assumption A.4 (Piecewise analytic)**.** For ϕ_i, there are K_i functions ϕ_i^k, $k = 1, \ldots, K_i$, each real analytic on the input space $\mathbb{R}^{d_{\phi_i}}$, where d_{ϕ_i} denotes the dimension of the input space, and agrees with ϕ_i on the closure of a set $\mathcal{D}^k$: $\phi_i(x) = \phi_i^k(x) \quad \forall x \in \overline{\mathcal{D}}_i^k$. The sets $\mathcal{D}_i^k$ are disjoint, and cover the whole input space: $\bigcup_{k=1}^{K_i} \overline{\mathcal{D}}_i^k = \mathbb{R}^{d_\pi}$. Each $\mathcal{D}_i^k$ is defined by $S_{i,k}$ real analytic functions $G_{i,k,s} : \mathbb{R}^{d_{\phi_i}} \to \mathbb{R}$ as

$$\mathcal{D}_i^k = \left\{ x \in \mathbb{R}^{d_{\phi_i}} \mid G_{i,k,s}(x) > 0 \quad \forall s = 1, \ldots, S_{i,k} \right\}.$$

Assumption A.5 (Differentiability)**.** Each parametric component of G_θ, $\psi_{\theta,j}$, is almost everywhere differentiable with respect to θ.

The following regularity conditions are summarized in Assumption A.14.

Assumption A.6 (Separability)**.** The reproducing kernel Hilbert space $\mathcal{H}_k$ is separable.

Assumption A.7. $\mathbb{E}_X \sqrt{k(X,X)} < \infty$, where $X \sim P_\star$.

The following regularity conditions are summarized in Assumption 2.3.

Assumption A.8. There exists a open and convex neighbourhood of $\theta_\star^{\mathrm{KS}}$ denoted $\mathcal{N}$ such that $\mathbb{E}_{Z,X} \sup_{\theta \in \mathcal{N}} \|\nabla_\theta k(G_\theta(Z), X)\| < \infty$ and $\mathbb{E}_{Z,Z'} \sup_{\theta \in \mathcal{N}} \|\nabla_\theta k(G_\theta(Z), k(G_\theta(Z')\| < \infty$ for all $\theta \in \mathcal{N}$.

Assumption A.9. $\mathbb{E}_{Z,X} \left\|\nabla_\theta k(G_\theta(Z), X)\right\|^2 |_{\theta=\theta_\star^{\mathrm{KS}}}, \mathbb{E}_{Z,Z'} \left\|\nabla_\theta k(G_\theta(Z), k(G_\theta(Z')\right\|^2 |_{\theta=\theta_\star^{\mathrm{KS}}} < \infty$.

Assumption A.10. $\mathbb{E}_{Z,X} \left[\left| \nabla_{\theta_r \theta_s} k(G_\theta(Z), X) |_{\theta=\theta_\star^{\mathrm{KS}}} \right| \log^+ \left(\left| \nabla_{\theta_r \theta_s} k(G_\theta(Z), X) |_{\theta=\theta_\star^{\mathrm{KS}}} \right| \right) \right]$,
$\mathbb{E}_{Z,Z'} \left[\left| \nabla_{\theta_r \theta_s} k(G_\theta(Z), G_\theta(Z')) |_{\theta=\theta_\star^{\mathrm{KS}}} \right| \right] < \infty$ for all $r, s \in \{1, \ldots, p\}$, where $\theta = (\theta_1, \ldots, \theta_p) \in \mathbb{R}^p$.

Assumption A.11. $\nabla_\theta k(G_\theta(z), x)$, $\nabla_\theta k(G_\theta(z), G_\theta(z'))$ are differentiable on $\mathcal{N}$ for all P-almost all $z, z' \in Z$ and $P_\star$-almost all $x \in \mathbb{R}^d$.

Assumption A.12. $\nabla_{\theta\theta} k(G_\theta(z), x)$, $\nabla_{\theta\theta} k(G_\theta(z), G_\theta(z'))$ are uniformly continuous at $\theta_\star^{\mathrm{KS}}$ for all P-almost all $z, z' \in Z$ and $P_\star$-almost all $x \in \mathbb{R}^d$, where $\nabla_{\theta\theta}(\cdot)$ denotes the Hessian with respect to θ

Assumption A.13. The Hessian $H = \mathbb{E}_{Z,Z'} \left[\nabla_{\theta\theta} k(G_\theta(Z), G_\theta(Z')) |_{\theta=\theta_\star^{\mathrm{KS}}} \right] - 2\mathbb{E}_{Z,X} \left[\nabla_{\theta\theta} k(G_\theta(Z), X) |_{\theta=\theta_\star^{\mathrm{KS}}} \right]$ is positive definite.

A.2.2 Known Results

When a stochastic simulation model is identifiable, the (kernel) optimum score estimator converges to the optimal simulation parameter almost surely. Oates et al. (2022) provides the following regularity conditions.

Assumption A.14 (Separability and Kernel Bound)**.** The reproducing kernel Hilbert space $\mathcal{H}_k$ is separable, and $\mathbb{E}_X \sqrt{k(X,X)} < \infty$, where $X \sim P_\star$.

Note that $\mathcal{H}_k$ is separable if the kernel k corresponding to the RKHS $\mathcal{H}_k$ is continuous and its input space is separable (Steinwart and Christmann, 2008). Since the input space $\mathbb{R}^d$ considered here is separable, $\mathcal{H}_k$ is separable for all continuous kernels on $\mathbb{R}^d$, which include the Gaussian, Laplacian, and Riesz kernel. Condition $\mathbb{E}_X \sqrt{k(X,X)} < \infty$ ensures that $P_\star$ is in the space of Borel probability measures on which the corresponding MMD of kernel k is a metric.

Proposition A.15 (Strong consistency (Oates et al., 2022)). *If Assumption A.14 hold, and the optimal simulation parameter $\theta_\star^{\mathrm{KS}}$ is unique, then $\widehat{\theta}_m^{\mathrm{KS}} \xrightarrow{\text{a.s.}} \theta_\star^{\mathrm{KS}}$, where $\xrightarrow{\text{a.s.}}$ denotes almost sure convergence.*

Proposition A.15 provides the convergence guarantee for the kernel optimum score estimator without the speed of convergence.

Proposition A.16 (Strong consistency under model non-identifiability (Oates et al., 2022)). *If Assumptions A.6 and A.7 hold, and the divergence function $d_{\mathrm{KS}}(P_\theta, P_\star)$ is continuous on Θ, then any accumulation point of $\{\widehat{\theta}_m^{\mathrm{KS}}\}_{m\in\mathbb{N}}$ is almost surely an element of $\arg\min_{\theta\in\Theta} d_{\mathrm{KS}}(P_\theta, P_\star)$.*

Note that Theorem 10 in Oates et al. (2022) gives the sufficient conditions for the existence of accumulation points of the sequence $\{\widehat{\theta}_m^{\mathrm{KS}}\}_{m\in\mathbb{N}}$.

The generalization bounds for MMD under bounded kernel require the following assumption (Briol et al., 2019).

Assumption A.17. 1. For every $Q \in \mathcal{P}_k$, there exists $c > 0$ such that the set $\{\theta \in \Theta : \mathrm{MMD}_k(P_\theta, Q) \leq \inf_{\theta'\in\Theta} \mathrm{MMD}_k(P_{\theta'}, Q) + c\}$ is bounded.

2. For every $n \in \mathbb{N}$ and $Q \in \mathcal{P}_k$, there exists $c_n > 0$ such that the set $\{\theta \in \Theta : \mathrm{MMD}_k\left(\widehat{P}_\theta^n, Q\right) \leq \inf_{\theta'\in\Theta} \mathrm{MMD}_k\left(\widehat{P}_{\theta'}^n, Q\right) + c_n\}$ is bounded, where $\widehat{P}_\theta^n$ and $\widehat{P}_{\theta'}^n$ are empirical measures.

Theorem A.18 (Generalization bounds (Briol et al., 2019)). *Suppose that the kernel k is bounded, and Assumption A.17 holds, then with probability at least $1-\delta$,*

$$\mathrm{MMD}\left(P_{\widehat{\theta}_m^{\mathrm{KS}}}, Q\right) \leq \inf_{\theta\in\Theta} \mathrm{MMD}\left(P_\theta, Q\right) + 2\sqrt{\frac{2}{m}\sup_{x\in\mathbb{R}^d} k(x,x)}\left(2 + \sqrt{\log\left(\frac{1}{\delta}\right)}\right),$$

and

$$\mathrm{MMD}\left(P_{\widehat{\theta}_{m,n}^{\mathrm{KS}}}, Q\right) \leq \inf_{\theta\in\Theta} \mathrm{MMD}\left(P_\theta, Q\right) + 2\left(\sqrt{\frac{2}{n}} + \sqrt{\frac{2}{m}}\right)\sqrt{\sup_{x\in\mathbb{R}^d} k(x,x)}\left(2 + \sqrt{\log\left(\frac{2}{\delta}\right)}\right).$$

The following results are used in proving Theorem 2.4.

Proposition A.19 (Differentiation lemma (Bińkowski et al., 2018)). *Let $\Theta \in \mathbb{R}^p$ be a non-trivial open set, Q be a probability measure on $\mathbb{R}^m$ and V is an m-dimensional random variable such that $V \sim Q$. Define a map $h : \mathbb{R}^m \times \Theta \mapsto \mathbb{R}^n$ with the following properties:*

1. *For any $\theta \in \Theta, \mathbb{E}_V\left[\|h_\theta(V)\|\right] < \infty$.*
2. *For Q-almost all $v \in \mathbb{R}^m$, the map $\Theta \to \mathbb{R}^n, \theta \mapsto h_\theta(v)$ is differentiable.*
3. *There exists a Q-integrable function $g : \mathbb{R}^m \mapsto \mathbb{R}$ such that $\|\nabla_\theta h_\theta(v)\| \leq g(v)$ for all $\theta \in \Theta$.*

Then, for any $\theta \in \Theta, \mathbb{E}_V\|\nabla_\theta h_\theta(V)\| < \infty$ and the function $\theta \mapsto \mathbb{E}_V[h_\theta(V)]$ is differentiable with differential

$$\nabla_\theta \mathbb{E}_V\left[h_\theta(V)\right] = \mathbb{E}_V\left[\nabla_\theta h_\theta(V)\right].$$

Theorem A.20 (Strong law of large numbers for generalized k-sample U-statistics (Sen, 1977)). *Let $\{X_{i,j}, i = 1,\ldots,k, j = 1,\ldots,n_i\}$ be a sequence of independent and identically distributed d-dimensional random variables with each $X_{i,j}$ having measure P_i. For $m = (m_1,\ldots,m_k)$, consider the generalized U-statistic*

$$U(n) = \prod_{i=1}^k \begin{pmatrix} n_i \\ m_i \end{pmatrix}^{-1} \sum \phi\left(X_{i,j_1},\ldots,X_{i,j_{m_i}}, i = 1,\ldots,k\right),$$

where $m_i \leq n_i$ is the degree vector and the summation $\sum$ extends over all possible $1 \leq j_1 < \ldots < j_{m_i} \leq n_i, i = 1,\ldots,k$. Suppose that $\mathbb{E}_{X_{i,j_1},\ldots,X_{i,j_{m_i}}}\left[\left|\phi\left(X_{i,j_1},\ldots,X_{i,j_{m_i}}\right)\right|\left(\log^+\left|\phi\left(X_{i,j_1},\ldots,X_{i,j_{m_i}}\right)\right|\right)^{k-1}\right] < \infty$, where $\log^+(x) = \max(0, \log(x))$. Then, as $n_1,\ldots,n_k \to \infty$,

$$U(n) \xrightarrow{\text{a.s.}} \mathbb{E}_{X_{i,j_1},\ldots,X_{i,j_{m_i}}}\left[\phi\left(X_{i,j_1},\ldots,X_{i,j_{m_i}}\right)\right].$$

A.2.3 Optimal and True Simulation Parameter

The following proposition gives the equivalence of the optimal simulation parameter and the true simulation parameter under model exactness and identifiability.

Proposition A.21. *Suppose that the stochastic simulation model is exact and identifiable. If the scoring rule S is strictly proper relative to $\mathcal{P}$, and $\{P_\theta : \theta \in \Theta\} \subset \mathcal{P}$, then $\theta_\star = \theta_0$.*

Proof. We prove by contradiction assuming $\theta_\star \neq \theta_0$. By uniqueness of θ_0, $P_\star \neq P_{\theta_\star}$. Since S is strictly proper, $\mathbb{E}_{X\sim P_\star} S(P_\star, X) < \mathbb{E}_{X\sim P_\star} S(P_{\theta^\star}, X)$. Thus if $\theta_\star \neq \theta_0$, we have

$$\mathbb{E}_{X\sim P_\star} S(P_\star, X) < \mathbb{E}_{X\sim P_\star} S(P_{\theta_\star}, X) = L_\star(\theta_\star) \leq L_\star(\theta_0) = \mathbb{E}_{X\sim P_\star} S(P_{\theta_0}, X) = \mathbb{E}_{X\sim P_\star} S(P_\star, X).$$

□

A.2.4 Comparison of Asymptotic Normality with Prior Work in Bayesian and Frequentist Frameworks

We want to emphasize the key distinctions between our asymptotic normality result (Theorem 2.4) and related work in both Bayesian and frequentist frameworks.

Comparison with Bayesian Approaches Our result differs fundamentally from Bayesian posterior consistency results such as Theorem 2 in Pacchiardi et al. (2024). While both approaches quantify parameter uncertainty, they operate within different statistical paradigms:

1. Our frequentist approach establishes that the distribution of the kernel optimum score estimator converges to a normal distribution centered at the optimal simulation parameter. In contrast, Pacchiardi et al. (2024) show that their posterior distribution concentrates around the optimal simulation parameter within a Bayesian framework.
2. Our result provides asymptotically correct coverage for our proposed confidence set procedure. This property is unattainable with posterior consistency results from Bayesian approaches. As Pacchiardi et al. (2024) explicitly note in Section 3.1 of their paper, their credible sets do not provide correct frequentist coverage, even under ideal conditions with strictly proper scoring rules and well-specified models.

