ETH zürich

EMULATION OF STOCHASTIC SIMULATORS USING GENERALIZED LAMBDA MODELS

X. Zhu and B. Sudret

CHAIR OF RISK, SAFETY AND UNCERTAINTY QUANTIFICATION
STEFANO-FRANSCINI-PLATZ 5
CH-8093 ZÜRICH

Risk, Safety & Uncertainty Quantification

Data Sheet

Journal:	-
Report Ref.:	RSUQ-2020-006
Arxiv Ref.:	https://arxiv.org/abs/2007.00996 [stat.CO, stat.ME]
DOI:	-
Date submitted:	July 2nd, 2020
Date accepted:	-

Emulation of stochastic simulators using generalized lambda models

Xujia Zhu [*1] and Bruno Sudret[†1]

[1] *Chair of Risk, Safety and Uncertainty Quantification, ETH Zürich, Stefano-Franscini-Platz 5, 8093 Zürich, Switzerland*

July 2, 2020

Abstract

Computer simulations are used in virtually all fields of applied science and engineering to predict the behavior of complex systems. In the context of uncertainty quantification and optimization, a large number of simulations are usually necessary, which may become intractable for high-fidelity numerical models. This problem is even more severe when it comes to stochastic simulators, which do not provide a deterministic outcome for a given set of input parameters, but rather a realization of a random variable. In a recent paper, we developed a novel surrogate model for stochastic simulators. In that framework, the response distribution is assumed to be a generalized lambda distribution. The associated distribution parameters are cast as functions of input variables and represented by polynomial chaos expansions. In this paper, we propose a new fitting procedure to construct such a surrogate model, which combines the maximum conditional likelihood estimation with (modified) feasible generalized least-squares. This method does not require repeated model evaluations for the same input parameters, so it is more versatile than the existing replication-based approaches. We compare the new method with the state-of-the-art nonparametric kernel estimator on two analytical examples and case studies. The performance of the proposed method is illustrated in terms of the accuracy of both the response distribution approximation and statistical quantities depending on the specific focus of the application.

1 Introduction

With increasing demands on the functionality and performance of modern engineering systems, design and maintenance of complex products and structures require advanced computational models, a.k.a. simulators. They help assess the reliability and optimize the behavior of the

[*]zhu@ibk.baug.ethz.ch
[†]sudret@ethz.ch

system already at the design phase. Classical simulators are usually deterministic because they implement solvers for the governing equation of the system. Thus, repeated model evaluations with the same input parameters consistently result in the same value of the output quantities of interest (QoI). In contrast, *stochastic simulators* contain intrinsic randomness, which leads to the QoI being a random variable conditioned on the given set of input parameters. In other words, each model evaluation with the same input values generates a realization of the response random variable that follows an unknown distribution. Formally, a stochastic simulator $\mathcal{M}_s$ can be expressed as

$$\begin{aligned} \mathcal{M}_s : \mathcal{D}_{\boldsymbol{X}} \times \Omega &\to \mathbb{R} \\ (\boldsymbol{x}, \omega) &\mapsto \mathcal{M}_s(\boldsymbol{x}, \omega), \end{aligned} \tag{1}$$

where $\boldsymbol{x}$ is the input vector that belongs to the input space $\mathcal{D}_{\boldsymbol{X}}$, and Ω denotes the sample space of the probability space $\{\Omega, \mathcal{F}, \mathbb{P}\}$ that represents the internal source of randomness.

Stochastic simulators are widely used in modern engineering, finance and medical sciences. Typical examples include evaluating the performance of a wind turbine under stochastic loads (Abdallah et al., 2019), predicting the price of an option in financial markets (Shreve, 2004), and the spread of a disease in epidemiology (Britton, 2010).

Due to the random nature of stochastic simulators, repeated model evaluations with the same input parameters, called hereinafter *replications*, are necessary to fully characterize the probability distribution of the corresponding QoI. In addition, uncertainty quantification and optimization problems typically require model evaluations for various sets of input parameters. Altogether, it is necessary to have a large number of model runs, which becomes intractable for costly models. To alleviate the computational burden, surrogate models, a.k.a. emulators, can be used to replace the original model. Such a model emulates the input-output relation of the simulator and is easy and cheap to evaluate.

Among several options for constructing surrogate models, this paper focuses on the so-called *non-intrusive* approaches. More precisely, the computational model is considered as a "black box" and only required to be evaluated on a limited number of input values (called the *experimental design*).

Three classes of methods can be found in the literature for emulating the entire response distribution of a stochastic code in a non-intrusive manner. The first one is the *random field approach*, which approximates the stochastic simulator by a random field. The definition in Eq. (1) implies that a stochastic simulator can be regarded as a random field indexed by its input variables. Controlling the intrinsic randomness allows one to get access to different trajectories of the simulator, which are deterministic functions of the input variables. In practice, this is achieved by fixing the *random seed* inside the simulator. Evaluations of the trajectories over the experimental design can then be extended to continuous trajectories, either by classical surrogate methods (Jimenez et al., 2017) or through Karhunen-Loève expansions (Azzi et al.,

2019). Since this approach requires the effective access to the random seed, it is only applicable to data generated in a specific way.

Another class of methods is the *replication-based approach*, which relies on using replications at all points of the experimental design to represent the response distribution through a suitable parametrization. The estimated distribution parameters are then treated as (noisy) outputs of a deterministic simulator. Then, conventional surrogate modeling methods, such as Gaussian processes (Rasmussen and Williams, 2006) and polynomial chaos expansions (PCEs) (Blatman and Sudret, 2011), can emulate these parameters as a function of the model input (Moutoussamy et al., 2015; Browne et al., 2016). Because this approach employs two separate steps, the surrogate quality depends on the accuracy of the distribution estimation from replicates in the first step (Zhu and Sudret, 2020b). Therefore, many replications are necessary, especially when non-parametric estimators are used for the local inference (Moutoussamy et al., 2015; Browne et al., 2016).

A third class of methods, known as the *statistical approach*, does not require replications or controlling the random seed. If the response distribution belongs to the exponential family, generalized linear models (McCullagh and Nelder, 1989) and generalized additive models (Hastie and Tibshirani, 1990) can be efficiently applied. When the QoI for a given set of input parameters follows an arbitrary distribution, nonparametric estimators can be considered, notably kernel density estimators (Fan and Gijbels, 1996; Hall et al., 2004) and projection estimators (Efromovich, 2010). However, it is well-known that nonparametric estimators suffer from the *curse of dimensionality* (Tsybakov, 2009), meaning that the necessary amount of data increases drastically with increasing input dimensionality.

In a recent paper (Zhu and Sudret, 2020b), we proposed a novel stochastic emulator called the *generalized lambda model* (GLaM). Such a surrogate model uses generalized lambda distributions (GLDs) to represent the response probability density function (PDF). The dependence of the distribution parameters on the input is modeled by polynomial chaos expansions. However, the methods developed in that paper (Zhu and Sudret, 2020b) rely on replications. In the present contribution, we propose a new statistical approach combining feasible generalized least-squares with maximum conditional likelihood estimations to get rid of the need for replications. Therefore, the proposed method is much more versatile, in the sense that replications and seed controls are not necessary anymore.

The paper is organized as follows. In Sections 2 and 3, we briefly review generalized lambda distributions and polynomial chaos expansions, which are the two main elements constituting the generalized lambda model. In Section 4, we recap the GLaM framework and introduce the maximum conditional likelihood estimator. Then, we present the algorithm developed to find an appropriate starting point to optimize the likelihood, and to design ad-hoc truncation schemes for the polynomial chaos expansions of distribution parameters. In Sections 5 and 6, we validate the proposed method on two analytical examples and two case studies in mathematical finance and epidemiology, respectively, to showcase its capability to tackle real problems. Finally, we

summarize the main findings of the paper and provide an outlook for future research in Section 7.

2 Generalized lambda distributions

2.1 Formulation

The generalized lambda distribution is a flexible probability distribution family. It is able to approximate most of the well-known parametric distributions (Freimer et al., 1988; Karian and Dudewicz, 2000), e.g., uniform, normal, Weibull and Student's distributions. The definition of a GLD relies on a parametrization of the *quantile function* $Q(u)$, which is a non-decreasing function defined on $[0, 1]$. In this paper, we consider the GLD of the Freimer-Kollia-Mudholkar-Lin family (Freimer et al., 1988), which is defined by

$$Q(u; \boldsymbol{\lambda}) = \lambda_1 + \frac{1}{\lambda_2}\left(\frac{u^{\lambda_3} - 1}{\lambda_3} - \frac{(1-u)^{\lambda_4} - 1}{\lambda_4}\right), \tag{2}$$

where $\boldsymbol{\lambda} = \{\lambda_l : l = 1, \ldots, 4\}$ are the four distribution parameters. More precisely, λ_1 is the location parameter, λ_2 is the scaling parameter, and λ_3 and λ_4 are the shape parameters. To ensure valid quantile functions (i.e., Q being non-decreasing on $[0, 1]$), it is required that λ_2 be positive. Based on the quantile function, the PDF $f_Y(y; \boldsymbol{\lambda})$ of a random variable Y following a GLD can be derived as

$$f_Y(y; \boldsymbol{\lambda}) = \frac{f_U(u)}{Q'(u; \boldsymbol{\lambda})} = \frac{\lambda_2}{u^{\lambda_3 - 1} + (1-u)^{\lambda_4 - 1}} \mathbb{1}_{[0,1]}(u), \text{ with } u = Q^{-1}(y; \boldsymbol{\lambda}), \tag{3}$$

where $\mathbb{1}_{[0,1]}$ is the indicator function. A closed-form expression of Q^{-1}, and therefore of f_Y, is in general not available, and thus the PDF is evaluated by solving the nonlinear equation Eq. (3) numerically.

2.2 Properties

GLDs cover a wide range of shapes determined by λ_3 and λ_4. For instance, $\lambda_3 = \lambda_4$ produces symmetric PDFs, and $\lambda_3, \lambda_4 < 1$ leads to bell-shaped distributions. Moreover, λ_3 and λ_4 are closely linked to the support and the tail properties of the corresponding PDF. $\lambda_3 > 0$ implies that the PDF support is left-bounded and $\lambda_4 > 0$ corresponds to right-bounded PDFs. Conversely, the distribution has lower infinite support for $\lambda_3 \leq 0$ and upper infinite support for $\lambda_4 \leq 0$. More precisely, the support of the PDF denoted by $\text{supp}\left(f_Y(y; \boldsymbol{\lambda})\right) = [B_l, B_u]$ is given by

$$B_l(\boldsymbol{\lambda}) = \begin{cases} -\infty, & \lambda_3 \leq 0 \\ \lambda_1 - \frac{1}{\lambda_2 \lambda_3}, & \lambda_3 > 0 \end{cases}, \quad B_u(\boldsymbol{\lambda}) = \begin{cases} +\infty, & \lambda_4 \leq 0 \\ \lambda_1 + \frac{1}{\lambda_2 \lambda_4}, & \lambda_4 > 0 \end{cases}. \tag{4}$$

Importantly, for $\lambda_3 < 0$ ($\lambda_4 < 0$), the left (resp. right) tail decays asymptotically as a power law, and thus the GLD family can also provide fat-tailed distributions. Due to this power law decay,

for $\lambda_3 \leq -\frac{1}{k}$ or $\lambda_4 \leq -\frac{1}{k}$, moments of order greater than k do not exist. For $\lambda_3, \lambda_4 > -0.5$, the mean and variance exist and are given by

$$\mu = \mathbb{E}[Y] = \lambda_1 - \frac{1}{\lambda_2}\left(\frac{1}{\lambda_3+1} - \frac{1}{\lambda_4+1}\right), \quad v = \mathrm{Var}[Y] = \frac{(d_2 - d_1^2)}{\lambda_2^2}, \tag{5}$$

where the two auxiliary variables d_1 and d_2 are defined by

$$\begin{aligned} d_1 &= \frac{1}{\lambda_3}\mathrm{B}(\lambda_3+1,1) - \frac{1}{\lambda_4}\mathrm{B}(1,\lambda_4+1), \\ d_2 &= \frac{1}{\lambda_3^2}\mathrm{B}(2\lambda_3+1,1) - \frac{2}{\lambda_3\lambda_4}\mathrm{B}(\lambda_3+1,\lambda_4+1) + \frac{1}{\lambda_4^2}\mathrm{B}(1,2\lambda_4+1), \end{aligned} \tag{6}$$

with B denoting the beta function.

3 Polynomial chaos expansions

Consider a deterministic computational model $\mathcal{M}_d(\boldsymbol{x})$ that maps a set of input parameters $\boldsymbol{x} = (x_1, x_2, \ldots, x_M)^T \in \mathcal{D}_{\boldsymbol{X}} \subset \mathbb{R}^M$ to the system response $y \in \mathbb{R}$. In the context of uncertainty quantification, the input variables are affected by uncertainty due to lack of knowledge or intrinsic variability (also called aleatory uncertainty). Therefore, they are modeled by random variables and grouped into a random vector $\boldsymbol{X}$ characterized by a joint PDF $f_{\boldsymbol{X}}$. The uncertainty in the input variables propagates through the the model $\mathcal{M}_d$ to the output, which becomes a random variable denoted by $Y = \mathcal{M}_d(\boldsymbol{X})$.

Provided that the output random variable Y has finite variance, $\mathcal{M}_d$ belongs to the Hilbert space $\mathcal{H}$ of square-integrable functions associated with the inner product

$$\langle u, v\rangle_{\mathcal{H}} \overset{\text{def}}{=} \mathbb{E}[u(\boldsymbol{X})v(\boldsymbol{X})] = \int_{\mathcal{D}_{\boldsymbol{X}}} u(\boldsymbol{x})v(\boldsymbol{x})f_{\boldsymbol{X}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{7}$$

If the joint PDF $f_{\boldsymbol{X}}$ fulfills certain conditions (Ernst et al., 2012), the space spanned by multivariate polynomials is dense in $\mathcal{H}$. In other words, $\mathcal{H}$ is a separable Hilbert space admitting a polynomial basis.

