International Journal of Statistics: Advances in Theory and Applications
Vol. 1, Issue 2, 2020, Pages 257-274
Published Online on December 11, 2019

http://jyotiacademicpress.org

ON CONVERGENCE OF MOMENTS FOR APPROXIMATING PROCESSES AND APPLICATIONS TO SURROGATE MODELS LIKE DEEP LEARNING NETWORKS

ANSGAR STELAND

Institute of Statistics
RWTH Aachen University
Aachen
Germany
e-mail: steland@stochastik.rwth-aachen.de

Abstract

We study criteria for a pair $(\{X_n\}, \{Y_n\})$ of approximating processes which guarantee closeness of moments by generalizing known results for the special case that $Y_n = Y$ for all n and X_n converges to Y in probability. This problem especially arises when working with surrogate models, e.g., to enrich observed data by simulated data, where the surrogates Y_n's are constructed to justify that they approximate the X_n's. We first discuss that case of sequences of random variables. Since this framework does not cover many applications where surrogate models such as deep neural networks are used to approximate more general stochastic processes, we extend the results to the more general framework of random fields of stochastic processes. This framework especially covers image data and sequences of images. We show that uniform integrability is sufficient, and this holds even for the case of processes provided they satisfy a weak stationarity condition.

2020 Mathematics Subject Classification: 68T01, 62M40, 62M45, 68T05.

Keywords and phrases: convergence of moments, data science, deep learning, surrogate model, stochastic approximation, machine learning, uncertainty quantification, uniform integrability.

Communicated by Shuzhen Yang.

Received March 14, 2018; Revised May 1, 2018

1. Introduction

Suppose we observe a random phenomenon, $X_n, n \geq 1,$ e.g., representing the outcome of a statistical experiment. Let us further assume that an approximation, Y_n, for X_n is available, such as a prediction of X_n based on a (estimated) prediction model or a computer simulation. The later case is receiving increasing interest in the field of uncertainty quantification, where observed data is enriched by data obtained from simulations, governed by so-called surrogate models, which are typically obtained from physical knowledge, by design-of-experiment methods, or (non-) parametric estimation from (small) random samples. Simulated data from surrogate models, which rely on random number generators as discussed in, e.g., [2], are often much cheaper to obtain than real data. Examples for surrogate models are linear models, Gaussian processes and deep learning networks. In this case, Y_n represents the (observable) output of the simulation and X_n the (unobserved) artificial random variable representing the outcome of the experiment not conducted. In such applications the connection between the true and the surrogate model and therefore between X_n and Y_n can be rather loose, such that the approximation error can not be analyzed rigorously and assumptions about it have to be made.

Assuming that (uniform) convergence in probability as a minimal requirement holds, the question arises under which conditions the moments of Y_n are close to the moments of X_n. We study this issue for the case of random variables and the substantially more general framework of random fields of stochastic processes, i.e., families of random variables indexed by a parameter $\lambda \in \Lambda$ (such as time) and an index $\boldsymbol{n}$ (such as discrete spatial location). In this way, the results are general enough to cover various applications including high-dimensional settings and image data.

The paper is organized as follows. Section 2 reviews the basic notions and provides the results for sequences of random variables. The results are applied to autoregressive processes of order 1 as arising, for instance, in signal processing when filtering input signals. Here the surrogate model is obtained by estimating unknowns arising in the true model equation describing the signal processing from disturbed observables and truncating the infinite series, in order to obtain a simple approximation. In Section 3, the general framework of random fields of stochastic processes indexed by a normed space is studied. Provided a weak stationarity condition is satisfied, the sufficiency of uniform integrability can be established. As two applications we study in Section 4 deep learning neural nets and Gaussian process kriging, two popular general purpose frameworks to construct surrogate models.

2. Criteria for Random Variable

Let us first consider the case of sequences of random variables.

2.1. Uniform integrability

Suppose that $\{X, X_n : n \geq 1\}$ is a sequence of random variables defined on a common probability space $(\Omega, \mathcal{A}, P)$. Suppose that $X_n \to X$, as $n \to \infty$, in probability. Then it is known that the convergence of the moments,

$$E(X_n) \to E(X), \quad n \to \infty,$$

follows, if $\{X_n\}$ is uniformly integrable; this result is usually stated for almost sure convergence, but it holds for convergence in probability as well. Recall that $\{X_n\}$ is called *uniformly integrable*, if and only if

$$\lim_{A \to \infty} \int_{|X_n| > A} |X_n| dP = E[|X_n| 1(|X_n| > A)] = 0.$$

Here and in what follows $1(\ )$ denotes the indicator function. Uniform integrability is equivalent to $\sup_{n \geq 1} E|X_n| < \infty$ and

For every $\varepsilon > 0$ there exists $\delta(\varepsilon) > 0$ such that for any

$$A \in \mathcal{A} : P(A) < \eta \Rightarrow \int_A |X_n| dP < \varepsilon. \tag{2.1}$$

It is well known that the above characterizations are optimal in the sense that if X_n converges to X in probability and the r-th absolute moments, $E|X_n|^r$, converge to $E|X|^r$, $0 < r < \infty$, then $\{X_n\}$ is uniformly integrable, [1]. Uniform integrability is what is needed to make the step from convergence in probability to convergence of moments. It is worth mentioning that a straightforward way to establish uniform integrability is to verify the sufficient condition

$$\sup_{n \geq 1} E|X_n|^{1+\delta} < \infty,$$

for some $\delta > 0$.