Comparison with Frequentist Approaches under Model Exactness Our work also extends beyond Briol et al. (2019), which establishes asymptotic normality for kernel optimum score estimators under model exactness. Key theoretical innovations in our approach include:

1. Different Statistical Tools: Model inexactness requires us to employ the strong law of large numbers for generalized k-sample U-statistics (Theorem A.20) rather than standard U-statistics theory (Van der Vaart, 2000). This necessitates additional technical conditions, specifically Assumption A.10 and the uniform continuity of the Hessian (Assumption A.12), that are not required in the exact model case.
2. Structural Differences in Target System Output Data: Under model inexactness, observations from the target system $(X_1, \ldots, X_m)$ lack the push-forward representation that exists in exact models. In other words, there exists no G_θ and Z such that $X = G_\theta(Z)$. This fundamental difference manifests in our U-statistic U_2 defined in Section A.2.5, requiring new conditions on the kernel function $k(G_\theta(Z), X)$ (Assumptions A.8, A.9 and A.11) that were unnecessary in previous work focused solely on $k(G_\theta(Z), G_\theta(Z'))$.
3. Weaker Differentiability Requirements: Our approach establishes asymptotic normality under less restrictive conditions, notably removing the requirement for trice differentiability of G_θ present in Briol et al. (2019).
4. Validated Confidence Set Construction: We provide the first result of consistent variance estimation in Theorem 2.6, which fully justifies the asymptotic validity of our confidence sets under model inexactness.

These technical advances allow us to develop the first complete asymptotic theory for kernel-based confidence sets in simulation models with inherent model inexactness, addressing a fundamental challenge in practical simulation calibration.

A.2.5 Proof of Theorem 2.4

Proof. We shall only prove the asymptotic normality of the kernel simulated score estimator $\widehat{\theta}^{\mathrm{KS}}_{m,n}$, as the proof for the asymptotic normality of the kernel optimum score estimator $\widehat{\theta}^{\mathrm{KS}}_{m}$ follows an entirely analogous way.

The first order optimality condition implies that $\nabla_\theta \widehat{L}^{\mathrm{KS}}_{m,n}(\theta) \mid_{\theta=\widehat{\theta}^{\mathrm{KS}}_{m,n}} = 0$. As $\widehat{\theta}^{\mathrm{KS}}_{m,n} \xrightarrow{\text{a.s.}} \theta^{\mathrm{KS}}_\star$ as $m,n \to \infty$, $\widehat{\theta}^{\mathrm{KS}}_{m,n} \in \mathcal{N}$ a.s. for sufficiently large m,n. Then the mean value theorem implies that there a.s. exists a $\tilde{\theta}^{\mathrm{KS}}_{m,n} \in \mathcal{N}$ between $\theta^{\mathrm{KS}}_\star$ and $\widehat{\theta}^{\mathrm{KS}}_{m,n}$ such that

$$0 = \nabla_\theta \widehat{L}^{\mathrm{KS}}_{m,n}(\theta) \mid_{\theta=\widehat{\theta}^{\mathrm{KS}}_{m,n}} = \nabla_\theta \widehat{L}^{\mathrm{KS}}_{m,n}(\theta) \mid_{\theta=\theta^{\mathrm{KS}}_\star} + \nabla_{\theta\theta} \widehat{L}^{\mathrm{KS}}_{m,n}(\theta) \mid_{\theta=\tilde{\theta}^{\mathrm{KS}}_{m,n}} (\widehat{\theta}^{\mathrm{KS}}_{m,n} - \theta^{\mathrm{KS}}_\star).$$

If $\nabla_{\theta\theta} \widehat{L}^{\mathrm{KS}}_{m,n}(\theta) \mid_{\theta=\tilde{\theta}^{\mathrm{KS}}_{m,n}}$ is non-singular, $\sqrt{m+n}(\widehat{\theta}^{\mathrm{KS}}_{m,n} - \theta^{\mathrm{KS}}_\star) = -\nabla_{\theta\theta} \widehat{L}^{\mathrm{KS}}_{m,n}(\theta)^{-1} \mid_{\theta=\tilde{\theta}^{\mathrm{KS}}_{m,n}} \cdot \sqrt{m+n} \nabla_\theta \widehat{L}^{\mathrm{KS}}_{m,n}(\theta) \mid_{\theta=\theta^{\mathrm{KS}}_\star}$.

We first show that $\nabla_{\theta\theta} \widehat{L}^{\mathrm{KS}}_{m,n}(\theta) \mid_{\theta=\tilde{\theta}^{\mathrm{KS}}_{m,n}} \xrightarrow{\mathrm{p}} H$. We prove the stronger almost sure convergence. Observe that

$$\nabla_{\theta\theta} \widehat{L}^{\mathrm{KS}}_{m,n}(\theta) \mid_{\theta=\tilde{\theta}^{\mathrm{KS}}_{m,n}} = \nabla_{\theta\theta} \widehat{L}^{\mathrm{KS}}_{m,n}(\theta) \mid_{\theta=\theta^{\mathrm{KS}}_\star} + \nabla_{\theta\theta} \widehat{L}^{\mathrm{KS}}_{m,n}(\theta) \mid_{\theta=\tilde{\theta}^{\mathrm{KS}}_{m,n}} - \nabla_{\theta\theta} \widehat{L}^{\mathrm{KS}}_{m,n}(\theta) \mid_{\theta=\theta^{\mathrm{KS}}_\star}, \tag{4}$$

where

$$\nabla_{\theta\theta} \widehat{L}^{\mathrm{KS}}_{m,n}(\theta) \mid_{\theta=\theta^{\mathrm{KS}}_\star} = \frac{1}{n(n-1)} \sum_{1\le i\ne j\le n} \nabla_{\theta\theta} k(G_\theta(Z_i), G_\theta(Z_j)) \mid_{\theta=\theta^{\mathrm{KS}}_\star} - \frac{2}{mn} \sum_{i=1}^{m}\sum_{j=1}^{n} \nabla_{\theta\theta} k(G_\theta(Z_j), X_i) \mid_{\theta=\theta^{\mathrm{KS}}_\star}.$$

Recall $\theta = (\theta_1, \ldots, \theta_p) \in \mathbb{R}^p$. As Assumption A.10 holds, the strong law of large numbers for U-statistics (Theorem A.20) implies that

$$\frac{1}{n(n-1)} \sum_{1\le i\ne j\le n} \nabla_{\theta_r\theta_s} k(G_\theta(Z_i), G_\theta(Z_j)) \mid_{\theta=\theta^{\mathrm{KS}}_\star} \xrightarrow{\text{a.s.}} \mathbb{E}_{Z,Z'} \nabla_{\theta_r\theta_s} k(G_\theta(Z), G_\theta(Z')) \mid_{\theta=\theta^{\mathrm{KS}}_\star},$$

and

$$\frac{2}{mn} \sum_{i=1}^{m}\sum_{j=1}^{n} \nabla_{\theta_r\theta_s} k(G_\theta(Z_j), X_i) \mid_{\theta=\theta^{\mathrm{KS}}_\star} \xrightarrow{\text{a.s.}} 2\mathbb{E}_{Z,X} \nabla_{\theta_r\theta_s} k(G_\theta(Z), X) \mid_{\theta=\theta^{\mathrm{KS}}_\star}.$$

Therefore,

$$\nabla_{\theta_r\theta_s} \widehat{L}^{\mathrm{KS}}_{m,n}(\theta) \mid_{\theta=\theta^{\mathrm{KS}}_\star} \xrightarrow{\text{a.s.}} \mathbb{E}_{Z,Z'} \nabla_{\theta_r\theta_s} k(G_\theta(Z), G_\theta(Z')) \mid_{\theta=\theta^{\mathrm{KS}}_\star} - 2\mathbb{E}_{Z,X} \nabla_{\theta_r\theta_s} k(G_\theta(Z), X) \mid_{\theta=\theta^{\mathrm{KS}}_\star}$$

for all $r,s \in \{1,\ldots,p\}$. Then $\nabla_{\theta\theta} \widehat{L}^{\mathrm{KS}}_{m,n}(\theta) \mid_{\theta=\theta^{\mathrm{KS}}_\star} \xrightarrow{\text{a.s.}} H$.

Now we just need to show that $\left\| \nabla_{\theta\theta} \widehat{L}^{\mathrm{KS}}_{m,n}(\theta) \mid_{\theta=\tilde{\theta}^{\mathrm{KS}}_{m,n}} - \nabla_{\theta\theta} \widehat{L}^{\mathrm{KS}}_{m,n}(\theta) \mid_{\theta=\theta^{\mathrm{KS}}_\star} \right\| \xrightarrow{\mathrm{p}} 0$.

Assumption A.12 implies that for any $\epsilon > 0$ there exists an open neighborhood $\mathcal{N}_\epsilon \subset \mathcal{N}$ such that

$$\left\| \nabla_{\theta\theta} k(G_\theta(z), x) - \nabla_{\theta\theta} k(G_\theta(z), x) \mid_{\theta=\theta^{\mathrm{KS}}_\star} \right\|, \left\| \nabla_{\theta\theta} k(G_\theta(z), G_\theta(z')) - \nabla_{\theta\theta} k(G_\theta(z), G_\theta(z')) \mid_{\theta=\theta^{\mathrm{KS}}_\star} \right\| < \epsilon$$

for all $\theta \in \mathcal{N}_\epsilon$, $z, z' \in \mathcal{Z}, x \in \mathbb{R}^d$. Since $\widehat{\theta}^{\mathrm{KS}}_{m,n} \xrightarrow{\text{a.s.}} \theta^{\mathrm{KS}}_\star$ as $m,n \to \infty$, $\tilde{\theta}^{\mathrm{KS}}_{m,n} \xrightarrow{\text{a.s.}} \theta^{\mathrm{KS}}_\star$ as $m,n \to \infty$, then $\tilde{\theta}^{\mathrm{KS}}_{m,n} \in \mathcal{N}_\epsilon$ a.s. for sufficiently large m,n. Therefore,

$$\left\| \nabla_{\theta\theta} k(G_\theta(z), x) \mid_{\theta=\tilde{\theta}^{\mathrm{KS}}_{m,n}} - \nabla_{\theta\theta} k(G_\theta(z), x) \mid_{\theta=\theta^{\mathrm{KS}}_\star} \right\|,$$
$$\left\| \nabla_{\theta\theta} k(G_\theta(z), G_\theta(z')) \mid_{\theta=\tilde{\theta}^{\mathrm{KS}}_{m,n}} - \nabla_{\theta\theta} k(G_\theta(z), G_\theta(z')) \mid_{\theta=\theta^{\mathrm{KS}}_\star} \right\| < \epsilon$$

a.s. for all $z, z' \in \mathcal{Z}, x \in \mathbb{R}^d$, which in turn implies that

$$\begin{aligned}
&\left\|\nabla_{\theta\theta}\widehat{L}^{\mathrm{KS}}_{m,n}(\theta)\mid_{\theta=\tilde{\theta}^{\mathrm{KS}}_{m,n}} - \nabla_{\theta\theta}\widehat{L}^{\mathrm{KS}}_{m,n}(\theta)\mid_{\theta=\theta^{\mathrm{KS}}_{\star}}\right\| \\
=&\left\| \frac{1}{n(n-1)} \sum_{1\le i\neq j\le n} \left(\nabla_{\theta\theta} k(G_\theta(Z_i), G_\theta(Z_j))\mid_{\theta=\tilde{\theta}^{\mathrm{KS}}_{m,n}} - k(G_\theta(Z_i), G_\theta(Z_j))\mid_{\theta=\theta^{\mathrm{KS}}_{\star}}\right)\right. \\
&\qquad \left. -\frac{2}{mn}\left(\sum_{i=1}^{m}\sum_{j=1}^{n} \nabla_{\theta\theta} k(G_\theta(Z_j), X_i)\mid_{\theta=\tilde{\theta}^{\mathrm{KS}}_{m,n}} - k(G_\theta(Z_j), X)\mid_{\theta=\theta^{\mathrm{KS}}_{\star}}\right)\right\| \\
\le& \frac{1}{n(n-1)} \sum_{1\le i\neq j\le n} \left\|\nabla_{\theta\theta} k(G_\theta(Z_i), G_\theta(Z_j))\mid_{\theta=\tilde{\theta}^{\mathrm{KS}}_{m,n}} - k(G_\theta(Z_i), G_\theta(Z_j))\mid_{\theta=\theta^{\mathrm{KS}}_{\star}}\right\| \\
&\qquad + \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} \left\|\nabla_{\theta\theta} k(G_\theta(Z_j), X_i)\mid_{\theta=\tilde{\theta}^{\mathrm{KS}}_{m,n}} - k(G_\theta(Z_j), X)\mid_{\theta=\theta^{\mathrm{KS}}_{\star}}\right\| \\
<& 3\epsilon.
\end{aligned}$$

Therefore, $\left\|\nabla_{\theta\theta}\widehat{L}^{\mathrm{KS}}_{m,n}(\theta)\mid_{\theta=\tilde{\theta}^{\mathrm{KS}}_{m,n}} - \nabla_{\theta\theta}\widehat{L}^{\mathrm{KS}}_{m,n}(\theta)\mid_{\theta=\theta^{\mathrm{KS}}_{\star}}\right\| \xrightarrow{\text{a.s.}} 0$. Given $\nabla_{\theta\theta}\widehat{L}^{\mathrm{KS}}_{m,n}(\theta)\mid_{\theta=\theta^{\mathrm{KS}}_{\star}} \xrightarrow{\text{a.s.}} H$, (4) implies that $\nabla_{\theta\theta}\widehat{L}^{\mathrm{KS}}_{m,n}(\theta)\mid_{\theta=\tilde{\theta}^{\mathrm{KS}}_{m,n}} \xrightarrow{\text{a.s.}} H$ as $m, n \to \infty$. Since H is positive-definite from Assumption A.13, $\nabla_{\theta\theta}\widehat{L}^{\mathrm{KS}}_{m,n}(\theta)\mid_{\theta=\tilde{\theta}^{\mathrm{KS}}_{m,n}}$ is non-singular for sufficiently large m, n.