In this study, we assume that $\boldsymbol{X}$ has mutually independent components, and thus the joint distribution $f_{\boldsymbol{X}}$ is expressed as

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{j=1}^{M} f_{X_j}(x_j). \tag{8}$$

Let $\left\{\phi_k^{(j)} : k \in \mathbb{N}\right\}$ be the orthogonal polynomial basis with respect to the marginal distribution of f_{X_j}, i.e.,

$$\mathbb{E}\left[\phi_k^{(j)}(X_j)\,\phi_l^{(j)}(X_j)\right] = \delta_{kl}, \tag{9}$$

with δ being the Kronecker symbol defined by $\delta_{kl} = 1$ if $k = l$ and $\delta_{kl} = 0$ otherwise. Then, multivariate orthogonal polynomial basis can be obtained as the tensor product of univariate

polynomials (Soize and Ghanem, 2004):

$$\psi_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \prod_{j=1}^{M} \phi_{\alpha_j}^{(j)}(x_j), \tag{10}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M) \in \mathbb{R}^M$ denotes the multi-index of degrees. Each component α_j indicates the polynomial degree of ϕ_{α_j} and thus of $\psi_{\boldsymbol{\alpha}}$ in the j-th variable x_j. For some classical distributions, e.g., normal, uniform, exponential, the associated univariate orthogonal polynomials are well-known as Hermite, Legendre and Laguerre polynomials (Xiu and Karniadakis, 2002). For arbitrary marginal distributions, such a basis can be computed numerically through the *Stieltjes procedure* (Gautschi, 2004).

Following the construction defined in Eq. (10), $\left\{\psi_{\boldsymbol{\alpha}}(\cdot), \boldsymbol{\alpha} \in \mathbb{N}^M\right\}$ forms an orthogonal basis for $\mathcal{H}$. Thus, the random output Y can be represented by

$$Y = \mathcal{M}_d(\boldsymbol{X}) = \sum_{\boldsymbol{\alpha} \in \mathbb{N}^M} c_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\boldsymbol{X}), \tag{11}$$

where $c_{\boldsymbol{\alpha}}$ is the coefficient associated with the basis function $\psi_{\boldsymbol{\alpha}}$. The spectral representation in Eq. (11) is a series with infinitely many terms. In practice, it is necessary to adopt truncation schemes to *approximate* $\mathcal{M}_d(\boldsymbol{x})$ with a finite series defined by a finite subset $\mathcal{A} \subset \mathbb{N}^M$ of multi-indices. A typical scheme is the hyperbolic (q-norm) truncation scheme (Blatman and Sudret, 2010):

$$\mathcal{A}^{p,q,M} = \left\{ \boldsymbol{\alpha} \in \mathbb{N}^M, \|\boldsymbol{\alpha}\|_q = \left(\sum_{i=1}^{M} |\alpha_i|^q \right)^{\frac{1}{q}} \leq p \right\}, \tag{12}$$

where p is the maximum total degree of polynomials, and $q \leq 1$ defines the quasi-norm $\|\cdot\|_q$. Note that with $q = 1$, we obtain the so-called full basis of total degree less than p.

4 Generalized lambda models (GLaM)

4.1 Introduction

Because of their flexibility, we assume that the response random variable of a stochastic simulator for a given input vector $\boldsymbol{x}$ follows a generalized lambda distribution. Hence, the distribution parameters $\boldsymbol{\lambda}$ are functions of the input variables:

$$Y(\boldsymbol{x}) \sim \mathrm{GLD}\left(\lambda_1(\boldsymbol{x}), \lambda_2(\boldsymbol{x}), \lambda_3(\boldsymbol{x}), \lambda_4(\boldsymbol{x})\right). \tag{13}$$

Under appropriate conditions discussed in Section 3, each component of $\boldsymbol{\lambda}(\boldsymbol{x})$ admits a spectral representation in terms of orthogonal polynomials. Recall that $\lambda_2(\boldsymbol{x})$ is required to be positive (see Section 2). Thus, we choose to build the associated PCE on the natural logarithm transform

$\log(\lambda_2(\boldsymbol{x}))$. This results in the following approximations:

$$\lambda_l(\boldsymbol{x}) \approx \lambda_l^{\mathrm{PC}}(\boldsymbol{x};\boldsymbol{c}) = \sum_{\boldsymbol{\alpha}\in\mathcal{A}_l} c_{l,\boldsymbol{\alpha}}\psi_{\boldsymbol{\alpha}}(\boldsymbol{x}), \quad l=1,3,4, \tag{14}$$

$$\lambda_2(\boldsymbol{x}) \approx \lambda_2^{\mathrm{PC}}(\boldsymbol{x};\boldsymbol{c}) = \exp\left(\sum_{\boldsymbol{\alpha}\in\mathcal{A}_2} c_{2,\boldsymbol{\alpha}}\psi_{\boldsymbol{\alpha}}(\boldsymbol{x})\right), \tag{15}$$

where $\boldsymbol{\mathcal{A}} = \{\mathcal{A}_l : l=1,\dots,4\}$ are the truncation sets defining the basis functions, and $\boldsymbol{c} = \{c_{l,\boldsymbol{\alpha}} : l=1,\dots,4,\ \boldsymbol{\alpha}\in\mathcal{A}_l\}$ are coefficients associated to the bases. For the purpose of clarity, we explicitly express $\boldsymbol{c}$ in the spectral approximations as in $\boldsymbol{\lambda}^{\mathrm{PC}}(\boldsymbol{x};\boldsymbol{c})$ to emphasize that $\boldsymbol{c}$ are the model parameters.

4.2 Estimation of the model parameters

Given the truncation sets $\boldsymbol{\mathcal{A}}$, the coefficients $\boldsymbol{c}$ need to be estimated from data to build the surrogate model. In this paper, as opposed to Zhu and Sudret (2020b) and the vast majority of the literature on stochastic simulators, the simulator is required to be evaluated *only once* on the experimental design $\mathcal{X} = \left\{\boldsymbol{x}^{(1)},\dots,\boldsymbol{x}^{(N)}\right\}$, and the associated model responses are collected in $\mathcal{Y} = \left\{y^{(1)},\dots,y^{(N)}\right\}$. To develop surrogate models in a non-intrusive manner, we propose to use the maximum conditional likelihood estimator:

$$\hat{\boldsymbol{c}} = \arg\min_{\boldsymbol{c}\in\mathcal{C}} \mathsf{L}(\boldsymbol{c}), \tag{16}$$

where

$$\mathsf{L}(\boldsymbol{c}) = \sum_{i=1}^{N} -\log\left(f^{\mathrm{GLD}}\left(y^{(i)};\boldsymbol{\lambda}^{\mathrm{PC}}\left(\boldsymbol{x}^{(i)};\boldsymbol{c}\right)\right)\right). \tag{17}$$

Here, f^{GLD} denotes the PDF of the GLD defined in Eq. (3), and $\mathcal{C}$ is the search space for $\boldsymbol{c}$. The estimator introduced in Eq. (17) can be derived from minimizing the Kullback-Leibler divergence between the surrogate PDF and the underlying true response PDF over $\mathcal{D}_{\boldsymbol{X}}$, see details in Zhu and Sudret (2020b). The advantages of this estimation method are twofold. On the one hand, it removes the need for replications in the experimental design. On the other hand, if a GLaM for a certain choice of $\boldsymbol{c}$ can exactly represent the stochastic simulator, the proposed estimator is *consistent* under mild conditions, as shown in Theorem 1 (see Appendix A.1 for a detailed proof).

Theorem 1. *Let* $\left(\boldsymbol{X}^{(1)},Y^{(1)}\right),\dots,\left(\boldsymbol{X}^{(N)},Y^{(N)}\right)$ *be independent and identically distributed random variables following* $\boldsymbol{X}\sim P_{\boldsymbol{X}}$ *and* $Y(\boldsymbol{x})\sim\mathrm{GLD}\left(\boldsymbol{\lambda}^{\mathrm{PC}}(\boldsymbol{x};\boldsymbol{c}_0)\right)$*. If the following conditions are fulfilled, the estimator defined in Eq. (16) is consistent, that is*

$$\hat{\boldsymbol{c}} \xrightarrow{a.s.} \boldsymbol{c}_0. \tag{18}$$

(i) $P_{\boldsymbol{X}}$ *is absolutely continuous with respect to the Lebesgue measure of* $\mathbb{R}^M$*, i.e., the joint PDF* $f_X(\boldsymbol{x})$ *is Lebesgue-measurable;*

(ii) f_X *has a compact support* $\mathcal{D}_{\boldsymbol{X}}$*;*

(iii) $\mathcal{C}$ *is compact, and* $\boldsymbol{c}_0 \in \mathcal{C}$*;*

(iv) There exists a set $A \subset \mathcal{D}_{\boldsymbol{X}}$ *with* $P_{\boldsymbol{X}}(\boldsymbol{X} \in A) > 0$ *such that* $\forall \boldsymbol{x} \in A$*,* $Y(\boldsymbol{x})$ *does not follow a uniform distribution.*

It is worth remarking that since a GLD can have very fat tails (see Section 2.2), solving the optimization problem may produce response PDFs with unexpected infinite moments when the model is trained on a small data set. To prevent too fat tails (if no prior knowledge suggests it), we apply the following threshold $\lambda_3^{\mathrm{PC}}(\boldsymbol{x}) = \max\left\{\lambda_3^{\mathrm{PC}}(\boldsymbol{x}; \hat{\boldsymbol{c}}), -0.3\right\}$ and $\lambda_4^{\mathrm{PC}}(\boldsymbol{x}) = \max\left\{\lambda_4^{\mathrm{PC}}(\boldsymbol{x}; \hat{\boldsymbol{c}}), -0.3\right\}$, which indicates that we enforce the surrogate PDFs to have finite moments up to order 3 (higher order moments may exist depending on $\hat{\boldsymbol{c}}$). Thresholds larger than -0.3 (e.g., from -0.1 to 0) can be used if the response PDF is known to be light-tailed. Note that when enough data are available, these operations are unnecessary because the resulting model does not exceed the threshold. Although the thresholdings could have been imposed in the model definition in Eq. (14), they change the regularity of the optimization problem, and do not generally improve the performance according to our experience. Therefore, we only use them for post-processing.

Remark 1. *Whilst we consider the simulator to be evaluated only once for each point of the experimental design in this paper, the estimator defined in Eq. (16) is not limited to this type of data. When replications are available, the objective function can be reformulated to*

$$\mathsf{L}(\boldsymbol{c}) = -\sum_{i=1}^{N} \frac{1}{R^{(i)}} \sum_{r=1}^{R^{(i)}} \log\left(f^{\mathrm{GLD}}\left(y^{(i,r)}; \boldsymbol{\lambda}^{\mathrm{PC}}\left(\boldsymbol{x}^{(i)}; \boldsymbol{c}\right)\right)\right), \tag{19}$$

where $R^{(i)}$ *denotes the number of replications at point* $\boldsymbol{x}^{(i)}$*, and* $y^{(i,r)}$ *is the model response for* $\boldsymbol{x}^{(i)}$ *at* r*-th replication. In addition, if* $R^{(i)}$ *is constant for all points* $\boldsymbol{x}^{(i)} \in \mathcal{X}$*, Eq. (19) provides the same estimator as in our previous work (Zhu and Sudret, 2020b).*

4.3 Fitting procedure

In practice, the evaluation of $\mathsf{L}(\boldsymbol{c})$ is not straightforward because the PDF of generalized lambda distributions does not have an explicit form as shown in Eq. (3). Details about the evaluation procedure are given in Zhu and Sudret (2020b). Note that the optimization problem Eq. (16) is subject to complex inequality constraints due to the dependence of the PDF support on $\boldsymbol{\lambda}$ (see Eq. (4)). Given a starting point, we follow the optimization strategy developed in Zhu and Sudret (2020b): We first apply the derivative-based *trust-region* optimization algorithm (Steihaug, 1983) without constraints. If none of the inequality constraints is activated at the optimum, we keep the results as the final estimates. Otherwise, the constrained (1+1)-CMA-ES algorithm (Arnold and Hansen, 2012) available in the software UQLab (Moustapha et al., 2019) is used instead.

Because $\mathsf{L}(\boldsymbol{c})$ is highly nonlinear, a good starting point is necessary to guarantee the convergence of the optimization algorithm. In this section, we introduce a robust method to find a suitable starting point.

According to Eq. (5), the mean $\mu(\boldsymbol{x})$ and the variance function $v(\boldsymbol{x})$ of a generalized lambda model satisfy

$$\begin{aligned} \mu(\boldsymbol{x}) &= \lambda_1^{\mathrm{PC}}(\boldsymbol{x}) + \frac{1}{\lambda_2^{\mathrm{PC}}(\boldsymbol{x})} g\left(\lambda_3^{\mathrm{PC}}(\boldsymbol{x}), \lambda_4^{\mathrm{PC}}(\boldsymbol{x})\right), \\ \log\left(v(\boldsymbol{x})\right) &= -2\log\left(\lambda_2^{\mathrm{PC}}(\boldsymbol{x})\right) + h\left(\lambda_3^{\mathrm{PC}}(\boldsymbol{x}), \lambda_4^{\mathrm{PC}}(\boldsymbol{x})\right), \end{aligned} \tag{20}$$

where we group the dependence of μ and $\log(v)$ on λ_3 and λ_4 into g and h, respectively, for the purpose of simplicity. If $\lambda_3^{\mathrm{PC}}(\boldsymbol{x})$ and $\lambda_4^{\mathrm{PC}}(\boldsymbol{x})$ do not vary strongly on $\mathcal{D}_{\boldsymbol{X}}$, we observe that the variation of the mean and the variance function are mostly dominated by the location parameter $\lambda_1^{\mathrm{PC}}(\boldsymbol{x})$ and the scale parameter $\lambda_2^{\mathrm{PC}}(\boldsymbol{x})$.