There is an interesting relationship to stochastic order relations. Let X_1 and X_2 be positive random variables. X_1 is less or equal than X_2 in the *increasing convex order*, denoted by

$$X_1 \leq_{ic} X_2,$$

if

$$E\varphi(X_1) \leq E\varphi(X_2),$$

for all non-decreasing convex functions $\varphi : [0, \infty) \to [0, \infty)$. Equivalently, $H_1(t) \leq H_2(t)$ for all $t \geq 0$, where $H_i(t) = \int_t^\infty (1 - F_i(u)) du, i = 1, 2,$ are the integrated survivor functions, see [6]. A sequence $\{X_n : n \geq 1\}$ of random variables is *ic-bounded* by a random variable Y, if

$$|X_n| \leq_{ic} Y \qquad \text{for all } n \geq 1.$$

In [5], it has been shown that $\{X_n : n \geq 1\}$ is uniformly integrable, if and only if $\{X_n\}$ is ic-bounded by an integrable random variable.

2.2. Convergence of moments for approximations

As explained in the Introduction, it is often not realistic to assume that a given sequence $\{X_n\}$ converges to some random variable X, but instead we can assume the existence of an approximating sequence $\{Y_n : n \geq 1\}$ of random variables. Then, at best, we can achieve *closeness* of the moments. Since in many present day real applications the connection between these sequences is somewhat loose in the sense that it is not possible to analyze the accuracy of the approximation rigorously, one has to *assume* appropriate conditions which allow to replace the moments of X_n by the moments of Y_n. The question arises, whether uniform integrability still suffices for this purpose.

The following result shows that the moments of Y_n are close to the moments of X_n, if Y_n approximates X_n and both series are uniformly integrable. Denote $\|X\|_r = (E|X|^r)^{1/r}$ for a random variable X and $0 < r < \infty$.

Theorem 2.1. *Let* $0 < r < \infty$. *Suppose that* $\{|X_n|^r : n \geq 1\}$ *and* $\{|Y_n|^r : n \geq 1\}$ *are uniformly integrable with*

$$|X_n - Y_n| \xrightarrow{P} 0,$$

as $n \to \infty$. *Then the following assertions hold*:

(i) $E|X_n - Y_n|^r \to 0$, *as* $n \to \infty$, *for* $0 < r \leq 1$.

(ii) $|E|X_n|^r - E|Y_n|^r| \to 0$, *as* $n \to \infty$, *for* $0 < r \leq 1$.

(iii) $|E(X_n) - E(Y_n)| \to 0$, *as* $n \to \infty$, *if* $r = 1$.

(iv) $|(E|X_n|^r)^{1/r} - (E|Y_n|^r)^{1/r}| \to$, *as* $n \to \infty$, *for* $0 < r < \infty$.

Let us consider the following example where we assume concrete models for X_n and Y_n. Suppose that X_n is a linear filter processing a random input sequence of i.i.d. innovations $\epsilon_t, t \geq 0,$ with mean zero, finite fourth moment and common variance $\sigma^2 \in (0, \infty)$. Further suppose that X_n is given by an autoregressive process of order 1 with known autoregressive parameter $\rho \in (-1, 1),$ given by

$$X_n = \mu + \sum_{j=0}^{\infty} \rho^j \epsilon_{n-j},$$

where the mean μ is unkown to us. The process $X_n,$ however, can only be observed with a (deterministic) uncertainty $e_n,$ i.e., we have at our disposal the process

$$X_{n,obs} = X_n + e_n,$$

where $e_n, n \geq 1,$ is assumed to be a sequence of constants with $\frac{1}{n}\sum_{i=1}^{n} e_i \to 0,$ as $n \to \infty.$ Consider the approximation Y_n following the surrogate model:

$$Y_n = \overline{X}_{n,obs} + \sum_{j=0}^{q_n} \rho^j \epsilon_{n-j},$$

for some sequence $q_n, n \geq 1,$ of natural numbers with $q_n \to \infty,$ where $\overline{X}_{n,obs} = \frac{1}{n}\sum_{i=1}^{n} X_{i,obs}.$ This means, the surrogate model is obtained by estimating the unknown mean by the average of the observed data and truncating the infinite sum to obtain a surrogate model from which allows for fast computations. Then

$$|X_n - Y_n| \leq \left|\frac{1}{n}\sum_{i=1}^{n} e_i\right| + |\overline{X}_n - \mu| + \left|\sum_{j>q_n} \rho^j \epsilon_{n-j}\right|.$$