We now show that $\sqrt{m+n}\nabla_\theta\widehat{L}^{\mathrm{KS}}_{m,n}(\theta)\mid_{\theta=\theta^{\mathrm{KS}}_{\star}} \xrightarrow{\mathrm{d}} N(0, C_\lambda)$. Observe that

$$\begin{aligned}
\nabla_\theta\widehat{L}^{\mathrm{KS}}_{m,n}(\theta)\mid_{\theta=\theta^{\mathrm{KS}}_{\star}} = &\frac{1}{n(n-1)} \sum_{1\le i\neq j\le n} \nabla_\theta k(G_\theta(Z_i), G_\theta(Z_j))\mid_{\theta=\theta^{\mathrm{KS}}_{\star}} \\
&- \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} \nabla_\theta k(G_\theta(Z_j), X_i)\mid_{\theta=\theta^{\mathrm{KS}}_{\star}}.
\end{aligned}$$

Denote

$$U_1 = \frac{1}{n(n-1)} \sum_{1\le i\neq j\le n} \nabla_\theta k(G_\theta(Z_i), G_\theta(Z_j))\mid_{\theta=\theta^{\mathrm{KS}}_{\star}}$$

and

$$U_2 = \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} \nabla_\theta k(G_\theta(Z_j), X_i)\mid_{\theta=\theta^{\mathrm{KS}}_{\star}}.$$

Both U_1 and U_2 are U-statistics. Assumption A.8 implies that there exists integrable functions $b_1 : Z \times \mathbb{R}^d \to \mathbb{R}$ and $b_2 : Z \times Z \to \mathbb{R}$ such that $\|\nabla_\theta k(G_\theta(Z), X)\| \le b_1(Z, X)$ and $\|\nabla_\theta k(G_\theta(Z), G_\theta(Z'))\| \le b_2(Z, Z')$, $\mathbb{E}_{Z,X}|b_1(Z, X)|, \mathbb{E}_{Z,Z'}|b_2(Z, Z')| < \infty$.

As Assumptions A.9 and A.11 hold, Proposition A.19 implies that

$$\mathbb{E}_{Z_1,\ldots,Z_n,X}\nabla_\theta\widehat{L}^{\mathrm{KS}}_{m,n}(\theta)\mid_{\theta=\theta^{\mathrm{KS}}_{\star}} = \nabla_\theta\mathbb{E}_{Z_1,\ldots,Z_n,X}\widehat{L}^{\mathrm{KS}}_{m,n}(\theta)\mid_{\theta=\theta^{\mathrm{KS}}_{\star}} = \nabla_\theta \operatorname{MMD}_k(P_\theta, P_\star)\mid_{\theta=\theta^{\mathrm{KS}}_{\star}}. \tag{5}$$

By the first order optimality condition, $\nabla_\theta \operatorname{MMD}_k(P_\theta, P_\star)\mid_{\theta=\theta^{\mathrm{KS}}_{\star}} = 0$.

Therefore, $\mathbb{E}_{Z_1,\ldots,Z_n,X}\nabla_\theta\widehat{L}^{\mathrm{KS}}_{m,n}(\theta)\mid_{\theta=\theta^{\mathrm{KS}}_{\star}} = 0$, which implies that

$$\begin{aligned}
\mathbb{E}U_1 &= \mathbb{E}_{Z_1,\ldots,Z_n}\frac{1}{n(n-1)} \sum_{1\le i\neq j\le n} \nabla_\theta k(G_\theta(Z_i), G_\theta(Z_j))\mid_{\theta=\theta^{\mathrm{KS}}_{\star}} \\
&= \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} \mathbb{E}_{Z_1,\ldots,Z_n,X}\nabla_\theta k(G_\theta(Z_j), X_i)\mid_{\theta=\theta^{\mathrm{KS}}_{\star}} \\
&= \mathbb{E}U_2.
\end{aligned}$$

Now denote $\mathbb{E}U_1 = \mathbb{E}U_2$ by M. We need to show that $\sqrt{n}(U_1 - \widehat{U}_1) \xrightarrow{\text{p}} 0$ and $\sqrt{m+n}(U_2 - \widehat{U}_2) \xrightarrow{\text{p}} 0$, where $\widehat{U}_1 = \frac{2}{n}\sum_{j=1}^{n} h(Z_j), \widehat{U}_2 = \frac{1}{m}\sum_{i=1}^{m} g_{1,0}(X_i) + \frac{1}{n}\sum_{j=1}^{n} g_{0,1}(Z_j)$ are the Hájek projections of U_1 and U_2, respectively. As $\sqrt{m+n}(U_1 - U_2) = \sqrt{m+n}\left(\widehat{U}_1 - \widehat{U}_2 + (U_1 - \widehat{U}_1) + (U_2 - \widehat{U}_2)\right)$, the asymptotic distribution of $\sqrt{m+n}(U_1 - U_2)$ is the same as the asymptotic distribution of $\sqrt{m+n}(\widehat{U}_1 - \widehat{U}_2)$ if $\sqrt{n}(U_1 - \widehat{U}_1) \xrightarrow{\text{p}} 0$ and $\sqrt{m+n}(U_2 - \widehat{U}_2) \xrightarrow{\text{p}} 0$. We only prove $\sqrt{m+n}(U_2 - \widehat{U}_2) \xrightarrow{\text{p}} 0$ as the proof of $\sqrt{n}(U_1 - \widehat{U}_1) \xrightarrow{\text{p}} 0$ follows an entirely analogous way. We prove the stronger L_2 convergence result $\mathbb{E}\left[\left(\sqrt{m+n}(U_2 - \widehat{U}_2)\right)^T \left(\sqrt{m+n}(U_2 - \widehat{U}_2)\right)\right] \to 0$ as $m, n \to \infty$. Trivial algebra gives that $\mathbb{E}\widehat{U}_1 = \mathbb{E}\widehat{U}_2 = M$. Then $\mathbb{E}(U_2 - \widehat{U}_2) = 0$, and hence

$$
\begin{aligned}
&\mathbb{E}\left[\left(\sqrt{m+n}(U_2 - \widehat{U}_2)\right)^T \left(\sqrt{m+n}(U_2 - \widehat{U}_2)\right)\right] \\
=&(m+n)\mathbb{E}\left[\left(U_2 - \widehat{U}_2\right)^T \left(U_2 - \widehat{U}_2\right)\right] \\
=&(m+n)\operatorname{tr}\left(\mathbb{C}\left[U_2 - \widehat{U}_2\right]\right) \\
=&(m+n)\operatorname{tr}\left(\mathbb{C}\left[U_2\right] + \mathbb{C}\left[\widehat{U}_2\right] - 2\mathbb{C}\left[U_2, \widehat{U}_2\right]\right).
\end{aligned}
$$

As Assumption A.10 holds, Theorem 12.6 in Van der Vaart (2000) implies that both $(m+n)\operatorname{tr}\left(\mathbb{C}\left[U_2\right]\right)$ and $(m+n)\operatorname{tr}\left(\mathbb{C}\left[\widehat{U}_2\right]\right)$ converges to $\frac{\operatorname{tr}(\mathbb{C}_Z[g_{0,1}(Z)])}{\lambda} + \frac{\operatorname{tr}(\mathbb{C}_X[g_{1,0}(X)])}{1-\lambda}$. We just need to show that $(m+n)\operatorname{tr}\left(\mathbb{C}\left[U_2, \widehat{U}_2\right]\right)$ converges to $\frac{\operatorname{tr}(\mathbb{C}_Z[g_{0,1}(Z)])}{\lambda} + \frac{\operatorname{tr}(\mathbb{C}_X[g_{1,0}(X)])}{1-\lambda}$ as well.

By definition,

$$
\begin{aligned}
&(m+n)\operatorname{tr}\left(\mathbb{C}\left[U_2, \widehat{U}_2\right]\right) \\
=&(m+n)\operatorname{tr}\left(\mathbb{E}\left[\left(\frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\nabla_\theta k(G_\theta(Z_j), X_i)\mid_{\theta=\theta^{\text{KS}}_\star} - M\right)^T \right.\right. \\
&\qquad\qquad\left.\left.\left(\frac{1}{m}\sum_{i=1}^{m} g_{1,0}(X_i) + \frac{1}{n}\sum_{j=1}^{n} g_{0,1}(Z_j) - M\right)\right]\right) \\
=&\operatorname{tr}\left(\frac{m+n}{m}\mathbb{E}_{X_1,\ldots,X_m}\left[\sqrt{m}\left(\frac{1}{m}\sum_{i=1}^{m} g_{1,0}(X_i) - \frac{1}{2}M\right)^T \sqrt{m}\left(\frac{1}{m}\sum_{i=1}^{m} g_{1,0}(X_i) - \frac{1}{2}M\right)\right]\right. \\
&\qquad\left.+\frac{m+n}{n}\mathbb{E}_{Y_1,\ldots,Y_n}\left[\sqrt{n}\left(\frac{1}{n}\sum_{j=1}^{n} g_{0,1}(Z_j) - \frac{1}{2}M\right)^T \sqrt{n}\left(\frac{1}{n}\sum_{j=1}^{n} g_{0,1}(Z_j) - \frac{1}{2}M\right)\right]\right) \\
\to&\frac{\operatorname{tr}\left(\mathbb{C}_Z\left[g_{0,1}(Z)\right]\right)}{\lambda} + \frac{\operatorname{tr}\left(\mathbb{C}_X\left[g_{1,0}(X)\right]\right)}{1-\lambda},
\end{aligned}
$$

leveraging the classical central limit theorem. Hence, we have $\sqrt{n}(U_1 - \widehat{U}_1) \xrightarrow{\text{p}} 0$ and $\sqrt{m+n}(U_2 - \widehat{U}_2) \xrightarrow{\text{p}} 0$ as $m, n \to \infty$.

Plug in the expressions of $\widehat{U}_1$ and $\widehat{U}_2$ to derive

$$\begin{aligned}\sqrt{m+n}(\widehat{U}_1-\widehat{U}_2) &= \sqrt{m+n}\left(\frac{2}{n}\sum_{j=1}^{n}h_1(Z_j)-\frac{1}{m}\sum_{i=1}^{m}g_{1,0}(X_i)-\frac{1}{n}\sum_{j=1}^{n}g_{0,1}(Z_j)\right)\\ &= \sqrt{m+n}\left(\frac{1}{n}\sum_{j=1}^{n}(2h_1(Z_j)-g_{0,1}(Z_j))-\frac{1}{m}\sum_{i=1}^{m}g_{1,0}(X_i)\right)\\ &= \sqrt{\frac{m+n}{n}}\sqrt{n}\left(\frac{1}{n}\sum_{j=1}^{n}(2h_1(Z_j)-g_{0,1}(Z_j))-\frac{1}{2}M\right)\\ &\quad -\sqrt{\frac{m+n}{m}}\sqrt{m}\left(\frac{1}{m}\sum_{i=1}^{m}g_{1,0}(X_i)-\frac{1}{2}M\right).\end{aligned}$$

As Assumption A.10 holds, the classical central limit theorem implies that

$$\begin{aligned}\sqrt{n}\left(\frac{1}{n}\sum_{j=1}^{n}(2h_1(Z_j)-g_{0,1}(Z_j))-\frac{1}{2}M\right) &\xrightarrow{\mathrm{d}} N\left(0,\mathbb{C}_Z\left[2h_1(Z)-g_{0,1}(Z)\right]\right)\\ \sqrt{m}\left(\frac{1}{m}\sum_{i=1}^{m}g_{1,0}(X_i)-\frac{1}{2}M\right) &\xrightarrow{\mathrm{d}} N\left(0,\mathbb{C}_X\left[g_{1,0}(X)\right]\right).\end{aligned}$$

Furthermore, as $\lim_{m,n\to\infty}\frac{n}{m+n}=\lambda$, X_i and Z_j are independent, we get

$$\sqrt{m+n}(\widehat{U}_1-\widehat{U}_2)\xrightarrow{\mathrm{d}} N\left(0,\frac{\mathbb{C}_Z\left[2h_1(Z)-g_{0,1}(Z)\right]}{\lambda}+\frac{\mathbb{C}_X\left[g_{1,0}(X)\right]}{1-\lambda}\right)=N(0,\Sigma_\lambda).$$

Hence $\sqrt{m+n}\nabla_\theta\widehat{L}^{\mathrm{KS}}_{m,n}(\theta)\mid_{\theta=\theta^{\mathrm{KS}}_\star}=\sqrt{m+n}(U_1-U_2)\xrightarrow{\mathrm{d}}N(0,\Sigma_\lambda)$. As $\nabla_{\theta\theta}\widehat{L}^{\mathrm{KS}}_{m,n}(\theta)\mid_{\theta=\tilde{\theta}^{\mathrm{KS}}_{m,n}}\xrightarrow{\mathrm{p}}H$, the Slutsky's theorem implies that $\sqrt{m+n}(\widehat{\theta}^{\mathrm{KS}}_{m,n}-\theta^{\mathrm{KS}}_\star)=-\nabla_{\theta\theta}\widehat{L}^{\mathrm{KS}}_{m,n}(\theta)^{-1}\mid_{\theta=\tilde{\theta}^{\mathrm{KS}}_{m,n}}\cdot\sqrt{m+n}\nabla_\theta\widehat{L}^{\mathrm{KS}}_{m,n}(\theta)\mid_{\theta=\theta^{\mathrm{KS}}_\star}\xrightarrow{\mathrm{d}}N(0,H^{-1}\Sigma_\lambda H^{-1})$. □

A.2.6 Proof of Theorem 2.6

Proof. To simplify notation, n in this proof is the simulation sample size for confidence set estimation that equals to n_c in Theorem 2.6. We first show that $\widehat{H}_{m,n}\xrightarrow{\mathrm{p}}H$ as $m,n\to\infty$. Recall that $\widehat{H}_{m,n}=\nabla_{\theta\theta}\widehat{L}^{\mathrm{KS}}_{m,n}(\theta)\mid_{\theta=\widehat{\theta}^{\mathrm{KS}}_m}$. The proof follows an entirely analogous way as the proof of $\nabla_{\theta\theta}\widehat{L}^{\mathrm{KS}}_{m,n}(\theta)\mid_{\theta=\tilde{\theta}^{\mathrm{KS}}_{m,n}}\xrightarrow{\mathrm{p}}H$ in A.2.5.