Recall that the spectral approximation for $\lambda_2(\boldsymbol{x})$ is on its logarithmic transform. If a PCE can be constructed for $\mu(\boldsymbol{x})$ and $-\frac{1}{2}\log\left(v(\boldsymbol{x})\right)$, the associated coefficients can be used as a preliminary guess for the coefficients of $\lambda_1^{\mathrm{PC}}(\boldsymbol{x})$ and $\lambda_2^{\mathrm{PC}}(\boldsymbol{x})$, respectively. As a result, we first focus on estimating the mean and the variance function as follows:

$$\mu(\boldsymbol{x}) = \sum_{\boldsymbol{\alpha}\in\mathcal{A}_\mu} c_{\mu,\boldsymbol{\alpha}}\psi_{\boldsymbol{\alpha}}(\boldsymbol{x}), \quad v(\boldsymbol{x}) = \exp\left(\sum_{\boldsymbol{\alpha}\in\mathcal{A}_v} c_{v,\boldsymbol{\alpha}}\psi_{\boldsymbol{\alpha}}(\boldsymbol{x})\right),$$

where the form of the variance function implies a multiplicative *heteroskedastic* effect (see Harvey (1976)).

The mean estimation is a classical regression problem. However, since the variance function is also unknown and needs to be estimated, the heteroskedastic effect should be taken into account. Many methods have been developed in statistics and applied science to tackle heteroskedastic regression problems. They can be classified into two groups: one class of methods relies on repeated measurements at given input values (Sadler and Smith, 1985; Ankenman et al., 2009; Murcia et al., 2018) (replication-based), whereas a second class of methods jointly estimates both quantities by optimizing certain functions without the need for replications (Nelder and Pregibon, 1987; Davidian and Carroll, 1987; Goldberg et al., 1997; Marrel et al., 2012). Some studies (Davidian and Carroll, 1987; Marrel et al., 2012) have shown higher efficiency of the second class of methods over the former. This finding supports our pursuit for a replication-free approach. In particular, we opt for feasible generalized least squares (FGLS) (Wooldridge, 2013), which iteratively fits the mean and variance functions in an alternative way.

The details are described in Algorithm 1. In this algorithm, OLS denotes the use of ordinary least-squares, and WLS is the weighted least-squares. $\hat{\boldsymbol{v}}$ corresponds to the set of estimated variances on the design points in $\mathcal{X}$, which are then used as weights in WLS to re-estimate $\boldsymbol{c}_\mu$.

Algorithm 1 Feasible generalized least-squares

$\hat{\boldsymbol{c}}_\mu \leftarrow \mathrm{OLS}\left(\mathcal{X}, \mathcal{Y}\right)$
for $i \leftarrow 1, \ldots, N_{\mathrm{FGLS}}$ **do**
  $\hat{\boldsymbol{\mu}} \leftarrow \sum_{\boldsymbol{\alpha}\in\mathcal{A}_\mu} c_{\mu,\boldsymbol{\alpha}}\psi_{\boldsymbol{\alpha}}(\mathcal{X})$
  $\tilde{\boldsymbol{r}} \leftarrow 2\log\left(|\mathcal{Y} - \hat{\boldsymbol{\mu}}|\right)$
  $\hat{\boldsymbol{c}}_v \leftarrow \mathrm{OLS}\left(\mathcal{X}, \tilde{\boldsymbol{r}}\right)$
  $\hat{\boldsymbol{v}} = \exp\left(\sum_{\boldsymbol{\alpha}\in\mathcal{A}_v} c_{v,\boldsymbol{\alpha}}\psi_{\boldsymbol{\alpha}}(\mathcal{X})\right)$
  $\hat{\boldsymbol{c}}_\mu \leftarrow \mathrm{WLS}\left(\mathcal{X}, \mathcal{Y}, \hat{\boldsymbol{v}}\right)$
end for
Output: $\hat{\boldsymbol{c}}_\mu$, $\hat{\boldsymbol{c}}_v$

After obtaining $\hat{\boldsymbol{c}}_\mu$ and $\hat{\boldsymbol{c}}_v$ from FGLS, we perform two rounds of the optimization procedure described at the beginning of this section to build the GLaM surrogate. First, we set the starting points as $\boldsymbol{c}_1 = \boldsymbol{c}_\mu$, $\boldsymbol{c}_2 = -\frac{1}{2}\boldsymbol{c}_v$ and $\lambda_3^{\mathrm{PC}}(\boldsymbol{x}) = \lambda_4^{\mathrm{PC}}(\boldsymbol{x}) = 0.13$, which corresponds to a normal-like shape. Then, we fit a GLaM with $\lambda_3^{\mathrm{PC}}(\boldsymbol{x})$ $\lambda_4^{\mathrm{PC}}(\boldsymbol{x})$ being only constant, i.e., the coefficients of non-constant basis functions are kept as zeros during the fitting. Finally, we use the resulting estimates as a starting point and construct a final GLaM with all the considered basis functions by solving Eq. (17).

4.4 Truncation schemes

Provided that the bases of $\boldsymbol{\lambda}^{\mathrm{PC}}(\boldsymbol{x})$ are given, we have presented a procedure to construct GLaMs from data in the previous section. However, there is generally no prior knowledge that would help select the truncation sets $\mathcal{A}_l$'s initio. In this section, we develop a method to determine a suitable hyperbolic truncation scheme $\mathcal{A}^{p,q,M}$ presented in Eq. (12) for each component of $\boldsymbol{\lambda}^{\mathrm{PC}}(\boldsymbol{x})$.

As discussed in Section 2, $\lambda_3^{\mathrm{PC}}(\boldsymbol{x})$ and $\lambda_4^{\mathrm{PC}}(\boldsymbol{x})$ control the shape variations of the response PDF. We assume that the shape does not vary in a strongly nonlinear way. Hence, the associated p can be set to a small value, e.g., $p = 1$ in practice. In contrast, $\lambda_1^{\mathrm{PC}}(\boldsymbol{x})$ and $\lambda_2^{\mathrm{PC}}(\boldsymbol{x})$ require possibly larger degree p since their behavior is associated with the mean and the variance function, which might vary nonlinearly over $\mathcal{D}_{\boldsymbol{X}}$. To this end, we modify Algorithm 1 to adaptively find appropriate truncation schemes for $\mu(\boldsymbol{x})$ and $v(\boldsymbol{x})$, which are then used for $\lambda_1(\boldsymbol{x})$ and $\lambda_2(\boldsymbol{x})$, respectively.

Algorithm 2 presents the modified feasible generalized least squares. Instead of using OLS, we apply the *adaptive ordinary least-squares* with degree and q-norm adaptivity (referred to as AOLS) (Marelli and Sudret, 2019). This algorithm builds a series PCEs, each of which is obtained by applying OLS with the truncation set $\mathcal{A}^{p,q,M}$ defined by a particular combination of $p \in \boldsymbol{p}$ and $q \in \boldsymbol{q}$. Then, it selects the truncation scheme for which the associated PCE has the lowest *leave-one-out* error. In the modified FGLS, the truncation set $\mathcal{A}_\mu$ for $\mu(\boldsymbol{x})$ is selected only

Algorithm 2 Modified feasible generalized least-squares

Input: $(\mathcal{X}, \mathcal{Y})$, $\boldsymbol{p}_1$, $\boldsymbol{q}_1$, $\boldsymbol{p}_2$, $\boldsymbol{q}_2$
$\mathcal{A}_\mu,\ \hat{\boldsymbol{c}}_\mu \leftarrow \mathrm{AOLS}\left(\mathcal{X}, \mathcal{Y}, \boldsymbol{p}_1, \boldsymbol{q}_1\right)$
for $i \leftarrow 1, \ldots, N_{\mathrm{FGLS}}$ **do**
    $\hat{\boldsymbol{\mu}} \leftarrow \sum_{\boldsymbol{\alpha} \in \mathcal{A}_\mu} c_{m,\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\mathcal{X})$
    $\tilde{\boldsymbol{r}} \leftarrow 2 \log\left(|\mathcal{Y} - \hat{\boldsymbol{\mu}}|\right)$
    $\mathcal{A}_v^i,\ \hat{\boldsymbol{c}}_v^i,\ \varepsilon_{\mathrm{LOO}}^i \leftarrow \mathrm{AOLS}\left(\mathcal{X}, \tilde{\boldsymbol{r}}, \boldsymbol{p}_2, \boldsymbol{q}_2\right)$
    $\hat{\boldsymbol{v}} \leftarrow \exp\left(\sum_{\boldsymbol{\alpha} \in \mathcal{A}_v} c_{v,\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\mathcal{X})\right)$
    $\hat{\boldsymbol{c}}_\mu \leftarrow \mathrm{WLS}\left(\mathcal{X}, \mathcal{Y}, \mathcal{A}_\mu, \hat{\boldsymbol{v}}\right)$
end for
$i^* = \arg\min\left\{\varepsilon_{\mathrm{LOO}}^i : i = 1, \ldots, N_{\mathrm{FGLS}}\right\}$
Output: $\mathcal{A}_\mu$, $\hat{\boldsymbol{c}}_\mu^{i^*}$, $\mathcal{A}_v^{i^*}$, $\hat{\boldsymbol{c}}_v^{i^*}$

once (before the loop), whereas several truncation schemes $\left\{\mathcal{A}_v^i : i = 1, \ldots, N_{\mathrm{FGLS}}\right\}$ are obtained. We select the one corresponding to the smallest leave-one-out error on the expansion of the variance as the truncation set $\mathcal{A}_v$ for $v(\boldsymbol{x})$. After running Algorithm 2, we apply the two-round optimization strategy described in the previous section to build the GLaM corresponding to the selected truncation schemes.

There are several parameters to be determined in Algorithm 2. In the following examples and applications, we set the candidate degrees $\boldsymbol{p}_1 = \{0, \ldots, 10\}$ for $\lambda_1^{\mathrm{PC}}(\boldsymbol{x})$, and $\boldsymbol{p}_2 = \{0, \ldots, 5\}$ for $\lambda_2^{\mathrm{PC}}(\boldsymbol{x})$. The lists of q-norm are $\boldsymbol{q}_1 = \boldsymbol{q}_2 = \{0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$, and the total number of FGLS iteration is set to $N_{\mathrm{FGLS}} = 10$.

5 Analytical examples

In this section, we validate the proposed algorithm on two analytical examples of input dimension $M = 2$ and $M = 5$, respectively. In both cases, the response PDF is known analytically but does belong to the GLD family, so as to test the flexibility of the proposed method. In addition, we compare the performance of GLaMs with the non-parametric kernel conditional density estimator from the package `np` (Hayfield and Racine, 2008) implemented in R. The latter performs a thorough leave-one-out cross-validation with a multi-start strategy to choose the bandwidths (Hall et al., 2004), which is one of the state-of-the-art kernel estimators. The surrogate model built by this method is referred to as KCDE.

We use Latin hypercube sampling (McKay et al., 1979) to generate the experimental design. The stochastic simulator is only evaluated once for each vector of input parameters. The associated QoI values are used to construct surrogate models with the proposed estimation procedure in Section 4.3.

To quantitatively assess the performance of the surrogate model, we define an error measure

between the underlying model and the emulator by

$$\varepsilon = \mathbb{E}\left[d\left(Y(\boldsymbol{X}), \hat{Y}(\boldsymbol{X})\right)\right], \tag{21}$$

where $Y(\boldsymbol{X})$ is the model response, $\hat{Y}(\boldsymbol{X})$ corresponds to that of the surrogate, and $d\left(Y_1, Y_2\right)$ denotes the contrast measure between the probability distributions of Y_1 and Y_2. In this study, we use the *normalized Wasserstein distance*, defined by

$$d\left(Y_1, Y_2\right) = \frac{d_{\mathrm{WS}}\left(Y_1, Y_2\right)}{\sigma\left(Y_1\right)}, \tag{22}$$

where d_{WS} is the *Wasserstein distance of order two* (Villani, 2000) defined by

$$d_{\mathrm{WS}}\left(Y_1, Y_2\right) \stackrel{\text{def}}{=} \|Q_1 - Q_2\|_2 = \sqrt{\int_0^1 \left(Q_1(u) - Q_2(u)\right)^2 \mathrm{d}u}, \tag{23}$$

where Q_1 and Q_2 are the quantile functions of Y_1 and Y_2, respectively.

Following this definition, the standard deviation σ_{Y_1} can be seen as the Wasserstein distance between the distribution of Y_1 and a degenerate distribution concentrated at the mean value μ_{Y_1}. As a result, the Wasserstein distance normalized by the standard deviation can be interpreted as the ratio of the error related to emulating the distribution of Y_1 by that of Y_2, and to using the mean value μ_{Y_1} as a proxy of Y_1.

Because d_{WS} is invariant under translation, the normalized Wasserstein distance is invariant under both translation and scaling, that is,

$$\forall a \in \mathbb{R} \setminus 0, b \in \mathbb{R} \quad \frac{d_{\mathrm{WS}}\left(a\, Y_1 + b, a\, Y_2 + b\right)}{\sigma(a\, Y_1 + b)} = \frac{d_{\mathrm{WS}}\left(Y_1, Y_2\right)}{\sigma(Y_1)}.$$

To calculate the expectation in Eq. (21), we use Latin hypercube sampling to generate a test set $\mathcal{X}_{\mathrm{test}}$ of size $N_{\mathrm{test}} = 1,000$ in the input space. The normalized Wasserstein distance is calculated for each $\boldsymbol{x} \in \mathcal{X}_{\mathrm{test}}$ and then averaged by N_{test}.

Experimental designs of various size $N \in \{250; 500; 1,000; 2,000; 4,000\}$ are investigated to study the convergence of the proposed method. Each scenario is run 50 times with independent experimental designs to account for statistical uncertainty in the random design. As a consequence, error estimates for each N are represented by box plots.

5.1 Example 1: a two-dimensional simulator

The first example is the *Black-Scholes* model used for stock prices (McNeil et al., 2005):

$$\mathrm{d}S_t = x_1\, S_t\, \mathrm{d}t + x_2\, S_t\, \mathrm{d}W_t, \tag{24}$$

where $\boldsymbol{x} = (x_1, x_2)^T$ are the input parameters, corresponding to the expected return rate and volatility of a stock, respectively. W_t is a standard Wiener process, which represents the source of stochasticity. Equation (24) is a stochastic differential equation whose solution $S_t(\boldsymbol{x})$ is a

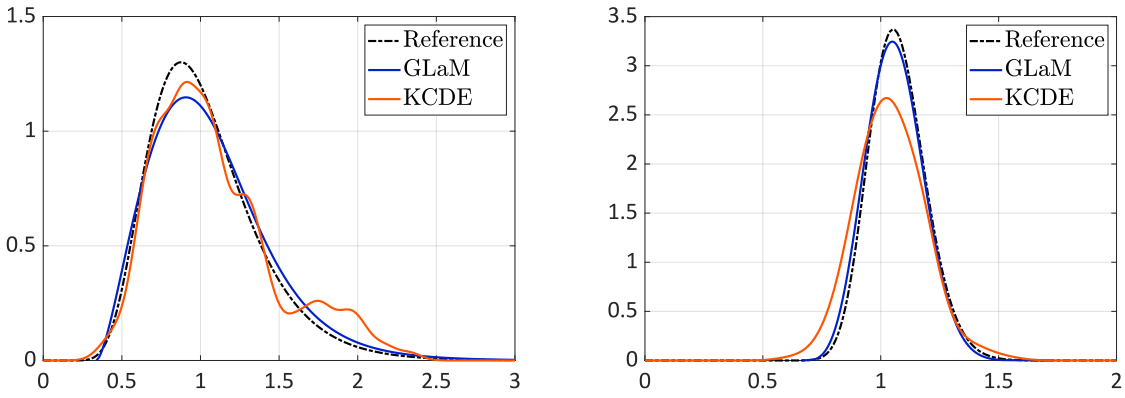
stochastic process for given parameters $\boldsymbol{x}$. Note that we explicitly express $\boldsymbol{x}$ in $S_t(\boldsymbol{x})$ to emphasize that $\boldsymbol{x}$ are input parameters, but the stochastic equation is defined with respect to time. Without loss of generality, we set the initial condition to $S_0(\boldsymbol{x}) = 1$.

In this example, we are interested in $Y(\boldsymbol{x}) = S_1(\boldsymbol{x})$, which corresponds to the stock value in one year i.e., $t = 1$. We set $X_1 \sim \mathcal{U}(0, 0.1)$ and $X_2 \sim \mathcal{U}(0.1, 0.4)$ to represent the input uncertainty, where the ranges are selected based on parameters calibrated from real data (Reddy and Clinton, 2016).

The solution to Eq. (24) can be derived using Itô calculus (Shreve, 2004): $Y(\boldsymbol{x})$ follows a lognormal distribution defined by

$$Y(\boldsymbol{x}) \sim \mathcal{LN}\left(x_1 - \frac{x_2^2}{2}, x_2\right). \tag{25}$$

As the distribution of $Y(\boldsymbol{x})$ is known, it is not necessary to simulate the whole process $S_t(\boldsymbol{x})$ with time integration to evaluate $S_1(\boldsymbol{x})$. Instead, we can directly generate samples from the distribution defined in Eq. (25).



(a) PDF for $\boldsymbol{x} = (0.03, 0.33)^T$ (b) PDF for $\boldsymbol{x} = (0.07, 0.11)^T$

Figure 1: Example 1 – Comparisons of the emulated PDF, $N = 500$.

Figure 1 shows two PDFs predicted by a GLaM and a KCDE built on an experimental design of size $N = 500$. We observe that with 500 model runs, the KCDE yields PDFs with spurious oscillations and demonstrates relatively poor representation of the bulk. In contrast, the GLaM can better approximate the underlying response PDF both in terms of magnitude and shape variations.

For quantitative comparisons, Figure 2 summarizes the error measure Eq. (21) with respect to the size of experimental design. KCDEs show a slow rate of convergence even in this example of dimension two. In contrast, GLaMs reveal high efficiency with a faster decrease of the errors. In terms of the average error, GLaMs outperforms KCDEs for all sizes of experimental design. Furthermore, GLaMs yield an average error near 0.1 for $N = 1,000$, which can be hardly achieved by KCDEs even with four times more model runs.

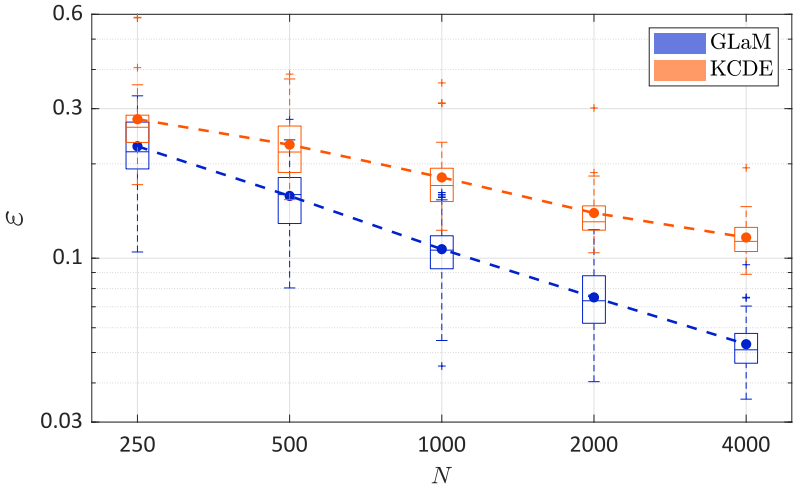


Figure 2: Example 1 – Comparison of the convergence between GLaMs and KCDEs in terms of the normalized Wasserstein distance as a function of the size of the experimental design. The dashed lines denote the average value over 50 repetitions of the full analysis.

5.2 Example 2: a five-dimensional simulator

The second example is given by

$$Y(\boldsymbol{x}) = \mathcal{M}_s(\boldsymbol{x}, \omega) = \mu(\boldsymbol{x}) + \sigma(\boldsymbol{x}) \cdot Z(\omega), \tag{26}$$

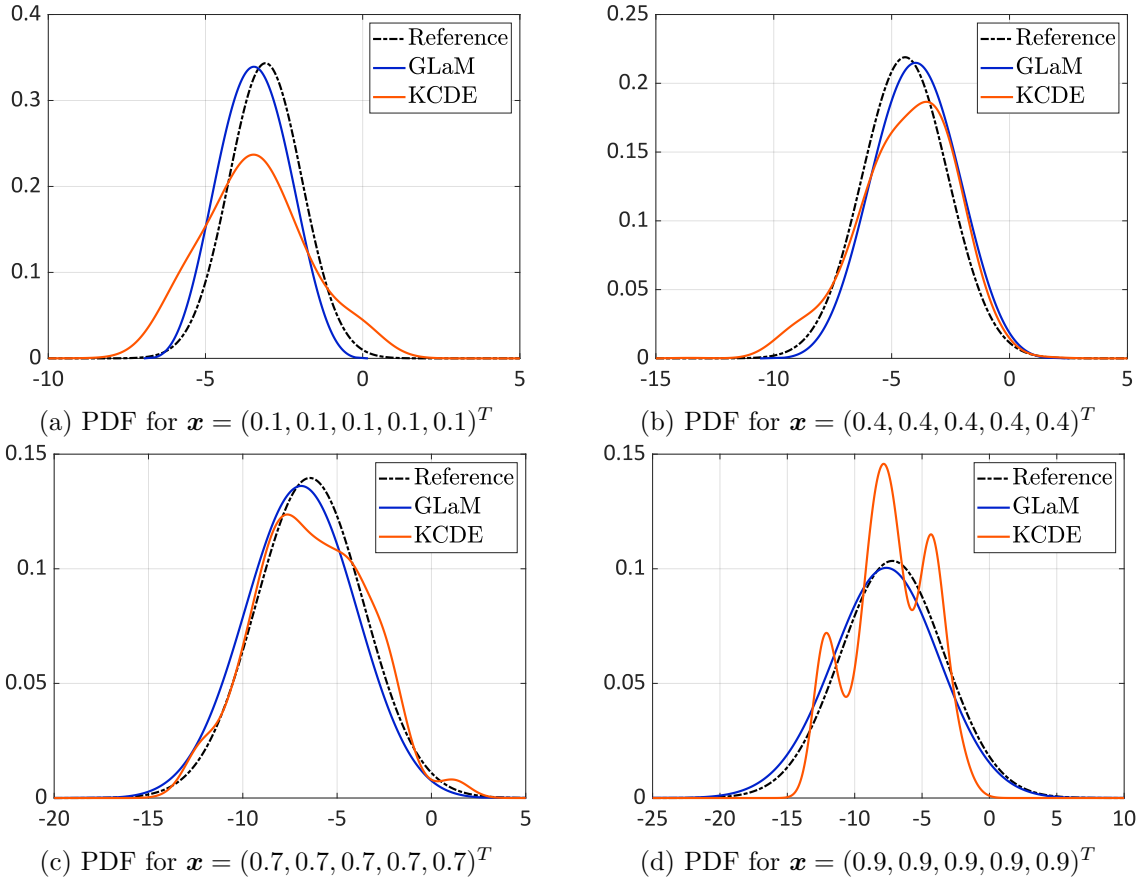
where $\boldsymbol{X} \sim \mathcal{U}\left([0,1]^5\right)$ are the input variables, and $Z \sim \mathcal{N}(0,1)$ is the latent variable that introduces the stochasticity. The simulator has an input dimension of $M = 5$, which is used to show the performance of the proposed method in a moderate-dimensional problem. By definition, $Y(\boldsymbol{x})$ is a Gaussian random variable with mean $\mu(\boldsymbol{x})$ and standard deviation $\sigma(\boldsymbol{x})$ which are defined by

$$\begin{aligned} \mu(\boldsymbol{x}) &= 3 - \sum_{j=1}^{5} j\, x_j + \frac{1}{5}\sum_{j=1}^{5} j\, x_j^3 + \frac{1}{15}\sum_{j=1}^{5} j\, \log\left((x_j^2 + x_j^4)\right) + x_1\, x_2^2 - x_5\, x_3 + x_2\, x_4, \\ \sigma(\boldsymbol{x}) &= \exp\left(\frac{1}{10}\sum_{j=1}^{5} j\, x_j\right), \end{aligned} \tag{27}$$

Thus, this example has a nonlinear mean function and a strong heteroskedastic effect: the variance varies between 1 and 20.

Figure 3 compares the model response PDFs (with different variances) for four input values with those predicted by a GLaM and a KCDE built upon $1,000$ model runs. The results show that the GLaM correctly identifies the shape of the underlying normal distribution among all possible shapes of the GLD. Moreover, it yields a better approximation to the reference PDF, whereas KCDE tends to "wiggle" in Figure 3d (high variance) and overestimate the spread in Figure 3a (low variance).

Similar to the first example, we perform a convergence study for $N \in \{250; 500; 1,000; 2,000; 4,000\}$, the results of which are shown in Fig. 4. In the case of small N, namely $N = 250$,



(a) PDF for $\boldsymbol{x} = (0.1, 0.1, 0.1, 0.1, 0.1)^T$

(b) PDF for $\boldsymbol{x} = (0.4, 0.4, 0.4, 0.4, 0.4)^T$

(c) PDF for $\boldsymbol{x} = (0.7, 0.7, 0.7, 0.7, 0.7)^T$

(d) PDF for $\boldsymbol{x} = (0.9, 0.9, 0.9, 0.9, 0.9)^T$

Figure 3: Example 2 – Comparisons of the emulated PDF, $N = 1,000$. Variance values 1.35, 3.32, 8.17, 14.88 from (a) to (d)

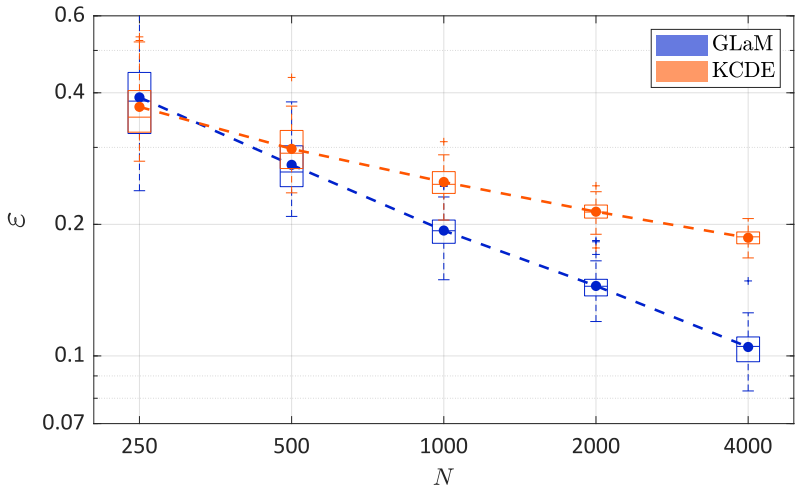


Figure 4: Example 2 – Comparison of the convergence between GLaMs and KCDEs in terms of the normalized Wasserstein distance as a function of the size of the experimental design. The dashed lines denote the average value over 50 repetitions of the full analysis.

both GLaMs and KCDEs perform poorly, with GLaMs showing a similar average error but higher variability. This is explained as follows. Because of the use of AOLS in the modified FGLS procedure, we observe that the total number of coefficients of GLaMs to be estimated varies between 19 to 39 for $N = 250$. Since the GLD is very flexible, a relatively large data set is necessary to provide enough evidence of the underlying PDF shape. Consequently, a small N can lead to overfitting for high-dimensional $\boldsymbol{c}$, but good surrogates can be obtained for more parsimonious models. In contrast, the KCDE always performs a thorough leave-one-out cross-validation strategy to select the bandwidths. Therefore, KCDEs show a slightly more stable estimate for $N = 250$. With N increasing, however, GLaMs converge with a much faster rate and outperforms KCDEs for $N \geq 500$ both in terms of the mean and median of the errors. For $N \geq 1,000$, the average performance of GLaM is even better than the best KCDE model among the 50 repetitions. Besides, in this example of moderate dimensionality, building a KCDE can be time-consuming for large experimental designs, namely $N = 4,000$, due to the bandwidth selection procedure. We observe that depending on the hardware, it can take up to hours for KCDEs, while constructing a GLaM is always in the order of tens of seconds.

6 Applications

In this section, we illustrate the performance of the proposed method on two case studies. Similar to the analytical examples, we use Latin hypercube sampling to generate experimental designs and test sets of $N_{\text{test}} = 1,000$. However, the analytical response PDF is unknown. To characterize the distribution of $Y(\boldsymbol{x})$, we repeatably evaluate the model 10^4 times for $\boldsymbol{x}$.