As Assumption A.10 holds, leveraging the strong law of large numbers for generalized k-sample U-statistics (Theorem A.20) gives $\nabla_{\theta\theta}\widehat{L}^{\mathrm{KS}}_{m,n}(\theta)\mid_{\theta=\theta^{\mathrm{KS}}_\star}\xrightarrow{\mathrm{a.s.}}H$. We just need to show that $\left\|\nabla_{\theta\theta}\widehat{L}^{\mathrm{KS}}_{m,n}(\theta)\mid_{\theta=\widehat{\theta}^{\mathrm{KS}}_{m,n}}-\nabla_{\theta\theta}\widehat{L}^{\mathrm{KS}}_{m,n}(\theta)\mid_{\theta=\theta^{\mathrm{KS}}_\star}\right\|\xrightarrow{\mathrm{a.s.}}0$.

Assumption A.12 implies that for any $\epsilon>0$ there exists an open neighborhood $\mathcal{N}_\epsilon\subset\mathcal{N}$ such that

$$\begin{aligned}&\left\|\nabla_{\theta\theta}k(G_\theta(z),x)-\nabla_{\theta\theta}k(G_\theta(z),x)\mid_{\theta=\theta^{\mathrm{KS}}_\star}\right\|,\\ &\left\|\nabla_{\theta\theta}k(G_\theta(z),G_\theta(z'))-\nabla_{\theta\theta}k(G_\theta(z),G_\theta(z'))\mid_{\theta=\theta^{\mathrm{KS}}_\star}\right\|<\epsilon\end{aligned}$$

for all $\theta\in\mathcal{N}_\epsilon$, $z,z'\in\mathcal{Z},x\in\mathbb{R}^d$. Since $\widehat{\theta}^{\mathrm{KS}}_{m,n}\xrightarrow{\mathrm{a.s.}}\theta^{\mathrm{KS}}_\star$ as $m,n\to\infty$, then $\widehat{\theta}^{\mathrm{KS}}_{m,n}\in\mathcal{N}_\epsilon$ a.s. for sufficiently large m,n. Therefore,

$$\begin{aligned}&\left\|\nabla_{\theta\theta}k(G_\theta(z),x)\mid_{\theta=\widehat{\theta}^{\mathrm{KS}}_{m,n}}-\nabla_{\theta\theta}k(G_\theta(z),x)\mid_{\theta=\theta^{\mathrm{KS}}_\star}\right\|,\\ &\left\|\nabla_{\theta\theta}k(G_\theta(z),G_\theta(z'))\mid_{\theta=\widehat{\theta}^{\mathrm{KS}}_{m,n}}-\nabla_{\theta\theta}k(G_\theta(z),G_\theta(z'))\mid_{\theta=\theta^{\mathrm{KS}}_\star}\right\|<\epsilon\end{aligned}$$

a.s. for all $z, z' \in \mathcal{Z}, x \in \mathbb{R}^d$, which in turn implies that

$$\left\| \nabla_{\theta\theta} \widehat{L}^{\mathrm{KS}}_{m,n}(\theta) \mid_{\theta=\widehat{\theta}^{\mathrm{KS}}_{m,n}} - \nabla_{\theta\theta} \widehat{L}^{\mathrm{KS}}_{m,n}(\theta) \mid_{\theta=\theta^{\mathrm{KS}}_{\star}} \right\| < 3\epsilon.$$

Therefore, $\left\| \nabla_{\theta\theta} \widehat{L}^{\mathrm{KS}}_{m,n}(\theta) \mid_{\theta=\widehat{\theta}^{\mathrm{KS}}_{m,n}} - \nabla_{\theta\theta} \widehat{L}^{\mathrm{KS}}_{m,n}(\theta) \mid_{\theta=\theta^{\mathrm{KS}}_{\star}} \right\| \xrightarrow{\text{a.s.}} 0$.

Now we show that $\widehat{\Sigma}_{m,n} \xrightarrow{\text{p}} \Sigma$ as $m, n \to \infty$. We slightly abuse notation to define $\widehat{\Sigma}_{m,n}(\theta)$ as a function of θ:

$$\widehat{\Sigma}_{m,n}(\theta) = \frac{4}{m-1} \sum_{i=1}^{m} \left(\widehat{\mu}_i(\theta) - \widehat{\mu}_{m,n}(\theta)\right)^T \left(\widehat{\mu}_i(\theta) - \widehat{\mu}_{m,n}(\theta)\right),$$

where

$$\widehat{\mu}_{m,n}(\theta) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \nabla_\theta k\left(G_\theta(Z_j), X_i\right)$$
$$\widehat{\mu}_i(\theta) = \frac{1}{n} \sum_{j=1}^{n} \nabla_\theta k\left(G_\theta(Z_j), X_i\right).$$

Then $\widehat{\Sigma}_{m,n}(\widehat{\theta}^{\mathrm{KS}}_{m,n}) = \widehat{\Sigma}_{m,n}$. We first prove that $\widehat{\Sigma}_{m,n}(\theta^{\mathrm{KS}}_{\star}) \xrightarrow{\text{p}} \Sigma$. By definition,

$$\begin{aligned}
&\widehat{\Sigma}_{m,n}(\theta^{\mathrm{KS}}_{\star}) \\
=&\frac{4}{m-1} \sum_{i=1}^{m} \left(\widehat{\mu}_i(\theta^{\mathrm{KS}}_{\star}) - \widehat{\mu}_{m,n}(\theta^{\mathrm{KS}}_{\star})\right)^T \left(\widehat{\mu}_i(\theta^{\mathrm{KS}}_{\star}) - \widehat{\mu}_{m,n}(\theta^{\mathrm{KS}}_{\star})\right) \\
=&\frac{4}{m-1} \sum_{i=1}^{m} \left(\widehat{\mu}_i(\theta^{\mathrm{KS}}_{\star})^T \widehat{\mu}_i(\theta^{\mathrm{KS}}_{\star}) - \widehat{\mu}_{m,n}(\theta^{\mathrm{KS}}_{\star})^T \widehat{\mu}_i(\theta^{\mathrm{KS}}_{\star}) - \widehat{\mu}_i(\theta^{\mathrm{KS}}_{\star})^T \widehat{\mu}_{m,n}(\theta^{\mathrm{KS}}_{\star}) + \widehat{\mu}_{m,n}(\theta^{\mathrm{KS}}_{\star})^T \widehat{\mu}_{m,n}(\theta^{\mathrm{KS}}_{\star})\right) \\
=&\frac{4}{(m-1)n^2} \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{l=1}^{n} \nabla_\theta k\left(G_\theta(Z_j), X_i\right)^T \mid_{\theta=\theta^{\mathrm{KS}}_{\star}} \nabla_\theta k\left(G_\theta(Z_l), X_i\right) \mid_{\theta=\theta^{\mathrm{KS}}_{\star}} \\
&\quad - \frac{4}{m(m-1)n^2} \sum_{i=1}^{m} \sum_{p=1}^{m} \sum_{j=1}^{n} \sum_{l=1}^{n} \nabla_\theta k\left(G_\theta(Z_j), X_i\right)^T \mid_{\theta=\theta^{\mathrm{KS}}_{\star}} \nabla_\theta k\left(G_\theta(Z_l), X_p\right) \mid_{\theta=\theta^{\mathrm{KS}}_{\star}}.
\end{aligned}$$

As Assumption A.9 holds, the strong law of large numbers for generalized k-sample U-statistics (Theorem A.20) implies that as $m, n \to \infty$,

$$\begin{aligned}
&\frac{4}{(m-1)n^2} \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{l=1}^{n} \nabla_\theta k\left(G_\theta(Z_j), X_i\right)^T \mid_{\theta=\theta^{\mathrm{KS}}_{\star}} \nabla_\theta k\left(G_\theta(Z_l), X_i\right) \mid_{\theta=\theta^{\mathrm{KS}}_{\star}} \xrightarrow{\text{a.s.}} \\
&\qquad 4\mathbb{E}_{X,Z,Z'}\left[\nabla_\theta k\left(G_\theta(Z), X\right)^T \mid_{\theta=\theta^{\mathrm{KS}}_{\star}} \nabla_\theta k\left(G_\theta(Z'), X\right) \mid_{\theta=\theta^{\mathrm{KS}}_{\star}}\right] \\
&\frac{4}{m(m-1)n^2} \sum_{i=1}^{m} \sum_{p=1}^{m} \sum_{j=1}^{n} \sum_{l=1}^{n} \nabla_\theta k\left(G_\theta(Z_j), X_i\right)^T \mid_{\theta=\theta^{\mathrm{KS}}_{\star}} \nabla_\theta k\left(G_\theta(Z_l), X_p\right) \mid_{\theta=\theta^{\mathrm{KS}}_{\star}} \xrightarrow{\text{a.s.}} \\
&\qquad 4\mathbb{E}_{X,X',Z,Z'}\left[\nabla_\theta k\left(G_\theta(Z), X\right)^T \mid_{\theta=\theta^{\mathrm{KS}}_{\star}} \nabla_\theta k\left(G_\theta(Z'), X'\right) \mid_{\theta=\theta^{\mathrm{KS}}_{\star}}\right],
\end{aligned}$$

where $X, X' \overset{\text{i.i.d.}}{\sim} P_\star$, $Z, Z' \overset{\text{i.i.d.}}{\sim} P$. Therefore,

$$
\begin{aligned}
&\widehat{\Sigma}_{m,n}(\theta_\star^{\mathrm{KS}}) \\
\overset{\text{a.s.}}{\to} &4\mathbb{E}_{X,Z,Z'}\left[\nabla_\theta k\left(G_\theta(Z),X\right)^T \mid_{\theta=\theta_\star^{\mathrm{KS}}} \nabla_\theta k\left(G_\theta(Z'),X\right) \mid_{\theta=\theta_\star^{\mathrm{KS}}}\right] \\
&\quad - 4\mathbb{E}_{X,X',Z,Z'}\left[\nabla_\theta k\left(G_\theta(Z),X\right)^T \mid_{\theta=\theta_\star^{\mathrm{KS}}} \nabla_\theta k\left(G_\theta(Z'),X'\right) \mid_{\theta=\theta_\star^{\mathrm{KS}}}\right] \\
=&4\mathbb{E}_{X,X'}\left[\mathbb{E}_Z\left[\nabla_\theta k\left(G_\theta(Z),X\right)^T \mid_{\theta=\theta_\star^{\mathrm{KS}}}\right]\left(\mathbb{E}_Z\left[\nabla_\theta k\left(G_\theta(Z),X\right) \mid_{\theta=\theta_\star^{\mathrm{KS}}}\right] - \mathbb{E}_Z\left[\nabla_\theta k\left(G_\theta(Z),X'\right) \mid_{\theta=\theta_\star^{\mathrm{KS}}}\right]\right)\right] \\
=&4\mathbb{E}_X\left[\mathbb{E}_Z\left[\nabla_\theta k\left(G_\theta(Z),X\right)^T \mid_{\theta=\theta_\star^{\mathrm{KS}}}\right]\left(\mathbb{E}_Z\left[\nabla_\theta k\left(G_\theta(Z),X\right) \mid_{\theta=\theta_\star^{\mathrm{KS}}}\right] - \mathbb{E}_{X,Z}\left[\nabla_\theta k\left(G_\theta(Z),X\right) \mid_{\theta=\theta_\star^{\mathrm{KS}}}\right]\right)\right] \\
=&\mathbb{C}_X\left[\mathbb{E}_Z \nabla_\theta k\left(G_\theta(Z),X\right) \mid_{\theta=\theta_\star^{\mathrm{KS}}}\right] \\
=&\Sigma.
\end{aligned}
$$

Now we just need to show that $\left\|\widehat{\Sigma}_{m,n}(\theta_\star^{\mathrm{KS}}) - \widehat{\Sigma}_{m,n}\right\| \xrightarrow{\text{p}} 0$. We only prove that

$$
\left\|\frac{4}{m-1}\sum_{i=1}^{m}\left[\widehat{\mu}_i(\theta_\star^{\mathrm{KS}})^T\widehat{\mu}_i(\theta_\star^{\mathrm{KS}}) - \widehat{\mu}_i(\widehat{\theta}_{m,n}^{\mathrm{KS}})^T\widehat{\mu}_i(\widehat{\theta}_{m,n}^{\mathrm{KS}})\right]\right\| \overset{\text{a.s.}}{\to} 0,
$$

as the almost sure convergence of all the other components of $\widehat{\Sigma}_{m,n}(\theta_\star^{\mathrm{KS}}) - \widehat{\Sigma}_{m,n}$ follows an entirely analogous way.