In addition to the accuracy of response distribution estimations, we also compare some summarizing statistical quantity $b(\boldsymbol{x})$ of the model response $Y(\boldsymbol{x})$, such as the mean $\mathbb{E}[Y(\boldsymbol{x})]$ or variance $\text{Var}[Y(\boldsymbol{x})]$, depending on the focus of the application. Note that $b(\boldsymbol{x})$ is a deterministic function of input variables, and we define the normalized mean-squared error by

$$\epsilon = \frac{\sum_{i=1}^{N_{\text{test}}} \left(b_S^{(i)} - \hat{b}^{(i)}\right)^2}{\sum_{i=1}^{N_{\text{test}}} \left(\hat{b}^{(i)} - \bar{\hat{b}}\right)^2}, \text{ with } \bar{\hat{b}} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \hat{b}^{(i)}, \tag{28}$$

where $b_S^{(i)}$ is the value predicted by the surrogate for $\boldsymbol{x}^{(i)} \in \mathcal{X}_{\text{test}}$, and $\hat{b}^{(i)}$ denotes the quantity estimated from 10^4 replicated runs of the original stochastic simulator for $\boldsymbol{x}^{(i)}$. The error ϵ defined in Eq. (28) indicates how much of the variance of $b(\boldsymbol{X})$ can be explained by $b_S(\boldsymbol{X})$ estimated from surrogate model.

6.1 Application 1: Asian options

In the first application, we apply the proposed method to a financial case study, namely an *Asian option* (Kemna and Vorst, 1990). Such an option, a.k.a. average value option, is a derivative

contract, the payoff of which is contingent on the average price of the underlying asset over a certain fixed time period. Due to the path-dependent nature, an Asian option has a complex behavior, and its valuation is not straightforward, as opposed to European options.

Recall the Black-Scholes model defined in Eq. (24) that represents the evolution of a stock price $S_t(\boldsymbol{x})$. Instead of relying on the stock price on the maturity date $t = T$, the payoff of an Asian call option reads

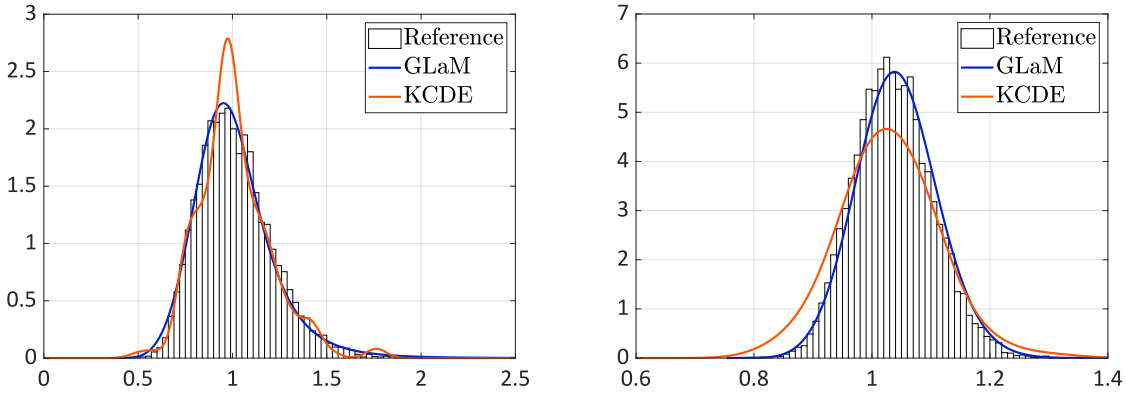
$$C(\boldsymbol{x}) = \max\left\{A_T(\boldsymbol{x}) - K, 0\right\}, \text{ with } A_t(\boldsymbol{x}) = \frac{1}{t}\int_0^t S_u(\boldsymbol{x})\mathrm{d}u. \tag{29}$$

where $A_t(\boldsymbol{x})$ is called the *continuous average process*, and K denotes the *strike price*. Because $A_T(\boldsymbol{x})$ plays an important role in the Asian option modeling Eq. (29), the PDF of $A_T(\boldsymbol{x})$ is of interest in this case study. As in Section 5.1, we set $T = 1$, which corresponds to a one-year inspection period. We choose $X_1 \sim \mathcal{U}(0, 0.1)$ and $X_2 \sim \mathcal{U}(0.1, 0.4)$ for the two input random variables. Unlike $S_1(\boldsymbol{x})$, the distribution of $A_1(\boldsymbol{x})$ cannot be derived analytically. It is necessary to simulate the trajectory of $S_t(\boldsymbol{x})$ to compute $A_1(\boldsymbol{x})$. Based on the Markovian and lognormal properties of $S_t(\boldsymbol{x})$, we apply the following recursive equations for the path simulation with a time step $\Delta t = 0.001$:

$$\begin{aligned} S_0(\boldsymbol{x}) &= 1, \\ S_{t+\Delta t}(\boldsymbol{x}) \mid S_t(\boldsymbol{x}) &\sim \mathcal{LN}\left(\log\left(S_t(\boldsymbol{x})\right) + \left(x_1 - \frac{x_2^2}{2}\right)\Delta t, x_2\sqrt{\Delta t}\right). \end{aligned}$$

Finally, the continuous average defined in Eq. (29) is approximated by the arithmetic mean, that is,

$$A_1(\boldsymbol{x}) = \frac{\sum_{k=1}^{1,000} S_{k\Delta t}(\boldsymbol{x})}{1,000}$$



(a) PDF for $\boldsymbol{x} = (0.03, 0.33)^T$ (b) PDF for $\boldsymbol{x} = (0.07, 0.11)^T$

Figure 5: Asian option – Comparisons of the emulated PDF, $N = 500$

Figure 5 shows two response PDFs predicted by the two surrogate models constructed on an experimental design of $N = 500$. The reference histograms are calculated from 10^4 repeated runs of the simulator for each set of input parameters. We observe that the KCDE is able to capture the

PDF shape for low volatility (in Figure 5b) but exhibits unrealistic fluctuations for high volatility (in Figure 5a). In comparison, the GLaM can well represent the PDF shape in both cases and also approximates more accurately the tails. Finally, Figure 6 confirms the superiority of GLaMs to KCDEs: GLaMs yield smaller average error for all $N \in \{250; 500; 1,000; 2,000; 4,000\}$ and demonstrate a better convergence rate. Moreover, for large experimental designs ($N \geq 2,000$), the average error of GLaMs is nearly half of that of KCDEs.

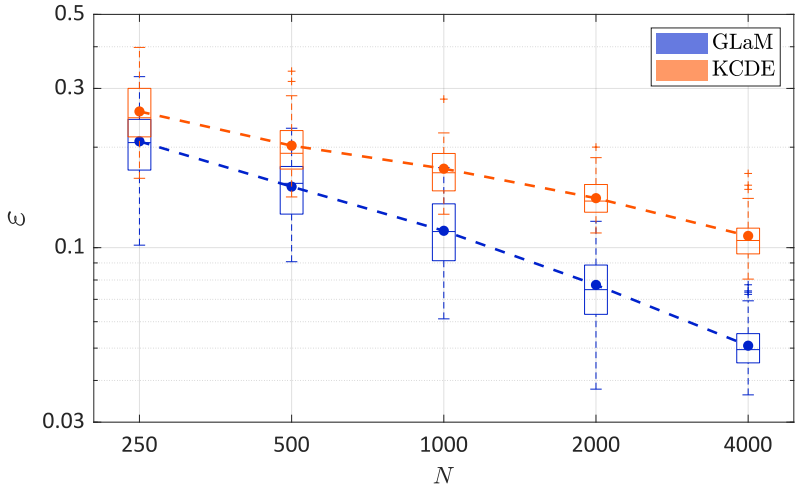


Figure 6: Asian option, average process $A_1(\boldsymbol{x})$ at $T = 1$ year – Comparison of the convergence of GLaMs and KCDEs in terms of the normalized Wasserstein distance as a function of the size of the experimental design. The dashed lines denote the average value over 50 repetitions of the full analysis.

As a second quantity of interest, we consider the expected payoff $\mu_C(\boldsymbol{x}) = \mathbb{E}[C(\boldsymbol{x})]$. This quantity not only is important for making investment decisions but also has a very similar form to the option price (Kemna and Vorst, 1990). The definition Eq. (29) implies that the payoff $C(\boldsymbol{x})$ is a mixed random variable, which has a probability mass at 0 and a continuous PDF on the positive line depending on the strike price K. In the following analysis, K is set to 1.

For GLaMs, $\mu_C(\boldsymbol{x})$ can be calculated by

$$\mu_C(\boldsymbol{x}) = \left(\lambda_1 - \frac{1}{\lambda_2\lambda_3} + \frac{1}{\lambda_2\lambda_4} - K\right)(1 - u_K) + \frac{1}{\lambda_2}\left(\frac{1 - u_K^{\lambda_3+1}}{\lambda_3\,(\lambda_3+1)} - \frac{(1-u_K)^{\lambda_4+1}}{\lambda_4\,(\lambda_4+1)}\right) \tag{30}$$

where λ's are the distribution parameters at $\boldsymbol{x}$, and u_K is the solution of the nonlinear equation

$$Q(u_K; \boldsymbol{\lambda}) = K. \tag{31}$$

with Q being the quantile function defined in Eq. (2).

Figure 7 shows the convergence of estimations of $\mu_C(\boldsymbol{x})$ in terms of the error defined in Eq. (28). The difference between the performance of GLaMs and KCDEs is not as significant as for the distribution estimation of $A_1(\boldsymbol{x})$ in Figure 6. For relatively small data sets, namely $N \leq 500$, both models work poorly: they are only able to explain on average no more than 70% of the

variance of $\mu_C(\boldsymbol{X})$. In addition, GLaMs demonstrate a higher variability of the errors. For larger experimental designs $N \geq 2,000$, however, the performance of GLaMs improves significantly over that of KCDEs. For $N = 4,000$, the average error of GLaMs is twice smaller than that of KCDEs, and the smallest error achieved by GLaMs is one order of magnitude smaller than the best KCDE.

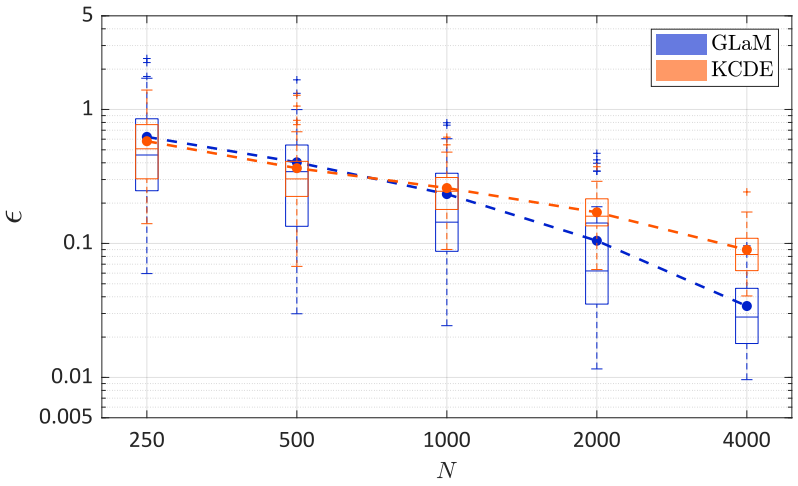


Figure 7: Asian option, expected payoff estimations – Comparison of the convergence of GLaMs and KCDEs in terms of the normalized mean squared error as a function of the size of the experimental design. The dashed lines denote the average value over 50 repetitions of the full analysis.

6.2 Application 2: Stochastic SIR model

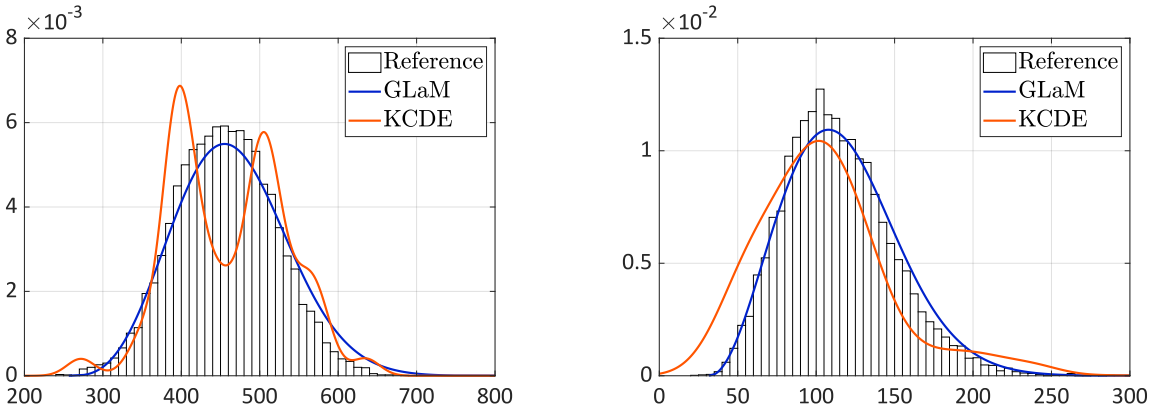
In the second application, we apply the proposed method to a *stochastic Susceptible-Infected-Recovered* (SIR) model in epidemiology (Britton, 2010). This model simulates the spread of an infectious disease, which can help find appropriate epidemiological interventions to minimize social and ethical impacts during the outbreak.

According to the standard SIR model, at time t a population of size P_t contains three groups of individuals: susceptible, infected and recovered, the counts of which are denoted by S_t, I_t and R_t, respectively. These three quantities fully characterize a population configuration at time t. Among the three groups, only susceptible individuals can get infected due to close contact with infected individuals, whereas an infected person can recover and becomes immune to future infections. We consider a fixed population without newborns and deaths, i.e., the total population size is constant, $P_t = P$. As a result, S_t, I_t and R_t satisfy the constraint $S_t + I_t + R_t = P$, and only the time evolution of (S_t, I_t) is necessary to characterize the spread of a disease.