By definition,

$$
\begin{aligned}
&\left\|\frac{4}{m-1}\sum_{i=1}^{m}\left[\widehat{\mu}_i(\theta_\star^{\mathrm{KS}})^T\widehat{\mu}_i(\theta_\star^{\mathrm{KS}}) - \widehat{\mu}_i(\widehat{\theta}_{m,n}^{\mathrm{KS}})^T\widehat{\mu}_i(\widehat{\theta}_{m,n}^{\mathrm{KS}})\right]\right\| \\
=&\frac{4}{m-1}\left\|\sum_{i=1}^{m}\left[\frac{1}{n}\sum_{j=1}^{n}\nabla_\theta k\left(G_\theta(Z_j),X_i\right)^T \mid_{\theta=\theta_\star^{\mathrm{KS}}} \cdot \frac{1}{n}\sum_{l=1}^{n}\nabla_\theta k\left(G_\theta(Z_l),X_i\right) \mid_{\theta=\theta_\star^{\mathrm{KS}}}\right.\right. \\
&\qquad\left.\left. -\frac{1}{n}\sum_{j=1}^{n}\nabla_\theta k\left(G_\theta(Z_j),X_i\right)^T \mid_{\theta=\widehat{\theta}_{m,n}^{\mathrm{KS}}} \cdot \frac{1}{n}\sum_{l=1}^{n}\nabla_\theta k\left(G_\theta(Z_l),X_i\right) \mid_{\theta=\widehat{\theta}_{m,n}^{\mathrm{KS}}}\right]\right\| \\
=&\frac{4}{(m-1)n^2}\left\|\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{n}\left[\nabla_\theta k\left(G_\theta(Z_j),X_i\right)^T \mid_{\theta=\theta_\star^{\mathrm{KS}}} \nabla_\theta k\left(G_\theta(Z_l),X_i\right) \mid_{\theta=\theta_\star^{\mathrm{KS}}}\right.\right. \\
&\qquad\left.\left. -\nabla_\theta k\left(G_\theta(Z_j),X_i\right)^T \mid_{\theta=\widehat{\theta}_{m,n}^{\mathrm{KS}}} \nabla_\theta k\left(G_\theta(Z_l),X_i\right) \mid_{\theta=\widehat{\theta}_{m,n}^{\mathrm{KS}}}\right]\right\| \\
\leq&\frac{4}{(m-1)n^2}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{n}\left\|\nabla_\theta k\left(G_\theta(Z_j),X_i\right)^T \mid_{\theta=\theta_\star^{\mathrm{KS}}} \nabla_\theta k\left(G_\theta(Z_l),X_i\right) \mid_{\theta=\theta_\star^{\mathrm{KS}}}\right. \\
&\qquad\left. -\nabla_\theta k\left(G_\theta(Z_j),X_i\right)^T \mid_{\theta=\widehat{\theta}_{m,n}^{\mathrm{KS}}} \nabla_\theta k\left(G_\theta(Z_l),X_i\right) \mid_{\theta=\widehat{\theta}_{m,n}^{\mathrm{KS}}}\right\| \\
\leq&\frac{4}{(m-1)n^2}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{n}\left\{\left\|\left(\nabla_\theta k\left(G_\theta(Z_j),X_i\right)^T \mid_{\theta=\theta_\star^{\mathrm{KS}}} - \nabla_\theta k\left(G_\theta(Z_j),X_i\right)^T \mid_{\theta=\widehat{\theta}_{m,n}^{\mathrm{KS}}}\right)\right.\right. \\
&\qquad\left.\cdot\nabla_\theta k\left(G_\theta(Z_l),X_i\right) \mid_{\theta=\theta_\star^{\mathrm{KS}}}\right\| \\
&+\left.\left\|\nabla_\theta k\left(G_\theta(Z_j),X_i\right)^T \mid_{\theta=\widehat{\theta}_{m,n}^{\mathrm{KS}}}\left(\nabla_\theta k\left(G_\theta(Z_l),X_i\right) \mid_{\theta=\theta_\star^{\mathrm{KS}}} - \nabla_\theta k\left(G_\theta(Z_l),X_i\right) \mid_{\theta=\widehat{\theta}_{m,n}^{\mathrm{KS}}}\right)\right\|\right\} \\
=&\frac{4}{(m-1)n^2}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{n}\left\{\left\|\nabla_\theta k\left(G_\theta(Z_l),X_i\right) \mid_{\theta=\theta_\star^{\mathrm{KS}}}\right\|\left\|\nabla_\theta k\left(G_\theta(Z_j),X_i\right) \mid_{\theta=\theta_\star^{\mathrm{KS}}}\right.\right. \\
&\qquad\left.\left. - \nabla_\theta k\left(G_\theta(Z_j),X_i\right) \mid_{\theta=\widehat{\theta}_{m,n}^{\mathrm{KS}}}\right\|\right. \\
&+\left.\left\|\nabla_\theta k\left(G_\theta(Z_j),X_i\right) \mid_{\theta=\widehat{\theta}_{m,n}^{\mathrm{KS}}}\right\|\left\|\nabla_\theta k\left(G_\theta(Z_l),X_i\right) \mid_{\theta=\theta_\star^{\mathrm{KS}}} - \nabla_\theta k\left(G_\theta(Z_l),X_i\right) \mid_{\theta=\widehat{\theta}_{m,n}^{\mathrm{KS}}}\right\|\right\}.
\end{aligned} \tag{6}
$$

As Assumption 2.5 holds, for any $\epsilon > 0$ there exists an open neighborhood $\mathcal{N}_\epsilon \subset \mathcal{N}$ such that

$$\left\| \nabla_\theta k\left(G_\theta(z), x\right) \mid_{\theta=\theta_\star^{\mathrm{KS}}} - \nabla_\theta k\left(G_\theta(z), x\right) \right\| < \epsilon,$$

and there exists a non-negative constant M such that

$$\left\| \nabla_\theta k\left(G_\theta(z), x\right) \mid_{\theta=\theta_\star^{\mathrm{KS}}} \right\| < M$$

for all $x \in \mathbb{R}^d$ and $z \in \mathcal{Z}$. Since $\widehat{\theta}_{m,n}^{\mathrm{KS}} \xrightarrow{\text{a.s.}} \theta_\star^{\mathrm{KS}}$ as $m, n \to \infty$, then $\widehat{\theta}_{m,n}^{\mathrm{KS}} \in \mathcal{N}_\epsilon$ a.s. for sufficiently large m, n. Then for sufficiently large m, n, the right-hand side of (6) denoted RHS a.s. has

$$\mathrm{RHS} \leq \frac{4}{(m-1)n^2} \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{l=1}^{n} 2M\epsilon \leq 8M\epsilon.$$

Therefore, we conclude that $\left\| \frac{4}{m-1} \sum_{i=1}^{m} \left[\widehat{\mu}_i(\theta_\star^{\mathrm{KS}})^T \widehat{\mu}_i(\theta_\star^{\mathrm{KS}}) - \widehat{\mu}_i(\widehat{\theta}_{m,n}^{\mathrm{KS}})^T \widehat{\mu}_i(\widehat{\theta}_{m,n}^{\mathrm{KS}}) \right] \right\| \xrightarrow{\text{a.s.}} 0$. Making the analogous arguments for all the other components of $\widehat{\Sigma}_{m,n}(\theta_\star^{\mathrm{KS}}) - \widehat{\Sigma}_{m,n}$ gives $\left\| \widehat{\Sigma}_{m,n}(\theta_\star^{\mathrm{KS}}) - \widehat{\Sigma}_{m,n} \right\| \xrightarrow{\text{p}} 0$. Hence $\widehat{\Sigma}_{m,n} \xrightarrow{\text{p}} \Sigma$ as $m, n \to \infty$. Similarly, $\widehat{C}_{m,n} = \widehat{H}_{m,n}^{-1} \widehat{\Sigma}_{m,n} \widehat{H}_{m,n}^{-1} \xrightarrow{\text{p}} H^{-1} \Sigma H^{-1}$ as $m, n \to \infty$ follows from the continuity of matrix inversion and multiplication. □

A.3 Implementation and Practical Considerations

A.3.1 Optimal Sample Size and Kernel Selection

Briol et al. (2019) discuss the choice of the simulation sample size n under model exactness by finding the best ratio $\lambda = \frac{n}{m+n}$ in terms of computational efficiency. Theorem 2 in Briol et al. (2019) establishes asymptotic variance-covariance of the kernel simulated score estimator $\widehat{\theta}_{m,n}^{\mathrm{KS}}$ to be $\frac{C}{\lambda(1-\lambda)}$. The asymptotic variance-covariance is hence minimized at $\lambda = \frac{1}{2}$ (i.e. the size of the simulated sample n equals that of the output data from the target system m). This means that using n much larger than m is not computationally efficient.

The choice of optimal λ in terms of computational efficiency under model inexactness is notably intricate. Recall that the asymptotic variance-covariance of the kernel simulated score estimator is $C_\lambda = H^{-1} \Sigma_\lambda H^{-1}$, where $\Sigma_\lambda = \frac{\mathbb{C}_Z[2h(Z) - g_{0,1}(Z)]}{\lambda} + \frac{\mathbb{C}_X[g_{1,0}(X)]}{1-\lambda}$. In addition to being dependent on λ and the unknown $\theta_\star^{\mathrm{KS}}$, Σ_λ is influenced by factors such as the chosen kernel, the reference measure of Z, the distribution of X (i.e., the distribution of the target system output), and the input model G_θ. Identifying the optimal λ is thus highly instance-dependent and computationally intractable.

Furthermore, Σ_λ represents a trade-off between the variance-covariance term arising from the simulated sample Z and that arising from data X. Let $c_1 = \mathbb{C}_Z\left[2h(Z) - g_{0,1}(Z)\right]$ and $c_2 = \mathbb{C}_X\left[g_{1,0}(X)\right]$. In the case where $p = 1$ (i.e., the parameter $\theta \in \Theta$ is one-dimensional), simple algebra yields that the asymptotic variance C_λ is minimized at $\lambda = \frac{\sqrt{c_1}}{\sqrt{c_1} + \sqrt{c_2}}$. When $c_1 = c_2$, including the model exactness case, $\lambda = \frac{1}{2}$. Heuristically, when the distributions of X and $G_\theta(Z) \mid_{\theta=\theta_\star^{\mathrm{KS}}} = Y(\theta_\star^{\mathrm{KS}})$ are relatively close (i.e., the model discrepancy is small), λ is close to $\frac{1}{2}$. Therefore, using n much larger than m is not computationally efficient. However, when the variance of $Y(\theta_\star^{\mathrm{KS}})$ is larger than that of X such that its corresponding variance-covariance c_1 dominates c_2, it becomes advantageous to use n much larger than m to reduce the variance of the estimator. Such a phenomenon usually happens at the beginning of training.

On the other hand, two main streams of approaches to kernel selection exist in the literature. The first stream focuses on investigating the asymptotic distribution of the estimator or the test statistics, such as in the kernel two-sample test (Gretton et al., 2012). The objective in this stream is to choose the kernel to maximize the power of the test (Sutherland et al., 2017; Liu et al., 2020) or minimize the asymptotic variance-covariance (Briol et al., 2019). In the latter case, while both C_λ and C are typically computationally intractable due to the unknown parameter $\theta_\star^{\mathrm{KS}}$, consistent estimations can be obtained (for example, $\widehat{C}_{m,n}$). The selection of the kernel for stochastic simulation models, guided by empirical estimations, is a subject for future research.

The second stream focuses on investigating the radial basis function (RBF) kernel, represented as $k(x, y) = r\left(\frac{\|x-y\|}{l}\right)$, where $r : \mathbb{R} \to \mathbb{R}^+$ is a positive-valued function, and l is the bandwidth of the kernel. The Gaussian

kernel is a notable instance of the RBF kernel. Several heuristics for RBF kernel selection, such as the median heuristic (Gretton et al., 2012) and the kernel mixture heuristic (Li et al., 2015), have demonstrated high power in kernel two-sample testing. However, the effectiveness of these heuristics in the context of stochastic simulation models requires further exploration in future research. Additionally, the choice of the parameter β for Riesz kernels presents another open problem to be investigated. The Riesz kernel with $1 \leq \beta \leq 2$ complies with Assumption A.2. However, there are currently no established guidelines for selecting an optimal value of β within the range of 1 to 2.

A.3.2 Differentiability and Unbiased Gradient for G/G/1 Queueing Models

We validate the differentiability of θ and gradient unbiasedness for each experiment in this section. We only discuss Experiment 4, as the case for all the other experiments follows an entirely analogous way. Recall in Experiment 4, the inter-arrival time of the G/G/1 model follows the Gamma distribution with rate λ, denoted Gamma$(0.5, \lambda)$, and the service time follows an exponential distribution with rate μ, denoted Exp(μ). The simulation parameter is denoted $\theta = (\mu, \lambda)$, and the input model is denoted G_θ.

For each θ, the procedure for generating a single sample $Y(\theta)$ is as follows. Sample $Y(\theta)$ is the average waiting times of a set of customers. To simulate $Y(\theta)$, we need to simulate a set of customer waiting times. We denote the waiting time of the j-th customer by $W_j(\theta)$. The Lindley equation provides a representation of the waiting time in a G/G/1 queueing system:

$$W_{j+1}(\theta) = \text{ReLU}\left(W_j(\theta) + S_j(\mu) - T_j(\lambda)\right),$$

where the service time of the j-th customer is denoted by $S_j(\mu)$, and the interarrival time between the j-th and $(j+1)$-th customer is denoted by $T_j(\lambda)$. We just need to simulate a set of customer service times $\{S_j(\mu)\}$ and waiting times $\{T_j(\lambda)\}$ and apply the Lindley equation to get $\{W_j(\theta)\}$ (Nelson and Pei, 2013).

The Lindley equation is akin to a ReLU activation function, hence simplifying validation of differentiability and unbiased gradient to Assumptions 3 and 4 in Assumption 2.1 on $S_j(\mu)$ and $T_j(\lambda)$. We focus solely on $T_j(\lambda)$, as the analysis for $S_j(\mu)$ follows an entirely analogous way.

Consider the 'pushforward' representation $T_j(\lambda) = T_\lambda(Z_j)$, $Z_j \overset{\text{i.i.d.}}{\sim} P$, where P is the user-specified reference measure. If we use inverse-transform sampling, then $P = U(0,1)$, $T_j(\lambda) = \frac{\log(1-Z_j)}{\lambda} = \psi_{\lambda,1}(\phi_1(Z_j))$, where $\psi_{\lambda,1}(z) = \frac{z}{\lambda}$ and $\phi_1(z) = \log(1-z)$. Such usual pseudo-random number generation with $P = U(0,1)$ violates Assumption 3 due to the non-M-Lipschitz inverse transform $\phi_1(z) = \log(1-z)$. Both assumptions hold if we select $P = \text{Exp}(1)$ and apply the linear transform $S_\mu = \frac{Q_n}{\mu}$. A similar decomposition holds for the rate parameter of a Gamma distribution: $P = \text{Gamma}(0.5, 0.5)$, $Z_j \sim P$, $S_j(\mu) = \frac{Z_j}{\mu}$. Hence, by picking a good reference measure P, G_θ is (almost everywhere) differentiable with respect to $\theta = (\mu, \lambda)$, and the gradient is unbiased.

Such decomposition of G_θ may not be universally applicable, as in the case of the shape parameter of a Gamma distribution. There exists no reference measure P such that the input model G_θ is differentiable when θ is the shape parameter of a Gamma distribution. Hence, the shape parameter is not learnable with our approach.

Furthermore, G_θ can also be a meta-model that models the service and arrival time distribution with learnable parameter θ. Such meta-models are usually deep neural networks (Chambers and Mount-Campbell, 2002; Ojeda et al., 2021) that satisfy Assumption 3 and 4 in Assumption 2.1.

A.3.3 Pathwise Gradients and Alternative Gradient Estimation Approaches

Our approach relies on pathwise gradients through backward propagation, which can be viewed as a form of infinitesimal perturbation analysis (IPA) in stochastic gradient estimation (Asmussen and Glynn, 2007; Fu, 2015; Mohamed et al., 2020). This connection to IPA highlights an important methodological consideration that our technique fundamentally depends on the existence of pathwise gradients for parameter estimation.

The pathwise gradient approach offers significant advantages when applicable, particularly in terms of computational efficiency and variance reduction. However, the existence of IPA estimators depends on how the stochastic process is represented (Fu, 2015). In other words, we need to find a good pair of reference measure P and input model G_θ as suggested in Section A.3.2. Hence, it faces limitations when simulation parameters do not admit pathwise differentiation, such as the shape parameter of the Gamma distribution discussed in Section A.3.2. For

such cases, score function/likelihood ratio estimators present a viable alternative. These methods can generate unbiased gradient estimates even when pathwise derivatives do not exist. Chapter VII of Asmussen and Glynn (2007), Section 5 of Fu (2015), and Section 4 of Mohamed et al. (2020) provide detailed discussions on using IPA and score function/likelihood ratio estimators for stochastic gradient estimation.

However, applying score estimators to likelihood-free stochastic simulation models presents a fundamental challenge: the inability to evaluate the log-likelihood of the model output distribution. One potential solution involves developing meta-models (e.g., neural network approximations) to estimate these likelihoods as suggested in Section A.3.2. However, this approach introduces additional complexity and computational considerations. The meta-model must approximate the likelihood function of the simulation parameter with sufficient accuracy while remaining computationally tractable. This limitation highlights an important direction for future research: developing efficient gradient estimation techniques that maintain the computational advantages of pathwise approaches while accommodating parameters that lack pathwise differentiability in likelihood-free settings.

A.3.4 Theoretical Applicability to Other Stochastic Simulation Models

Our approach can be applied to other stochastic simulation models, such as (s, S) stochastic inventory models, financial market simulators (Bai et al., 2021), and time-series models with latent variables, such as the stochastic volatility model (Kim et al., 1998) described in Section 5.2 of Briol et al. (2019). However, as in the case of G/G/1 queueing models, learning the simulation parameter θ indeed requires careful choice of reference measure P to ensure differentiability of θ and unbiasedness of the gradient estimate.

Even when the simulation parameter is differentiable, there is no guarantee that the U-statistic gradient estimate is unbiased if Assumption 2.1 is violated. Consider the stochastic volatility model described in Section 5.2 of (Briol et al., 2019). While the simulation parameters in this model are differentiable, the transformation from unobserved latent variables h_t to observed data y_t involves a non-M-Lipschitz exponential function, violating Assumption 3 in Assumption 2.1. Nevertheless, exploring the applicability of our approach to such models is of great interest to us and we plan to address this in our future work.

A.3.5 Validation of Asymptotic Normality Conditions

Besides the differentiability and unbiased gradient issue, the applicability of our confidence set procedure relies on verifying Assumption 2.3. Despite the fact that the assumptions in Assumption 2.3 are standard assumptions for asymptotic normality (Briol et al., 2019; Oates et al., 2022), they are technical and unintuitive in nature and hard to verify. Note that both Briol et al. (2019); Oates et al. (2022) do not verify these conditions in practice.

One case where Assumption 2.3 holds is when X and Z have bounded support (or are almost surely bounded), G_θ is an (almost everywhere, or P-almost surely) smooth and bounded function, and the kernel k is bounded (e.g. the Gaussian kernel). In queueing systems, this translates to bounded waiting and arrival times (e.g. from a $U(0, 1)$ distribution) and using the Gaussian kernel for KOSE. While we do not provide an explicit example where Assumption 2.3 does not hold, potential violations might occur with unbounded X and Z, when G_θ involves exponential transforms, or when using unbounded kernels like the Riesz or Polynomial kernels.

A.3.6 Scalability in Parameter Dimension

We acknowledge the limitation of our experiments to small parameter dimensions. This stems from implementation constraints and the need to ensure differentiability and gradient unbiasedness for each parameter.

Regarding scalability, while concentration and generalization bounds for MMD (Briol et al., 2019) suggest that convergence speed does not depend on the parameter dimension p, we expect the marginal asymptotic variance for each parameter to increase with p. However, the computationally intractable nature of the asymptotic variance-covariance C limits further theoretical insights. Larger p (and data dimension d) will also introduce challenges in Hessian and Jacobian estimation. Our plug-in estimators might produce confidence sets of incorrect coverages and the memory needed to compute the Hessian significantly increases.

Besides, with increasing p, the (terminal) SGD estimate $\widehat{\theta}_{m,n}^{(t)}$ we use to construct confidence set may deviate from the kernel optimum score estimator $\widehat{\theta}_m^{KS}$, similar to the challenges in deep learning optimization where local minima and saddle points are prevalent. Even at convergence, we may only reach a local optimum, and uniqueness of the kernel optimum score estimator $\widehat{\theta}_m^{KS}$ is not guaranteed.

To address uncertainties from SGD estimation, we can use the plug-in estimate for C_λ (variance-covariance matrix for kernel simulated score estimator defined in Theorem 2.4) rather than the plug-in estimate for C. By fixing the sample path w when simulating from P_θ, we can define $\widehat{\theta}_m^{KS}(w)$ and employ plug-in estimation for C_λ to construct confidence sets.

A.4 Experimental Details

We provide details on the experimental setup required to produce Tables 2 - 5.

In Table 1, we carefully design the experiment settings to study both stationary and non-stationary queueing systems. For Experiments 1 and 2, we set the target service rate to 1.2 to ensure system stationarity. Experiment 3 replicates the non-stationary queueing system setup from Bai and Lam (2020), allowing us to compare our results with established benchmarks. For Experiment 4, we selected a target service rate of 2.5 to maintain system stationarity while testing in a higher parameter dimension. Please refer to Section A.1.3 for more details regarding our experiment design.

Unless otherwise specified, all the experiments are run individually on an NVIDIA GeForce RTX 2080 Ti GPU. For Experiments 1, 2, the waiting time from the target system X and $Y(\theta)$ are the averages of 50 Monte Carlo draws following a burn-in period of 10. For Experiment 4, the waiting time from the target system X and $Y(\theta)$ are the averages of 20 Monte Carlo draws following a burn-in period of 10. The target system in Experiment 3 is a non-stationary queueing system, and the waiting times are thus the averages of 10 Monte Carlo draws with no burn-in period.

Kernel Optimum Score Estimation (KOSE) and Confidence Set Estimation For all experiments with our method, we use PyTorch (Paszke et al., 2019) for kernel optimum score estimation and confidence set estimation, and the Adam optimizer (Kingma and Ba, 2015) for minimizing the kernel optimum score. The learning rate at the k-th iteration is $\frac{\eta}{\sqrt{1+k}}$, where η is the initial learning rate. We set η to be 1 throughout our experiments. We set the stopping criterion at 200 iterations for Experiments 1 and 2, and 800 iterations for Experiments 3 and 4. We use $n_c = 5,000$ for confidence set estimation. The Hessian estimator $\widehat{H}_{m,n_c}$ is obtained with finite difference with interval 0.1 in Experiments 1 and 2, where the parameter space is one-dimensional, and $\widehat{H}_{m,n_c}$ is obtained with the built-in torch.autograd.functional.hessian within Pytorch in Experiments 3 and 4. For Experiments 2 and 4, where the model is inexact, the optimal simulation parameter $\theta_\star^{\mathrm{KS}}$ used to compute MSE and coverage for all methods except KOSE-Riesz is estimated with a sample of size $5,000$ from the target system using KOSE-Gaussian with learning rate $\eta = 1$ and stopping at iteration 800. The obtained optimal simulation parameters are displayed in the last column of Tables 2 - 5. For KOSE-Riesz, the optimal simulation parameter $\theta_\star^{\mathrm{KS}}$ used to compute MSE and coverage is estimated with a sample of size $5,000$ from the target system using KOSE-Riesz with learning rate $\eta = 1$ and stopping at iteration $1,000$. We reinforce numerical stability of the estimated optimal simulation parameters with Polyak-Ruppert averaging (Polyak and Juditsky, 1992) for the last 100 iteration. However, in Experiment 4, for $a = 0.2$, numerical instability causes $\theta_\star^{\mathrm{KS}}$ for KOSE-Gaussian to be $[0.00772042, 1.02951885]$, which is on the border of the parameter space. To address this issue, we report $\theta_\star^{\mathrm{KS}}$ for KOSE-Riesz instead.

MMD Posterior Bootstrap (NPL-MMD) For the MMD posterior bootstrap method (Dellaporta et al., 2022), we use the public GitHub repository by Dellaporta et al. (2024) that contains the experiment implementations. For all experiments with the MMD posterior bootstrap method, the number of points sampled from the model at each MMD optimization iteration is fixed at 30 (except it is set at 20 for $a = 0.4, 0.2$ in Experiment 4 due to memory issues), and the number of bootstrap iterations is fixed at $1,000$. The bandwidth of the Gaussian kernel used in MMD is picked with the median heuristic (Gretton et al., 2012), and the prior is the non-informative Dirichlet Process (DP) prior. For Experiments 1 and 2, where the parameter space is one-dimensional, the 95% credible set is the equal-tailed interval. The 95% joint credible set is constructed as the union of two 97.5% equal-tailed intervals on each dimension of the parameter space, for Experiments 3 and 4, from the joint posterior draw.

Accept-Reject ABC with MMD (ABC-AR) For the accept-reject ABC with MMD, we set the tolerance to be the 1% quantile of the simulated distances, as used in Experiment 1 of Frazier et al. (2020b). The bandwidth of the Gaussian kernel used in MMD is picked with the median heuristic (Gretton et al., 2012). The prior is the

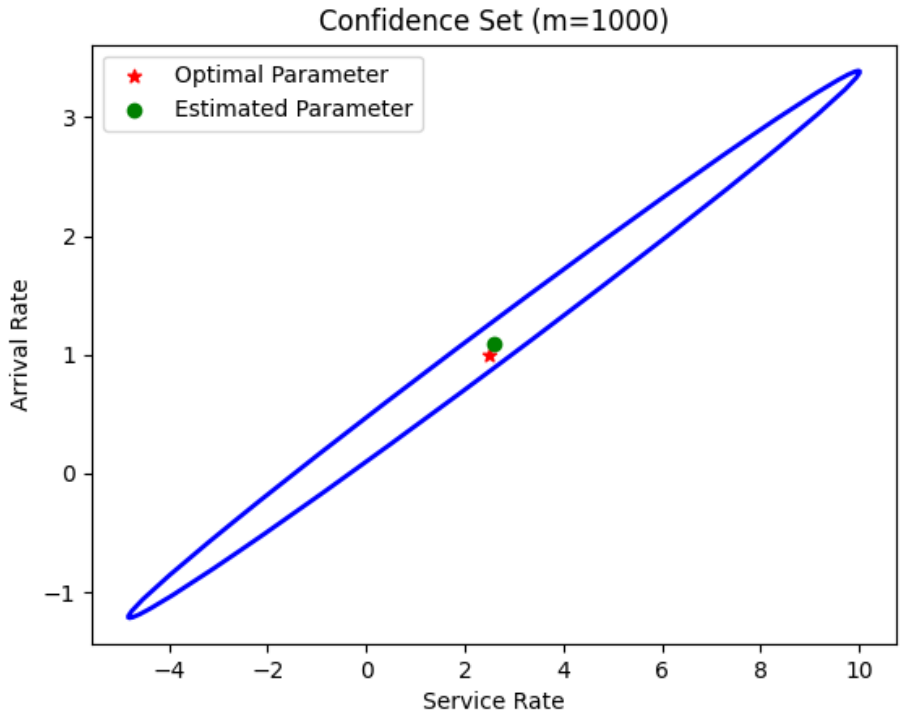

Figure 1: Confidence set for KOSE-Riesz, Experiment 4, $a = 1$.

uniform distribution on the set of parameters that are all within 1 in each dimension from the optimal simulation parameter, except in Experiment 1, where the optimal simulation parameter $\theta_\star^{\mathrm{KS}} = 1.2$ and the prior is $U(0.5, 2.5)$. We construct the 95% credible sets in the same way as the MMD posterior bootstrap method. The number of draws from the prior is $1,000$ for Experiments 1 and 2, and the number of draws from the prior is $9,000$ for Experiments 3 and 4 to account for the extra dimension of the parameter.