To account for random recoveries and interactions among individuals, stochastic SIR models are usually preferred to represent the epidemic evolution. Without going into details, the model dynamics is briefly summarized as follows. The pair (I_t, S_t) evolves as a continuous-time Markov

process following mutual transition rates β and γ, which denote the contact rate and recovery rate, respectively. The epidemic stops at time $t = T$ where $I_T = 0$, indicating that no further infections can occur. The evolution process is simulated by the *Gillespie algorithm* (Gillespie, 1977). The reader is referred to Britton (2010) for a more detailed presentation of stochastic SIR models.

In this case study, we set the total population equal to $P = 2,000$ and $\beta = \gamma = 0.5$ as in Binois et al. (2018). The initial configuration $\boldsymbol{x} = (S_0, I_0)$ is the vector of input parameters. To account for different scenarios, the input variables $\boldsymbol{X}$ are modeled as $X_1 \sim \mathcal{U}(1200, 1800)$ (initial number of susceptible individuals) and $X_2 \sim \mathcal{U}(20, 200)$ (initial number of infected individuals). The QoI is the total number of newly infected individuals during the outbreak, i.e., $Y(\boldsymbol{x}) = S_T - S_0$.



(a) PDF for $\boldsymbol{x} = (1714, 165)^T$ (b) PDF for $\boldsymbol{x} = (1364, 61)^T$

Figure 8: SIR model – Comparisons of the emulated PDF, $N = 500$

Figure 8 compares two response PDFs estimated by a GLaM and by a KCDE for two sets of initial configurations, using an experimental design of size $N = 500$. The reference histograms are obtained by 10^4 repeated model runs for each $\boldsymbol{x}$. We observe that the PDF shape varies: it changes from symmetric to right-skewed distributions depending on the input variables. This would be difficult to approximate with a simple distribution family such as normal or lognormal. However, the GLaM is able to accurately capture this variation, while KCDE exhibits relatively poor shape representations. More detailed comparisons of the two surrogate models are shown in Figure 9. For all experimental designs, GLaMs clearly outperform KCDEs. For $N \geq 500$, the biggest error of GLaMs is smaller than the smallest error of KCDEs among the 50 repetitions. Finally, to achieves the same accuracy as GLaMs, KCDEs require around 7 times more model runs.

In epidemiological management, the expected value $\mu(\boldsymbol{x}) = \mathbb{E}[Y(\boldsymbol{x})]$ is crucial for decision making (Merl et al., 2009). Therefore, we investigate the accuracy of $\mu(\boldsymbol{x})$ estimations, and the results are in Figure 10. First of all, both GLaM and KCDE can explain more than 90% of the variance in $\mu(\boldsymbol{X})$ for $N = 250$, which implies an overall accurate approximation to the mean function. With increasing N, GLaM shows a more rapid decay of the error. Furthermore, GLaMs built on

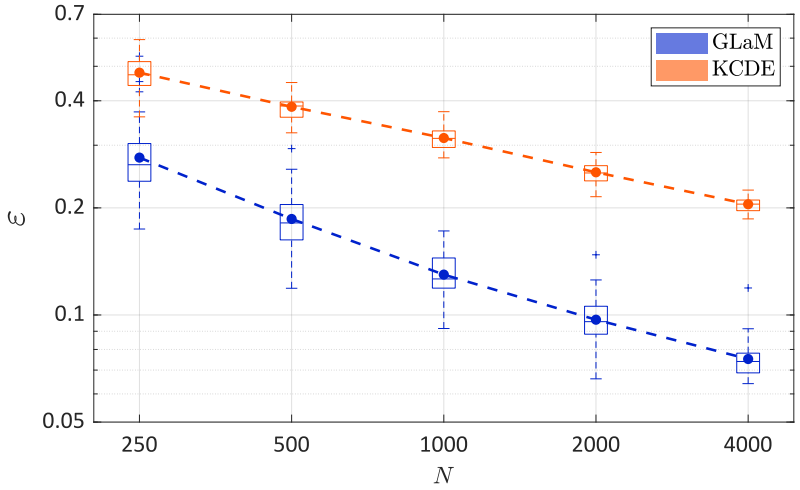


Figure 9: SIR model – Comparison of the convergence between GLaMs and KCDEs in terms of the normalized Wasserstein distance as a function of the size of the experimental design. The dashed line denotes the average value over 50 repetitions of the full analysis.

$N = 1,000$ have a similar (or even slightly better) performance to KCDEs with $N = 4,000$.

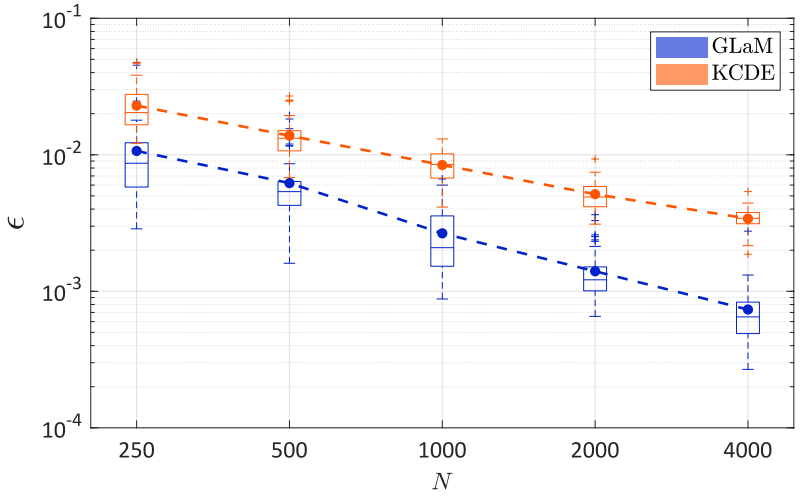


Figure 10: SIR model, mean value estimations – Comparison of the convergence between GLaMs and KCDEs in terms of the normalized mean-squared error as a function of the size of the experimental design. The dashed line denotes the average value over 50 repetitions of the full analysis.

7 Conclusions

This paper presented an efficient and accurate non-intrusive surrogate modeling method for stochastic simulators that does not require replicated runs of the latter. We follow the setting of Zhu and Sudret (2020b), where the generalized lambda distribution is used to flexibly approximate

the response probability density function. The distribution parameters, as functions of the input variables, are approximated by polynomial chaos expansions. In this paper, however, we do not require replicated runs of the stochastic simulator, which provides a more general and versatile approach. We propose the maximum conditional likelihood estimator to construct such a model for given bases functions. This estimation method is shown to be consistent and applicable to data with or without replications. In addition, we modify the feasible generalized least-squares algorithm to select suitable truncation schemes for the distribution parameters, which also provides a good starting point for the subsequent optimization of the likelihood function.

The performance of the new method is illustrated on analytical examples and cases studies in mathematical finance and epidemics. The results show that with a reasonable number of model runs, the developed algorithm can produce surrogate models that accurately approximate the response probability density function and capture the shape variations of the latter with $\boldsymbol{x}$. Considering the normalized Wasserstein distance as an error metric, generalized lambda models always show a better convergence rate than the nonparametric kernel conditional density estimator with adaptive bandwidth selections (from the package `np` in R). Furthermore, the proposed method generally yields more reliable estimates of certain important quantities.

Possible interesting applications of the proposed method to be investigated in future studies include reliability analysis and sensitivity analysis (Zhu and Sudret, 2020a). In terms of the estimation method, we plan to develop algorithms that select only important basis functions based on appropriate model selection criteria to improve the performance of the generalized lambda surrogate model for small data sets.

Acknowledgments

This paper is a part of the project "Surrogate Modeling for Stochastic Simulators (SAMOS)" funded by the Swiss National Science Foundation (Grant #200021_175524), whose support is gratefully acknowledged.

References

Abdallah, I., C. Lataniotis, and B. Sudret (2019). Parametric hierarchical Kriging for multi-fidelity aero-servo-elastic simulators – application to extreme loads on wind turbines. *Prob. Eng. Mech. 55*, 67–77.

Ankenman, B., B. Nelson, and J. Staum (2009). Stochastic Kriging for simulation metamodeling. *Oper. Res. 58*, 371–382.

Arnold, D. and N. Hansen (2012). A (1+1)-CMA-ES for constrained optimisation. In T. Soule and J. H. Moore (Eds.), *Proc. of the Genetic and Evolutionary Computation Conference 2012 (GECCO 2012)*, pp. 297–304.

Azzi, S., B. Sudret, and J. Wiart (2019). Surrogate modeling of stochastic functions-application to computational electromagnetic dosimetry. *Int. J. Uncertainty Quantification 9*, 351–363.

Binois, M., R. Gramacy, and M. Ludkovski (2018). Practical heteroscedastic Gaussian process modeling for large simulation experiments. *J. Comput. Graph. Stat. 27*, 808–821.

Blatman, G. and B. Sudret (2010). An adaptive algorithm to build up sparse polynomial chaos expansions for stochastic finite element analysis. *Prob. Eng. Mech. 25*, 183–197.

Blatman, G. and B. Sudret (2011). Adaptive sparse polynomial chaos expansion based on Least Angle Regression. *J. Comput. Phys. 230*, 2345–2367.

Britton, T. (2010). Stochastic epidemic models: a survey. *Math. Biosci. 225*, 24–35.

Browne, T., B. Iooss, L. Le Gratiet, J. Lonchampt, and E. Rémy (2016). Stochastic simulators based optimization by Gaussian process metamodels – application to maintenance investments planning issues. *Quality Reliab. Eng. Int. 32*(6), 2067–2080.

Davidian, M. and R. Carroll (1987). Variance function estimation. *J. Amer. Stat. Assoc. 400*, 1079–1091.

Efromovich, S. (2010). Dimension reduction and adaptation in conditional density estimation. *J. Amer. Stat. Assoc. 105*, 761–774.

Ernst, O., A. Mugler, H. Starkloff, and E. Ullmann (2012). On the convergence of generalized polynomial chaos expansions. *ESAIM: Math. Model. and Num. Anal. 46*, 317–339.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and Its Applications*, Volume 66 of *Monographs on Statistics and Applied Probability*. Chapman and Hall/CRC.

Freimer, M., G. Kollia, G. Mudholkar, and C. Lin (1988). A study of the generalized Tukey lambda family. *Comm. Stat. Theor. Meth. 17*, 3547–3567.

Gautschi, W. (2004). *Orthogonal polynomials: computation and approximation*. Oxford University Press.

Gillespie, D. (1977). Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem. 81*, 2340–2361.

Goldberg, P., C. Williams, and C. M. Bishop (1997). Regression with input-dependent noise: a gaussian process treatment. In *Proc. 10th Int. Conf. on Advances in neural information processing systems (NIPS10), Colorado, USA*, pp. 493–499.

Hall, P., J. Racine, and Q. Li (2004). Cross-validation and the estimation of conditional probability densities. *J. Amer. Stat. Assoc. 99*, 1015–1026.

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica 50*, 1029–1054.

Harvey, A. (1976). Estimating regression models with multiplicative heteroscedasticity. *Econometrica 44*, 461–465.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models*, Volume 43 of *Monographs on Statistics and Applied Probability*. Chapman and Hall/CRC.

Hayfield, T. and J. Racine (2008). Nonparametric Econometrics: The np Package. *J. Stat. Softw. 27*, 1015–1026.

Jimenez, M., O. Le Maître, and O. Knio (2017). Nonintrusive polynomial chaos expansions for sensitivity analysis in stochastic differential equations. *SIAM J. Uncer. Quant. 5*, 387–402.

Karian, Z. and E. Dudewicz (2000). *Fitting Statistical Distributions: The Generalized Lambda Distribution and Generalized Bootstrap Methods*. Chapman and Hall/CRC.

Kemna, A. and A. Vorst (1990). A pricing method for options based on average asset values. *J. Bank. Finance 14*, 113–129.

Marelli, S. and B. Sudret (2019). UQLab user manual – Polynomial chaos expansions. Technical report, Chair of Risk, Safety and Uncertainty Quantification, ETH Zurich, Switzerland. Report # UQLab-V1.3-104.

Marrel, A., B. Iooss, S. Da Veiga, and M. Ribatet (2012). Global sensitivity analysis of stochastic computer models with joint metamodels. *Stat. Comput. 22*, 833–847.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models*, Volume 37 of *Monographs on Statistics and Applied Probability*. Chapman and Hall/CRC.

McKay, M., R. Beckman, and W. Conover (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics 21*(2), 239–245.

McNeil, A., R. Frey, and P. Embrechts (2005). *Quantitative Risk Management: Concepts, Techniques, and Tools*. Princeton Series in Finance. Princeton, New Jersey: Princeton University Press.

Merl, D., L. R. Johnson, R. B. Gramacy, and M. Mangel (2009). A statistical framework for the adaptive management of epidemiological interventions. *PLoS ONE 4*(6), 5089.

Moustapha, M., C. Lataniotis, P. Wiederkehr, P.-R. Wagner, D. Wicaksono, S. Marelli, and B. Sudret (2019). UQLib user manual. Technical report, Chair of Risk, Safety and Uncertainty Quantification, ETH Zurich, Switzerland. Report # UQLab-V1.3-201.

Moutoussamy, V., S. Nanty, and B. Pauwels (2015). Emulators for stochastic simulation codes. *ESAIM: Math. Model. Num. Anal. 48*, 116–155.

Murcia, J., P. Réthoré, N. Dimitrov, A. Natarajan, J. D. Sørensen, P. Graf, and T. Kim (2018). Uncertainty propagation through an aeroelastic wind turbine model using polynomial surrogates. *Renewable Energy 119*, 910–922.

Nelder, J. and D. Pregibon (1987). An extended quasi-likelihood function. *Biometrika 74*, 221–232.

Newey, W. and D. McFadden (1994). *Large sample estimation and hypothesis testing*, Chapter 36, pp. 2111–2245. Elsevier.

Rasmussen, C. and C. Williams (2006). *Gaussian processes for machine learning* (Internet ed.). Adaptive computation and machine learning. Cambridge, Massachusetts: MIT Press.

Reddy, K. and V. Clinton (2016). Simulating stock prices using geometric Brownian motion: Evidence from Australian companies. *Australasian Accounting, Business and Finance Journal 10*(3), 23–47.

Sadler, W. and M. Smith (1985). Estimation of the response error relationship in immunoassay. *Clinical Chemistry 31*, 1802–1805.

Shreve, S. (2004). *Stochastic Calculus for Finance II.* New York: Springer.

Soize, C. and R. Ghanem (2004). Physical systems with random uncertainties: chaos representations with arbitrary probability measure. *SIAM J. Sci. Comput. 26*(2), 395–410.

Steihaug, T. (1983). The conjugate gradient method and trust regions in large scale optimization. *SIAM J. Num. Anal. 20*(3), 626–637.

Talagrand, M. (1987). The Glivenko-Cantelli problem. *Ann. Prob. 15*, 837–870.

Tsybakov, A. (2009). *Introduction to Nonparametric Estimation.* Springer Series in Statistics. Cambridge, Massachusetts: Springer.

van de Geer, S. (2000). *Empirical processes in M-estimation.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

Villani, C. (2000). *Optimal transport, old and new.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Springer.

Wooldridge, J. (2013). *Introductory Econometrics: A Modern Approach* (5th ed.). Cengage Learning.

Xiu, D. and G. Karniadakis (2002). The Wiener-Askey polynomial chaos for stochastic differential equations. *SIAM J. Sci. Comput. 24*(2), 619–644.

Zhu, X. and B. Sudret (2020a). Global sensitivity analysis for stochastic simulators based on generalized lambda surrogate models. *Reliab. Eng. Sys. Safety.* (Submitted).

Zhu, X. and B. Sudret (2020b). Surrogate models of stochastic simulators using generalized lambda distributions. *Int. J. Uncertainty Quantification.* (Accepted).

A Appendix

A.1 Consistency of the maximum likelihood estimator

In this section, we prove the consistency of the maximum likelihood estimator, as described in Theorem 1. For the ease of derivation, we introduce the following notations:

$$q_{\boldsymbol{c}}(\boldsymbol{x}, y) = f_{Y|\boldsymbol{X}}\left(y|\boldsymbol{\lambda}^{\mathrm{PC}}(\boldsymbol{x}; \boldsymbol{c})\right), \quad p_{\boldsymbol{c}}(\boldsymbol{x}, y) = f_{\boldsymbol{X},Y}(\boldsymbol{x}, y) = f_X(\boldsymbol{x}) q_{\boldsymbol{c}}(\boldsymbol{x}, y),$$

where $q_{\boldsymbol{c}}$ denotes the conditional PDF with model parameters $\boldsymbol{c}$, and $p_{\boldsymbol{c}}$ corresponds to the associated joint PDF. Under this setting, we assume that the true distribution q_0 belongs to the family for a particular set of coefficients $\boldsymbol{c}_0$, i.e., $q_0 = q_{\boldsymbol{c}_0}$ and $p_0 = p_{\boldsymbol{c}_0}$. We denote the probability measure of the probability space of $(\boldsymbol{X}, Y)$ by P_0 and the Lebesgue measure by μ.

The maximum likelihood estimation defined in Eq. (16) belongs to the generalized method of moments (GMM) (Hansen, 1982) for which we define the *loss function* by

$$\ell_{\boldsymbol{c}}(\boldsymbol{x}, y) = -\log\left(q_{\boldsymbol{c}}(\boldsymbol{x}, y)\right) \mathbb{1}_{q_0(\boldsymbol{x},y)>0}(\boldsymbol{x}, y). \tag{32}$$

It holds that

$$\boldsymbol{c}_0 = \arg\min_{\boldsymbol{c}} l(\boldsymbol{c}), \text{ where } l(\boldsymbol{c}) = \mathbb{E}\left[\ell_{\boldsymbol{c}}(\boldsymbol{X}, Y)\right].$$

The maximum likelihood estimator is then defined by

$$\hat{\boldsymbol{c}} = \arg\min_{\boldsymbol{c}} l_n(\boldsymbol{c}), \text{ where } l_n(\boldsymbol{c}) = \frac{1}{n}\sum_{i=1}^{n} \ell_{\boldsymbol{c}}\left(\boldsymbol{X}^{(i)}, Y^{(i)}\right),$$

where l_n is the empirical version of l.

To prove the consistency of a GMM estimator, the *uniform law of large numbers* is usually used. In the case of MLE for the generalized lambda model, classical methods (Newey and McFadden, 1994) to prove the uniform law of large numbers cannot be applied directly, due to the fact that the support of $q_{\boldsymbol{c}}$ can depend on the model parameters $\boldsymbol{c}$, as shown in Eq. (4). To circumvent this problem, we use the techniques suggested by van de Geer (2000) for the proof.

Lemma 1. *Under the conditions described in Theorem 1, we have*

(i) Boundedness: $\sup_{\boldsymbol{c}\in\mathcal{C}} q_{\boldsymbol{c}}(\boldsymbol{x}, y) < +\infty$

(ii) Continuity: $\forall\, \tilde{\boldsymbol{c}} \in \mathcal{C}$*, the map* $\boldsymbol{c} \mapsto q_{\boldsymbol{c}}$ *is continuous at* $\tilde{\boldsymbol{c}}$ *for* μ*-almost all* $(\boldsymbol{x}, y) \in \mathcal{D}_{\boldsymbol{x}} \times \mathbb{R}$

Proof. (i) As the conditions of Theorem 1 indicate that $\mathcal{D}_{\boldsymbol{X}}$ and $\mathcal{C}$ are compact, the two sets are bounded according to the *Heine-Borel theorem*. Hence, the value of $\boldsymbol{\lambda}^{\mathrm{PC}}(\boldsymbol{x}; \boldsymbol{c})$ is also bounded. We denote respectively $\left\{\overline{C}_i, i = 1, \ldots, 4\right\}$ and $\{\underline{C}_i, i = 1, \ldots, 4\}$ as the upper and lower bounds for each component of $\boldsymbol{\lambda}$:

$$\underline{C}_i \leq \lambda_i \leq \overline{C}_i, \quad \forall i = 1, \ldots, 4. \tag{33}$$

In addition, Eq. (15) guarantees that $\lambda_2^{\mathrm{PC}}(\boldsymbol{x};\boldsymbol{c})$ is bounded away from 0, i.e., $\underline{C}_2 > 0$. Consider now Eq. (3) to evaluate the PDF of GLDs. If u in Eq. (3) does not exist in $[0,1]$, $q_{\boldsymbol{c}} = 0$ and thus bounded. For $u \in [0,1]$, we have

$$\frac{\lambda_2}{u^{\lambda_3-1}+(1-u)^{\lambda_4-1}} \leq \frac{\overline{C}_2}{u^{\overline{k}}+(1-u)^{\overline{k}}}, \tag{34}$$

where

$$\overline{k} = \max\left\{\overline{C}_3 - 1, \overline{C}_4 - 1\right\}.$$

Define the function $m(u) = u^{\overline{k}} + (1-u)^{\overline{k}}$, which corresponds to the denominator of Eq. (34). For $\overline{k} = 0$ and 1, $m(u)$ is a constant function equal to 2 and 1, respectively. If $k \neq 0, 1$, the derivative $m'(u) = \overline{k}\left(u^{\overline{k}-1} - (1-u)^{\overline{k}-1}\right)$ is equal to 0 only at $u = 0.5$ in $[0,1]$. As a result, $\min m(u) = \min\{m(0), m(0.5), m(1)\}$. For $\overline{k} < 0$, $\min m(u) = m(0.5) = 2^{1-\overline{k}}$. While for $\overline{k} > 0$, $\min m(u) = \min\{m(0), m(0.5), m(1)\} = \min\left\{1, 2^{1-\overline{k}}\right\}$. Hence, we have $\min m(u) \geq \min\left\{1, 2^{1-\overline{k}}\right\} = C_m$. Taking this property into account, Eq. (34) becomes

$$\frac{\lambda_2}{u^{\lambda_3-1}+(1-u)^{\lambda_4-1}} \leq \frac{\overline{C}_2}{C_m} = C_q. \tag{35}$$

Therefore, $\sup_{\boldsymbol{c}\in\mathcal{C}} q_{\boldsymbol{c}}(\boldsymbol{x}, y) \leq C_q$.

(ii) Next, we prove the continuity. For any $\tilde{\boldsymbol{c}} \in \mathcal{C}$, we classify the points $(\boldsymbol{x}, y) \in \mathcal{D}_{\boldsymbol{x}} \times \mathbb{R}$ into three groups based on their corresponding latent variable $\tilde{u}$: (1) $\tilde{u} \in (0,1)$, (2) $\tilde{u}$ does not exist within $[0,1]$ and (3) $\tilde{u} = 0$ or 1.

For $(\boldsymbol{x}, y)$ in the first class, y is a interior point of the support of the conditional distribution $q_{\tilde{\boldsymbol{c}}}(\boldsymbol{x}, \cdot)$. Thereby, the following equation holds:

$$y = Q(\tilde{u}; \tilde{\boldsymbol{\lambda}}) = \tilde{\lambda}_1 + \frac{1}{\tilde{\lambda}_2}\left(\frac{\tilde{u}^{\tilde{\lambda}_3}-1}{\tilde{\lambda}_3} - \frac{(1-\tilde{u})^{\tilde{\lambda}_4}-1}{\tilde{\lambda}_4}\right), \tag{36}$$

where the distribution parameters $\tilde{\boldsymbol{\lambda}}$ are obtained by evaluating $\boldsymbol{\lambda}^{\mathrm{PC}}(\boldsymbol{x}; \tilde{\boldsymbol{c}})$. The partial derivatives of $Q(u; \boldsymbol{\lambda})$ with respect to all the relevant parameters are

$$\frac{\partial Q}{\partial u} = \frac{1}{\lambda_2}\left(u^{\lambda_3-1} + (1-u)^{\lambda_4-1}\right), \tag{37}$$

$$\frac{\partial Q}{\partial \lambda_1} = 1, \tag{38}$$

$$\frac{\partial Q}{\partial \lambda_2} = -\frac{1}{\lambda_2^2}\left(\frac{u^{\lambda_3}-1}{\lambda_3} - \frac{(1-u)^{\lambda_4}-1}{\lambda_4}\right), \tag{39}$$

$$\frac{\partial Q}{\partial \lambda_3} = \frac{1}{\lambda_2\lambda_3^2}\left(u^{\lambda_3}\ln(u)\lambda_3 - \left(u^{\lambda_3}-1\right)\right), \tag{40}$$

$$\frac{\partial Q}{\partial \lambda_4} = \frac{1}{\lambda_2\lambda_4^2}\left(\left((1-u)^{\lambda_4}-1\right) - (1-u)^{\lambda_4}\ln(1-u)\lambda_4\right). \tag{41}$$

It can be easily observed that Eq. (37) and Eq. (38) are continuous functions of $u \in (0,1)$ and $\boldsymbol{\lambda}$. Although Eq. (39) is undefined for $\lambda_3 = 0$ and $\lambda_4 = 0$, the limit exists according to

l'Hôpital's rule. The same holds for Eq. (40) and Eq. (41). As a result, we can extent Eqs. (39) to (41) by continuity, and thus they become continuous function of $u \in (0,1)$ and $\boldsymbol{\lambda}$. Therefore $Q(u, \boldsymbol{\lambda})$ is continuously differentiable. In addition, Eq. (37) is bounded away from 0. These two properties allows one to apply the *implicit function theorem*, and thus u is a continuous function of $\boldsymbol{\lambda}$ in a neighborhood of $\tilde{\boldsymbol{\lambda}}$, which implies that u is continuous at $\tilde{\boldsymbol{\lambda}}$. According to Eq. (3), the PDF is a continuous function of both u and $\boldsymbol{\lambda}$. Hence, using the continuity shown before, $f_Y(y; \boldsymbol{\lambda})$ is continuous at $\tilde{\boldsymbol{\lambda}}$. Furthermore, $\boldsymbol{\lambda}^{\mathrm{PC}}(\boldsymbol{x}; \boldsymbol{c})$ are C^∞ functions of $\boldsymbol{c}$, and thus $\boldsymbol{\lambda}^{\mathrm{PC}}(\boldsymbol{x}; \boldsymbol{c})$ is continuous at $\tilde{\boldsymbol{c}}$. Combining both the continuity of $f_Y(y; \boldsymbol{\lambda})$ and $\boldsymbol{\lambda}^{\mathrm{PC}}(\boldsymbol{x}; \boldsymbol{c})$, we have that $q_{\boldsymbol{c}}(\boldsymbol{x}, y)$ is continuous at $\tilde{\boldsymbol{c}}$ for the point $(\boldsymbol{x}, y)$.

Now consider a point $(\boldsymbol{x}, y)$ in the second class, which implies that y is outside the support of $q_{\tilde{\boldsymbol{c}}}(\boldsymbol{x}, \cdot)$, say, y is smaller than the lower bound of the support of $q_{\tilde{\boldsymbol{c}}}(\boldsymbol{x}, \cdot)$. In this case, $q_{\tilde{\boldsymbol{c}}}(\boldsymbol{x}, y) = 0$. According to Eq. (4), if the lower bound is finite, it is a continuous function of $\boldsymbol{\lambda}$ and thus continuous at $\tilde{\boldsymbol{c}}$. As a result, for $\boldsymbol{c}$ within a certain neighborhood of $\tilde{\boldsymbol{c}}$, the lower bound is larger than y, which implies $q_{\boldsymbol{x}}(\boldsymbol{x}, y) = 0$ for $\boldsymbol{c}$ in this neighborhood. Thereby, $q_{\boldsymbol{c}}(\boldsymbol{x}, y)$ is continuous at $\tilde{\boldsymbol{c}}$. The analogous reasoning holds for the case where y is bigger than the upper bound of the support.