Eligibility Set (ES) For the eligibility set method (Bai and Lam, 2020), we sample $1,000$ candidate parameters $\theta \sim U(0.5, 2.5)$ in Experiment 1, and from the uniform distribution on the set of parameters that are all within 1 in each dimension from the optimal simulation parameter for Experiments 2 and 4.

A.4.1 Extra Experiments on the Choice of β in KOSE-Riesz

We conduct extra experiments on the choice of β in the Riesz kernel for $\beta = 1.25, 1.5, 1.75, 2$, where $\beta = 2$ corresponds to a non-strictly proper scoring rule. All experiments, unless otherwise specified, are run with an NVIDIA TITAN Xp GPU. The results under the setting of Table 2, Experiment 2: $R = 1,000$ are in Table 6. The results under the setting of Table 3, Experiment 4 are in Table 7.

Unlike the Gaussian kernel, there are no existing guidelines for selecting β in the Riesz kernel to the best of our knowledge. For non-integer values of β, the energy score is not differentiable, necessitating a smoothing of the energy score computation.

From Table 6, it appears that when β is closer to 1, the produced confidence set has better coverage. This trend holds in Table 7 except for $\alpha = 0.6$. Notably, KOSE-Riesz for all the values $\beta = 1.25, 1.5, 1.75, 2$ are capable of producing a valid confidence set for $\alpha = 0.4$, the most extreme model inexact scenario.

A.4.2 Extra Experiments on the Sensitivity of n

We experiment with the choice of n on the convergence and the coverage of the confidence sets for $n = 2, 10, 50, 100, 200$. The results under the setting of Table 3, Experiment 4 ($p = 2$, inexact model), run with an

Figure 2: Confidence set for KOSE-Riesz, Experiment 4, $a = 0.6$.

NVIDIA TITAN Xp GPU, are included in Table 8. The results for KOSE-Riesz under the setting of Experiment 1 ($p = 1$, exact model), run with an NVIDIA 1080Ti GPU, are included in Table 9.

From these tables, it is evident that smaller n leads to greater MSE in estimation and worse confidence set coverage if the number of iterations is unchanged. However, performance improves as the number of iterations increases.

A.4.3 Extra Experiments on the Bias in Uncertainty Evaluation

We conduct additional experiments to investigate the bias in uncertainty evaluation. This bias stems from two primary sources: estimation bias and variance-covariance approximation bias. The estimation bias arises from the learning procedure, where the distribution of the kernel optimum score estimator (KOSE) deviates from the asymptotic normal distribution. The variance-covariance approximation bias is introduced during the estimation of the variance-covariance matrix.

Quantifying the bias in uncertainty evaluation presents challenges. For one-dimensional examples (i.e., $p = 1$), the bias is considered small if the mean squared error (MSE) closely approximates the estimated average asymptotic variance. However, for two-dimensional examples (i.e., $p = 2$), we lack a definitive measure for this bias. One approach to assess the magnitude of the bias is to observe whether the asymptotic variance estimate stabilizes. This stabilization can be inferred by examining the average width and height of the produced confidence sets (scaled up by a factor of $\sqrt{m}$).

Table 11 presents the results of Experiment 2, including the estimated average asymptotic variance. Table 10 shows the results of KOSE-Riesz under the conditions of Table 3, Experiment 4. All experiments, unless otherwise specified, are run with an NVIDIA TITAN Xp GPU.

From Table 11, it is evident that the MSE closely approximates the asymptotic variance, particularly for KOSE-Riesz. Table 10 demonstrates that the width and height (scaled by $\sqrt{m}$) of the confidence set remain stable across $m = 1000, 2000$, and 5000, suggesting a stable asymptotic variance estimate.

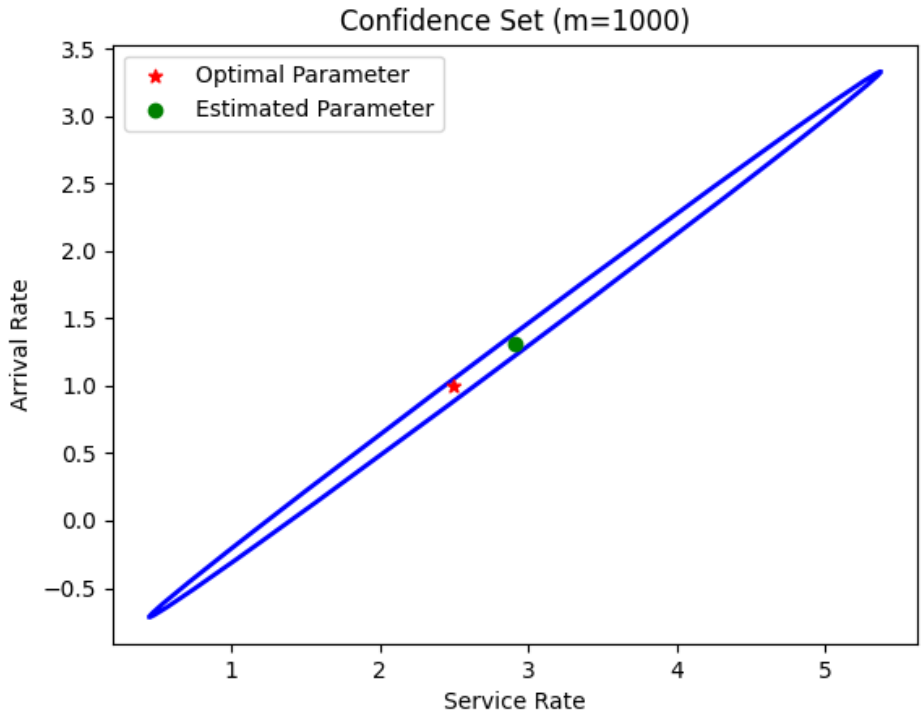

Figure 3: Confidence set for KOSE-Gaussian, Experiment 4, $a = 1$.

A.4.4 Extra Experiments on Contamination Models

We conduct an additional experiment on contamination models under the setting of Experiment 3. We add white noise drawn from a $N(0, 0.01)$ distribution to $(100\epsilon)\%$ of the output data from the target system, where $\epsilon = 0.01, 0.05, 0.1, 0.2, 0.5$. The results are reported in Table 12. From Table 12, we observe that our method is robust to noise, as expected from the robustness of the MMD and, consequently, the kernel score (Briol et al., 2019).

A.4.5 Extra Experiments on Stochastic Volatility Model

We conduct additional experiments on the stochastic volatility model (Kim et al., 1998) described in Section 5.2 of Briol et al. (2019). Following the notations of Briol et al. (2019), the model is defined as follows:

Given initial volatility $h_1 \sim N(0, \sigma^2/(1-\phi^2))$, the model follows the following set of equations:

$$h_t = \phi h_{t-1} + \eta_t, \eta_t \sim N(0, \sigma^2)$$
$$y_t = \epsilon_t \kappa \exp(0.5 h_t), \epsilon_t \sim N(0, 1).$$

where y_t is the observed mean corrected return on holding an asset at time t and h_t is the hidden log-volatility at time t. The observed data is $\{y_t\}_{t=1}^d$, and the simulation parameter here is (ϕ, κ, σ).

In our experiment, following Briol et al. (2019), we set the data dimension is $d = 30$, and we use the following reparameterization in our algorithm to avoid numerical issues: $\theta_1 = \log \frac{1+\phi}{1-\phi}, \theta_2 = \log \kappa, \theta_3 = \log(\sigma^2)$. We only consider the model exactness case i.e. M-closed case as in Briol et al. (2019).

Unless otherwise specified, all the experiments are run individually on an NVIDIA A100 Tensor Core GPU. For optimization, we use mini-batch stochastic gradient descent here with an initial learning rate of 10^{-3}, a batch size of 100, and train for 500 epochs. We consider two different settings of optimal simulation parameter: $(\phi_\star, \kappa_\star, \sigma_\star) = (0.98, 0.65, 0.15)$ and $(\phi_\star, \kappa_\star, \sigma_\star) \approx (0.45, 1.92, 1.08)$, corresponding to $\theta_\star \approx (2.00, -0.19, -1.65)$ and $\theta_\star = (0.98, 0.65, 0.15)$, respectively. The results for the former case are reported in Table 13, and the results for

Figure 4: Confidence set for KOSE-Gaussian, Experiment 4, $a = 0.6$.

the latter case are reported in Table 14.

From Tables 13 and 14, it is evident that KOSE-Riesz is able to produce valid confidence sets in both cases when $m \geq 500$, and KOSE-Gaussian is able to produce valid confidence sets for the second case when $m \geq 500$. The underperformance of KOSE-Gaussian in the first case likely stems from optimization challenges, as indicated by its substantially higher MSE values. These results demonstrate the capability of our method to construct valid confidence sets for simulation parameters in stochastic volatility models under model exactness.

Table 4: Experiment 1 results; R is the number of independent runs; $m = n$.

	KOSE-Riesz, $R = 1,000$				KOSE-Gaussian, $R = 1,000$			
m, n	MSE	Cov.	Width	T	MSE	Cov.	Width	T
10	$6.0 \cdot 10^{-3}$	**0.973**	0.45	6	$2.0 \cdot 10^{-1}$	**0.959**	0.51	2
50	$1.2 \cdot 10^{-3}$	**0.990**	0.20	7	$1.5 \cdot 10^{-3}$	**0.986**	0.21	2
100	$6.4 \cdot 10^{-4}$	**0.995**	0.14	6	$6.0 \cdot 10^{-4}$	**1.000**	0.15	3
500	$1.4 \cdot 10^{-4}$	**0.993**	0.06	8	$1.4 \cdot 10^{-4}$	**0.991**	0.07	8
$1,000$	$8.6 \cdot 10^{-5}$	**0.984**	0.05	10	$7.4 \cdot 10^{-5}$	**0.983**	0.04	15

	ABC-AR, $R = 1,000$			NPL-MMD, $R = 100$			ESet, $R = 1,000$		
m, n	Cov.	Width	T	Cov.	Width	T	Cov.	Width	T
10	0.885	0.29	3	0.79	0.77	30	**1.000**	0.42	3
50	0.854	0.12	3	0.80	0.38	30	**1.000**	0.19	3
100	**0.977**	0.08	3	0.67	0.27	31	**1.000**	0.13	3
500	**0.990**	0.05	3	0.44	0.14	34	**0.998**	0.05	3
$1,000$	**0.992**	0.03	3	0.43	0.09	67	**0.998**	0.04	3

Table 5: Experiment 3 results; R is the number of independent runs; $m = n$; MSE, μ and MSE, λ are the average mean-squared errors for the service and arrival rates. NR means that the experiment is not run due to excessive computation time.

	KOSE-Riesz, $R = 1,000$				KOSE-Gaussian, $R = 1,000$			
m, n	MSE, μ	MSE, λ	Cov.	T	MSE, μ	MSE, λ	Cov.	T
10	$9.0 \cdot 10^{-3}$	1.09	**1**	3	$1.1 \cdot 10^{-2}$	$1.6 \cdot 10^{-1}$	**0.999**	2
50	$2.5 \cdot 10^{-3}$	$2.5 \cdot 10^{-3}$	**1**	3	$1.0 \cdot 10^{-3}$	$1.5 \cdot 10^{-2}$	**1.000**	3
100	$2.1 \cdot 10^{-4}$	$2.0 \cdot 10^{-2}$	**1**	3	$1.9 \cdot 10^{-3}$	$4.9 \cdot 10^{-3}$	**1.000**	4
500	$3.4 \cdot 10^{-3}$	$6.5 \cdot 10^{-3}$	**1**	2	$7.1 \cdot 10^{-4}$	$1.1 \cdot 10^{-2}$	**1.000**	15
$1,000$	$2.0 \cdot 10^{-4}$	$6.0 \cdot 10^{-4}$	**1**	3	$4.1 \cdot 10^{-4}$	$9.1 \cdot 10^{-5}$	**1.000**	36

	ABC-AR, $R = 100$			NPL-MMD, $R = 100$		
m, n	Cov.	Width, μ, λ	T	Cov.	Width, μ, λ	T
10	0	$0.62, 0.59$	66	0.88	$1.06, 12.8$	30
50	NR	NR	NR	0.86	$0.62, 7.07$	32
100	**1**	$0.19, 0.65$	66	0.81	$0.50, 5.86$	38
500	NR	NR	NR	0.73	$0.30, 3.29$	69
$1,000$	NR	NR	NR	0.57	$0.22, 2.41$	161

Table 6: Under the setting of Table 2, Experiment 2: $R = 1,000$. Top: $\beta = 1.25$ (left entry), $\beta = 1.5$ (right entry). Bottom: $\beta = 1.75$ (left entry), $\beta = 2$ (right entry). Asymp.Var. is the estimated average asymptotic variance. NR means that the experiment is not run due to excessive computation time.