The last class corresponds to the case where y is located on the endpoint of the support of $q_{\tilde{\boldsymbol{c}}}(\boldsymbol{x}, \cdot)$. By taking $\tilde{u} = 0$ and 1 in Eq. (36) or consider directly Eq. (4), we obtain two associated deterministic functions between $\boldsymbol{x}$ and y. As a result, points of the third class can be represented by two curves in $\mathcal{D}_x \times \mathbb{R}$, whose Lebesgue measure is zero. This closes the proof of continuity. □

Lemma 2. *The class $\mathcal{G}$ defined below satisfies the uniform strong law of large numbers*

$$\mathcal{G} = \left\{ g_{\boldsymbol{c}} = \log\left(\frac{q_{\boldsymbol{c}} + q_0}{2q_0}\right) \mathbb{1}_{q_0>0} : \boldsymbol{c} \in \mathcal{C} \right\} \tag{42}$$

Proof. According to the continuity property in Lemma 1, it is obvious that for all $\tilde{\boldsymbol{c}} \in \mathcal{C}$, the map $\boldsymbol{c} \mapsto g_{\boldsymbol{c}}$ is continuous at $\tilde{\boldsymbol{c}}$ for μ-almost all $(\boldsymbol{x}, y) \in \mathcal{D} \times \mathbb{R}$. By assumption, the probability measure P_0 is absolutely continuous with respect to μ, and thus $g_{\boldsymbol{c}}$ is continuous for P_0-almost all $(\boldsymbol{x}, y) \in \mathcal{D} \times \mathbb{R}$.

Define G as the envelope function of the class $\mathcal{G}$, i.e., $G(\boldsymbol{x}, y) = \sup_{\boldsymbol{c}\in\mathcal{C}} |g_{\boldsymbol{c}}(\boldsymbol{x}, y)|$. Let us prove that $G \in L_1(P_0)$, where $L_1(P_0)$ denotes the set of absolutely integrable functions with respect to P_0.

Taking the boundedness property in Lemma 1 into account, we obtain

$$g_{\boldsymbol{c}}(\boldsymbol{x}, y) \le \log\left(\frac{2C_q}{q_0(\boldsymbol{x}, y)}\right) = \log(2C_q) - \log(q_0(\boldsymbol{x}, y)). \tag{43}$$

Obviously, $g_{\boldsymbol{c}}(\boldsymbol{x}, y) \ge -\log(2)$. Therefore,

$$\begin{aligned} |g_{\boldsymbol{c}}(\boldsymbol{x}, y)| &\le \max\left\{\log(2), |\log(2C_q)| + |\log(q_0(\boldsymbol{x}, y))|\right\} \\ &\le \log(2) + |\log(C_q)| + |\log(q_0(\boldsymbol{x}, y))| \end{aligned}. \tag{44}$$

Because the inequality is independent of $\boldsymbol{c}$, we have

$$\begin{aligned} G(\boldsymbol{x}, y) &\leq \log(2) + |\log(C_q)| + |\log(q_0(\boldsymbol{x}, y))|, \\ \mathbb{E}\left[G(\boldsymbol{X}, Y)\right] &\leq \log(2) + |\log(C_q)| + \mathbb{E}\left[|\log(q_0(\boldsymbol{X}, Y)|\right]. \end{aligned} \tag{45}$$

Now consider the last term in Eq. (45):

$$\begin{aligned} \mathbb{E}\left[|\log(q_0(\boldsymbol{X}, Y)|\right] &= \int_{\mathcal{D}_{\boldsymbol{x}} \times \mathbb{R}} |\log\left(q_0(\boldsymbol{x}, y)\right)| p_0(\boldsymbol{x}, y) \mathrm{d}\boldsymbol{x} \mathrm{d}y \\ &= \int_{\mathcal{D}_{\boldsymbol{x}}} \left(\int_{\mathbb{R}} |\log\left(q_0(\boldsymbol{x}, y)\right)| q_0(\boldsymbol{x}, y) \mathrm{d}y \right) f_{\boldsymbol{X}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \end{aligned} \tag{46}$$

Through a change of variables, the integral within the parenthesis of Eq. (46) can be calculated as

$$B(\boldsymbol{x}) = \int_{\mathbb{R}} |\log\left(q_0(\boldsymbol{x}, y)\right)| q_0(\boldsymbol{x}, y) \mathrm{d}y = \int_0^1 \left| \log\left(\frac{\lambda_2}{u^{\lambda_3 - 1} + (1-u)^{\lambda_4 - 1}} \right) \right| \mathrm{d}u, \tag{47}$$

where $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{\mathrm{PC}}(\boldsymbol{x}; \boldsymbol{c}_0)$. According to Eq. (33), we have

$$\begin{aligned} B(\boldsymbol{x}) &\leq \int_0^1 |\log(\lambda_2)| + \left| \log\left(u^{\lambda_3 - 1} + (1-u)^{\lambda_4 - 1} \right) \right| \mathrm{d}u \\ &\leq k_2 + \int_0^1 \max\left\{ \left| \log\left(u^{\underline{k}} + (1-u)^{\underline{k}} \right) \right|, \left| \log\left(u^{\overline{k}} + (1-u)^{\overline{k}} \right) \right| \right\} \mathrm{d}u \end{aligned}, \tag{48}$$

where

$$k_2 = \max\left\{ \left| \log\left(\overline{C}_2 \right) \right|, \left| \log\left(\underline{C}_2 \right) \right| \right\}, \ \underline{k} = \min\left\{ \underline{C}_3 - 1, \underline{C}_4 - 1 \right\}, \ \overline{k} = \max\left\{ \overline{C}_3 - 1, \overline{C}_4 - 1 \right\}.$$

Using the symmetry of the integrand, we get

$$\begin{aligned} B(\boldsymbol{x}) &\leq k_2 + 2 \cdot \max\left\{ \int_0^{\frac{1}{2}} \left| \log\left(u^{\underline{k}} + (1-u)^{\underline{k}} \right) \right| \mathrm{d}u, \int_0^{\frac{1}{2}} \left| \log\left(u^{\overline{k}} + (1-u)^{\overline{k}} \right) \right| \mathrm{d}u \right\} \\ &\leq k_2 + 2 \cdot \left(\int_0^{\frac{1}{2}} \left| \log\left(u^{\underline{k}} + (1-u)^{\underline{k}} \right) \right| \mathrm{d}u + \int_0^{\frac{1}{2}} \left| \log\left(u^{\overline{k}} + (1-u)^{\overline{k}} \right) \right| \mathrm{d}u \right). \end{aligned} \tag{49}$$

Without loss of generality, we now study the property of the integral

$$\int_0^{\frac{1}{2}} \left| \log\left(u^k + (1-u)^k \right) \right| \mathrm{d}u. \tag{50}$$

For $k = 0$, Eq. (50) is equal to $\frac{1}{2}\log(2)$. For $k > 0$, we have $u^k \leq (1-u)^k$, and thus

$$\begin{aligned} \int_0^{\frac{1}{2}} \left| \log\left(u^k + (1-u)^k \right) \right| \mathrm{d}u \leq \int_0^{\frac{1}{2}} \left| \log\left(2(1-u)^k \right) \right| \mathrm{d}u &\leq \frac{1}{2}\log(2) - \int_0^{\frac{1}{2}} k \log(1-u) \mathrm{d}u \\ &= \frac{1}{2}\log(2) + \frac{k}{2}\left(1 - \log(2)\right) \end{aligned}. \tag{51}$$

Through similar calculation, for $k < 0$, we have

$$\begin{aligned} \int_0^{\frac{1}{2}} \left| \log\left(u^k + (1-u)^k \right) \right| \mathrm{d}u \leq \int_0^{\frac{1}{2}} \left| \log\left(2u^k \right) \right| \mathrm{d}u &\leq \frac{1}{2}\log(2) + \int_0^{\frac{1}{2}} k \log(u) \mathrm{d}u \\ &= \frac{1}{2}\log(2) + \frac{-k}{2}\left(\log(2) + 1\right) \end{aligned}. \tag{52}$$

As a result, Eq. (50) is finite. More precisely,

$$\int_0^{\frac{1}{2}} \left|\log\left(u^k + (1-u)^k\right)\right| \mathrm{d}u \leq \frac{1}{2}\log(2) + \frac{|k|}{2}\left(\log(2)+1\right). \tag{53}$$

Equation (53) implies

$$B(\boldsymbol{x}) \leq k_2 + \log(2) + \left(|\underline{k}| + \left|\overline{k}\right|\right)\left(\log(2)+1\right) = C_B. \tag{54}$$

By inserting Eq. (54) into Eq. (46), we obtain

$$\mathbb{E}\left[\left|\log(q_0(\boldsymbol{X}, Y)\right|\right] \leq C_B. \tag{55}$$

Then, according to Eq. (45), the envelope function G fulfills

$$\begin{aligned} \mathbb{E}\left[G(\boldsymbol{X}, Y)\right] &\leq \log(2) + \left|\log(C_q)\right| + \mathbb{E}\left[\left|\log\left(q_0(\boldsymbol{X}, Y)\right)\right|\right] \\ &= \log(2) + \left|\log(C_q)\right| + C_B < +\infty \end{aligned}. \tag{56}$$

Since G is always positive according to its definition, Eq. (56) means $G \in L_1(P_0)$. The continuity and the property of the envelope function G shown above allow applying van de Geer (2000, Lemma 3.10), which guarantees that $\mathcal{G}$ satisfies the uniform weak law of large numbers:

$$\sup_{\boldsymbol{c}\in\mathcal{C}} \left(\frac{1}{n}\sum_{i=1}^{n} g_{\boldsymbol{c}}\left(\boldsymbol{X}^{(i)}, Y^{(i)}\right) - \mathbb{E}\left[g_{\boldsymbol{c}}\left(\boldsymbol{X}, Y\right)\right]\right) \xrightarrow[n\to+\infty]{P} 0. \tag{57}$$

Finally, Talagrand (1987, Theorem 22) extends the convergence to *almost surely*, which is the uniform strong law of large numbers. □

Now, we have all the ingredients to prove Theorem 1.

Proof. Following van de Geer (2000, Lemma 4.1, 4.2), it can be easily shown that

$$0 \leq \int_{\mathcal{D}_{\boldsymbol{x}}} h^2\left(q_{\hat{\boldsymbol{c}}}, q_0 \mid \boldsymbol{x}\right) f_{\boldsymbol{X}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \leq 8\left(\sum_{i=1}^{N} g_{\hat{\boldsymbol{c}}}\left(\boldsymbol{X}^{(i)}, Y^{(i)}\right) - \mathbb{E}\left[g_{\hat{\boldsymbol{c}}}\left(\boldsymbol{X}, Y\right)\right]\right), \tag{58}$$

where the Hellinger distance is given by

$$h^2\left(q_{\hat{\boldsymbol{c}}}, q_0 \mid \boldsymbol{x}\right) = \frac{1}{2}\int_{\mathbb{R}} \left(\sqrt{q_{\hat{\boldsymbol{c}}}(\boldsymbol{x}, y)} - \sqrt{q_0(\boldsymbol{x}, y)}\right)^2 \mathrm{d}y.$$

According to Lemma 2, Eq. (58) implies

$$\int_{\mathcal{D}_{\boldsymbol{x}}} h^2\left(q_{\hat{\boldsymbol{c}}}, q_0 \mid \boldsymbol{x}\right) f_{\boldsymbol{X}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \xrightarrow{\text{a.s.}} 0, \tag{59}$$

which is called the *Hellinger consistency*.

We define the function

$$R(\boldsymbol{c}) = \int_{\mathcal{D}_{\boldsymbol{x}}} h^2\left(q_{\boldsymbol{c}}, q_0 \mid \boldsymbol{x}\right) f_{\boldsymbol{X}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{60}$$

According to Lemma 1, $\forall \tilde{\boldsymbol{c}} \in \mathcal{C}$, the map $\boldsymbol{c} \mapsto \left(\sqrt{q_{\boldsymbol{c}}} - \sqrt{q_0}\right)^2$ is continuous at $\tilde{\boldsymbol{c}}$ for $\forall \boldsymbol{x} \in \mathcal{D}_x$ and almost all $y \in \mathbb{R}$. Since $\left(\sqrt{q_{\boldsymbol{c}}} - \sqrt{q_0}\right)^2 \leq q_{\boldsymbol{c}} + q_0$, and $\int_{\mathbb{R}} (q_{\boldsymbol{c}} + q_0)\, \mathrm{d}y = 2 < +\infty$, the map $\boldsymbol{c} \mapsto h^2\left(q_{\boldsymbol{c}}, q_0 \mid \boldsymbol{x}\right)$ is continuous for all $\boldsymbol{x} \in \mathcal{D}_{\boldsymbol{x}}$, which is guaranteed by the *generalized Lebesgue dominated convergence theorem*. Similarly, the map $\boldsymbol{c} \mapsto R(\boldsymbol{c})$ is also continuous.

Without going into lengthy discussions, it can be shown that the GLD is *not identifiable* only for $\lambda_3 = \lambda_4 = 1$ and $\lambda_3 = \lambda_4 = 2$. In other words, by excluding two points in the $\lambda_3 - \lambda_4$ plane, different values of $\boldsymbol{\lambda}$ lead to different distributions. Note that the two exceptions are the only two cases where the corresponding distributions are uniform distributions. As a result, the last condition in Theorem 1 excludes the non-identifiable cases. Furthermore, $\boldsymbol{\lambda}^{\mathrm{PC}}(\boldsymbol{x}; \boldsymbol{c})$ are polynomials in $\boldsymbol{x}$ and linear in $\boldsymbol{c}$. Therefore, for $\boldsymbol{c} \neq \tilde{\boldsymbol{c}}$, $\boldsymbol{\lambda}^{\mathrm{PC}}(\boldsymbol{x}; \boldsymbol{c})$ and $\boldsymbol{\lambda}^{\mathrm{PC}}\left(\boldsymbol{x}; \tilde{\boldsymbol{c}}\right)$ are not identical for μ-almost all $\boldsymbol{x} \in \mathbb{R}^M$, and thus for $P_{\boldsymbol{X}}$-almost all $\boldsymbol{x} \in \mathcal{D}_{\boldsymbol{X}}$. Hence, there exists a set $\Omega_{\boldsymbol{x}}$ with $P_{\boldsymbol{X}}(\Omega_{\boldsymbol{x}}) > 0$ such that as long as $\boldsymbol{c} \neq \boldsymbol{c}_0$, $h\left(q_{\boldsymbol{c}}, q_0 \mid \boldsymbol{x}\right) > 0\ \forall \boldsymbol{x} \in \Omega_{\boldsymbol{x}}$, which implies the uniqueness. Finally, combining Eq. (59) with the continuity and uniqueness of $R(\boldsymbol{c})$, we have $\hat{\boldsymbol{c}} \xrightarrow{\text{a.s.}} \boldsymbol{c}_0$. $\square$