m, n	a	MSE	Cov.	Width	Asymp.Var.	T
500, 500	1.0	3.0×10^{-4}, 1.2×10^{-4}	0.683, 0.729	0.04, 0.02	1.1×10^{-4}, 4.1×10^{-5}	8, 8
500, 500	0.8	1.5×10^{-4}, 4.2×10^{-4}	**0.979**, 0.638	0.06, 0.04	2.2×10^{-4}, 1.3×10^{-4}	8, 8
500, 500	0.6	2.3×10^{-4}, 3.4×10^{-4}	**0.990**, **0.961**	0.09, 0.08	4.9×10^{-4}, 4.4×10^{-4}	8, 8
500, 500	0.4	1.4×10^{-3}, 9.9×10^{-4}	**0.999**, **1.000**	0.18, 0.19	2.1×10^{-3}, 2.3×10^{-3}	8, 8
500, 500	0.2	NR, NR	NR, NR	NR, NR	NR, NR	NR, NR
500, 500	1.0	1.3×10^{-4}, 2.6×10^{-4}	0.533, 0.154	0.02, 0.01	1.7×10^{-5}, 6.6×10^{-6}	8, 8
500, 500	0.8	3.3×10^{-4}, 2.3×10^{-4}	0.564, 0.527	0.03, 0.02	6.3×10^{-5}, 3.5×10^{-5}	8, 8
500, 500	0.6	2.5×10^{-4}, 2.0×10^{-4}	**0.966**, **0.955**	0.07, 0.06	3.3×10^{-4}, 2.6×10^{-4}	8, 8
500, 500	0.4	1.7×10^{-3}, 3.3×10^{-4}	**1.000**, **1.000**	0.21, 0.21	2.8×10^{-3}, 2.9×10^{-3}	9, 9
500, 500	0.2	1.7×10^{-2}, 4.2×10^{-3}	**1.000**, **1.000**	0.71, 1.12	3.4×10^{-2}, 8.8×10^{-2}	8, 9

Table 7: Under the setting of Table 3, Experiment 4. $R = 1,000$. Top: $\beta = 1.25$ (left entry), $\beta = 1.5$ (right entry). Bottom: $\beta = 1.75$ (left entry), $\beta = 2$ (right entry). NR means that the experiment is not run due to excessive computation time.

m, n	a	MSE, μ	MSE, λ	Cov.	T
1000, 1000	1.0	1.8×10^{-2}, 1.1×10^{-2}	2.4×10^{-2}, 7.6×10^{-3}	**1**, **1**	8.2, 8.3
1000, 1000	0.8	6.3×10^{-3}, 6.1×10^{-3}	1.4×10^{-3}, 1.7×10^{-3}	**1**, **1**	8.1, 8.5
1000, 1000	0.6	3.98×10^{-1}, 4.85×10^{-1}	1.55×10^{-1}, 2.12×10^{-1}	0, 0	7.5, 8.4
1000, 1000	0.4	8.7×10^{-4}, 7.5×10^{-4}	9.6×10^{-3}, 2.1×10^{-3}	**1**, **1**	7.5, 8.5
1000, 1000	0.2	3.0×10^{-3}, NR	2.3×10^{-3}, NR	**1**, NR	7.7, NR
1000, 1000	1.0	4.87×10^{-1}, 6.91	5.00×10^{-1}, 28.88	0, 0	8.4, 8.3
1000, 1000	0.8	1.7×10^{-2}, 2.8×10^{-3}	1.0×10^{-2}, 2.1×10^{-4}	**1**, **1**	8.5, 8.1
1000, 1000	0.6	3.28×10^{-1}, 4.9×10^{-4}	1.36×10^{-1}, 1.2×10^{-2}	0, **1**	8.4, 8.1
1000, 1000	0.4	3.7×10^{-4}, 9.0×10^{-3}	1.4×10^{-3}, 1.0×10^{-2}	**1**, **1**	8.4, 8.1
1000, 1000	0.2	NR, 6.9×10^{-4}	NR, 3.7×10^{-3}	NR, 0	NR, 8.3

Table 8: Under the setting of Table 3, Experiment 4. $R = 1,000$. Sensitivity to n for KOSE-Riesz (left entry) and KOSE-Gaussian (right entry). Iterations is the number of SGD iterations.

n, m	a	MSE, μ	MSE, λ	Cov.	T	Iterations
2, 1000	0.8	8.91, 10.29	9.92, 10.95	0, 0	7.7, 25.5	800
10, 1000	0.8	4.09, 3.63	3.22, 2.22	0, 0	7.7, 25.3	800
50, 1000	0.8	1.21, 0.57	0.82, 0.34	0, 0	7.6, 25.2	800
100, 1000	0.8	0.48, 0.16	0.37, 0.13	0, 0	7.4, 25.1	800
500, 1000	0.8	0.07, 0.04	0.03, 0.01	0, **1**	7.9, 25.3	800
2, 1000	0.8	9.71, 6.33	8.40, 4.89	0, 0	18.0, 36.5	2500
10, 1000	0.8	1.82, 1.63	1.25, 0.93	0, 0	14.9, 33.3	2000
50, 1,000	0.8	0.44, 0.12	0.29, 0.07	0, 0	12.4, 30.2	1500
100, 1000	0.8	0.35, 0.12	0.18, 0.06	0, **1**	9.2, 26.9	1000
500, 1000	0.8	0.02, 0.02	0.01, 0.01	0, **1**	9.3, 26.8	1000

Table 9: KOSE-Riesz under the setting of Experiment 1. $R = 1,000$. Asymp.Var. is the estimated average asymptotic variance.

	KOSE-Riesz, $R = 1,000$					
n, m	a	MSE	Cov.	Width	Asymp.Var.	T
2,500	1	0.015	0.289	0.26	0.185	5
10,500	1	0.003	0.597	0.34	0.035	5
50,500	1	0.001	0.688	0.07	0.000	5
100,500	1	0.001	0.896	0.07	0.000	5
200,500	1	0.000	**0.976**	0.07	0.000	5

Table 10: KOSE-Riesz under the setting of Table 3, Experiment 4; $R = 100$. Width and Height are the average width and height of the estimated confidence sets, respectively.

	KOSE-Riesz, $R = 100$				
n, m	a	Cov.	Width	Height	T
$1000, 1000$	1.0	**1.00**	0.05	2.61	7.9
$1000, 1000$	0.8	0.00	0.05	2.71	7.9
$1000, 1000$	0.6	0.00	0.03	1.28	8.0
$1000, 1000$	0.4	0.00	0.01	1.25	8.1
$1000, 1000$	0.2	0.00	0.00	1.04	8.2
$1000, 2000$	1.0	**1.00**	0.03	1.56	11.4
$1000, 2000$	0.8	0.00	0.04	2.39	11.3
$1000, 2000$	0.6	0.00	0.03	1.37	10.9
$1000, 2000$	0.4	0.00	0.01	1.29	10.7
$1000, 2000$	0.2	0.00	0.00	1.13	10.7
$1000, 5000$	1.0	**0.99**	0.04	1.83	20.4
$1000, 5000$	0.8	**1.00**	0.05	2.65	20.2
$1000, 5000$	0.6	0.00	0.02	1.27	19.7
$1000, 5000$	0.4	0.00	0.01	1.27	19.6
$1000, 5000$	0.2	0.00	0.00	1.08	19.5

Table 11: Experiment 2 results; $R = 1,000$. Asymp.Var. is the estimated average asymptotic variance.

	KOSE-Riesz, $R = 1,000$					
a	MSE	Cov.	Width	Asymp.Var.	T	$\theta_\star^{\mathrm{KS}}$
1	$4.0 \cdot 10^{-4}$	0.920	0.07	$3.1 \cdot 10^{-4}$	3	1.2
0.8	$1.6 \cdot 10^{-4}$	**0.996**	0.07	$3.6 \cdot 10^{-4}$	3	1.5
0.6	$2.6 \cdot 10^{-4}$	**0.999**	0.10	$6.7 \cdot 10^{-4}$	3	1.9
0.4	$6.7 \cdot 10^{-4}$	**0.999**	0.15	$1.5 \cdot 10^{-3}$	3	2.5
0.2	$3.0 \cdot 10^{-3}$	**1.000**	0.34	$8.0 \cdot 10^{-3}$	3	4.1
	KOSE-Gaussian, $R = 1,000$					
a	MSE	Cov.	Width	Asymp.Var.	T	$\theta_\star^{\mathrm{KS}}$
1	$1.8 \cdot 10^{-4}$	**0.990**	0.06	$2.8 \cdot 10^{-4}$	4	1.2
0.8	$7.4 \cdot 10^{-4}$	0.842	0.08	$4.0 \cdot 10^{-4}$	4	1.5
0.6	$8.0 \cdot 10^{-7}$	**0.942**	0.10	$6.6 \cdot 10^{-4}$	3	1.9
0.4	$1.0 \cdot 10^{-3}$	**0.997**	0.17	$1.8 \cdot 10^{-3}$	4	2.5
0.2	$1.9 \cdot 10^{-2}$	0.833	0.37	$9.8 \cdot 10^{-3}$	4	4.1

Table 12: Contamination Models under the setting of Experiment 3; $R = 100$; ϵ is the noise level; $m = n = 1,000$. MSE, μ and MSE, λ are the average mean-squared errors for the service and arrival rates.

	KOSE-Riesz, $R = 100$		KOSE-Gaussian, $R = 100$	
ϵ	MSE, μ	MSE, λ	MSE, μ	MSE, λ
0.01	0.190	0.093	0.071	0.024
0.05	0.222	0.100	0.064	0.023
0.1	0.221	0.099	0.332	0.118
0.2	0.227	0.102	0.270	0.135
0.5	0.244	0.111	0.058	0.029

Table 13: Stochastic volatility model with $(\phi_\star, \kappa_\star, \sigma_\star) = (0.98, 0.65, 0.15)$, $\theta_\star \approx (2.00, -0.19, -1.65)$; R is the number of independent runs; $n = 100$; MSE, ϕ, MSE, κ, and MSE, σ are the average mean-squared errors for ϕ, κ and σ, respectively.

	KOSE-Riesz, $R = 100$					KOSE-Gaussian, $R = 100$				
m	MSE, ϕ	MSE, κ	MSE, σ	Cov.	T	MSE, ϕ	MSE, κ	MSE, σ	Cov.	T
10	$4.5 \cdot 10^{-3}$	$7.1 \cdot 10^{-3}$	$2.0 \cdot 10^{-3}$	0.43	4	$4.8 \cdot 10^{-3}$	$5.6 \cdot 10^{-3}$	$2.4 \cdot 10^{-3}$	0.30	4
50	$2.6 \cdot 10^{-3}$	$1.6 \cdot 10^{-3}$	$1.2 \cdot 10^{-3}$	0.17	4	$4.3 \cdot 10^{-3}$	$1.6 \cdot 10^{-3}$	$2.2 \cdot 10^{-3}$	0.06	4
100	$2.3 \cdot 10^{-3}$	$1.2 \cdot 10^{-3}$	$1.0 \cdot 10^{-3}$	0.03	5	$4.2 \cdot 10^{-3}$	$1.2 \cdot 10^{-3}$	$2.1 \cdot 10^{-3}$	0.00	5
500	$3.1 \cdot 10^{-5}$	$1.1 \cdot 10^{-4}$	$4.6 \cdot 10^{-4}$	**0.99**	23	$4.4 \cdot 10^{-4}$	$4.3 \cdot 10^{-4}$	$3.0 \cdot 10^{-4}$	0.00	28
1,000	$2.7 \cdot 10^{-5}$	$7.4 \cdot 10^{-5}$	$4.7 \cdot 10^{-4}$	**0.96**	45	$2.0 \cdot 10^{-4}$	$2.0 \cdot 10^{-4}$	$2.4 \cdot 10^{-3}$	0.00	51

Table 14: Stochastic volatility model with $(\phi_\star, \kappa_\star, \sigma_\star) \approx (0.45, 1.92, 1.08)$, $\theta_\star = (0.98, 0.65, 0.15)$; R is the number of independent runs; $n = 100$; MSE, ϕ, MSE, κ, and MSE, σ are the average mean-squared errors for ϕ, κ and σ, respectively.

	KOSE-Riesz, $R = 100$					KOSE-Gaussian, $R = 100$				
m	MSE, ϕ	MSE, κ	MSE, σ	Cov.	T	MSE, ϕ	MSE, κ	MSE, σ	Cov.	T
10	$2.1 \cdot 10^{-3}$	$5.8 \cdot 10^{-2}$	$8.0 \cdot 10^{-3}$	0.51	4	$1.6 \cdot 10^{-3}$	$4.7 \cdot 10^{-2}$	$5.0 \cdot 10^{-3}$	0.50	4
50	$2.7 \cdot 10^{-3}$	$1.2 \cdot 10^{-2}$	$7.3 \cdot 10^{-3}$	0.90	4	$1.9 \cdot 10^{-3}$	$9.8 \cdot 10^{-3}$	$4.1 \cdot 10^{-3}$	0.89	4
100	$3.0 \cdot 10^{-3}$	$6.3 \cdot 10^{-3}$	$7.4 \cdot 10^{-3}$	**0.96**	5	$2.0 \cdot 10^{-3}$	$5.1 \cdot 10^{-3}$	$4.1 \cdot 10^{-3}$	**0.96**	5
500	$3.0 \cdot 10^{-3}$	$3.5 \cdot 10^{-3}$	$6.7 \cdot 10^{-3}$	**0.99**	23	$1.9 \cdot 10^{-3}$	$2.4 \cdot 10^{-3}$	$3.8 \cdot 10^{-3}$	**0.99**	23
1,000	$3.1 \cdot 10^{-3}$	$2.5 \cdot 10^{-3}$	$6.0 \cdot 10^{-3}$	**0.96**	45	$1.6 \cdot 10^{-3}$	$2.0 \cdot 10^{-4}$	$3.5 \cdot 10^{-3}$	**0.97**	46