The first term on the right-hand is $o_P(1)$ by virtue of the weak law of large numbers for time series, and the second term can be bounded by

$$P\left(\left|\sum_{j>q_n} \rho^j \epsilon_{n-j}\right| > \varepsilon\right) \leq \frac{\sigma^2}{\varepsilon^2} \sum_{j>q_n} |\rho|^{2j} \to 0,$$

for any $\varepsilon > 0$. Hence,

$$|X_n - Y_n| \xrightarrow{P} 0,$$

as $n \to \infty$. Further, by our moment conditions, X_n and Y_n are uniformly integrable, and $\frac{1}{n}\sum_{i=1}^{n} X_i$ is uniformly integrable, if X_n has this property, see, e.g., [1]. Hence, the above theorem applies. It is worth mentioning that the i.i.d. assumption for the ϵ_t's can be relaxed.

2.3. Proof

Proof of Theorem 2.1. By the c_r-inequality (which follows from the inequality $|x - y|^r \leq 2^r(|x| + |y|)$ for real numbers x, y and $r > 0$),

$$|X_n - Y_n|^r \leq 2^r(|X_n|^r + |Y_n|^r),$$

we may conclude that $\{|X_n - Y_n|^r : n \geq 1\}$ is uniformly integrable. Let $\varepsilon > 0$. We have

$$E|X_n - Y_n|^r \leq \varepsilon^r + E[|X_n - Y_n|^r 1(|X_n - Y_n| > \varepsilon)],$$

leading to

$$\limsup_{n\to\infty} E|X_n - Y_n|^r \leq \varepsilon^r + \limsup_{n\to\infty} E[|X_n - Y_n|^r 1(|X_n - Y_n| > \varepsilon)].$$

We will show that the second term vanishes. By uniform integrability of $|X_n - Y_n|$, there exists $\delta(\varepsilon) > 0$ such that for any event A with $P(A) < \delta(\varepsilon)$ we have $E[|X_n - Y_n|^r 1_A] < \varepsilon$. Since $|X_n - Y_n| \xrightarrow{P} 0$, as $n \to \infty$, there exists $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$

$$P(|X_n - Y_n| > \varepsilon) < \delta(\varepsilon).$$

Therefore,

$$E[|X_n - Y_n|^r 1(|X_n - Y_n| > \varepsilon)] < \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, (i) follows. To show (ii), apply the inequality $|x + y|^r \leq |x|^r + |y|^r$ to obtain $||x|^r - |y|^r| \leq |x - y|^r$, such that $|E|X_n|^r - E|Y_n|^r| \leq E|X_n - Y_n|^r$, which establishes (ii). (iii) follows by linearity, $|E(X_n) - E(Y_n)| \leq E|X_n - Y_n|$. Lastly, apply Minkowski's inequality to obtain

$$\|X_n\|_r - \|Y_n\|_r = \|X_n - Y_n + Y_n\|_r - \|Y_n\|_r \leq \|X_n - Y_n\|_r,$$

and

$$\|Y_n\|_r - \|X_n\|_r = \|Y_n - X_n + X_n\|_r - \|X_n\|_r \leq \|Y_n - X_n\|_r.$$

□

3. Criteria for Random Fields of Stochastic Processes

In surrogate modelling applications, one often considers models for high-dimensional objects, e.g., images or sequences of images, stochastic differential equations whose numerical solution can be intractable due to excessive computational costs, or nonlinear regressions with a large number of regressors as arising in genetics and finance. Here surrogate models are constructed which allow for efficient computations and have good approximation properties. Examples are deep learning neural networks and the Gaussian process framework, which we discuss in the next section.

Therefore, let us now study a more general theoretical framework, namely, random fields of stochastic processes, which covers those special cases.

3.1. Preliminaries

Recall that a stochastic process is a family $\{X(\lambda) : \lambda \in \Lambda\}$ of random variables $X(\lambda)$ defined on a probability space $(\Omega, \mathcal{A}, P)$. Here Λ is an arbitrary index set. Such a process is called (strictly) stationary, if the (multivariate) distributions associated to $\lambda_1, \ldots, \lambda_k \in \Lambda$, $k \in \mathbb{N}$ arbitrary, are shift invariant in the sense that for all h such that $\lambda_j + h \in \Lambda$, $j = 1, \ldots, k$, it holds

$$P_{(X(\lambda_1), \ldots, X(\lambda_k))} = P_{(X(\lambda_1 + h), \ldots, X(\lambda_k + h))}.$$

This clearly implies stationarity of the one-dimensional marginal distribution $P_{X(\lambda)}$, but not vice-versa.

A random field of dimension $q \in \mathbb{N}$ is a family of random elements indexed by a multiindex $\boldsymbol{i} = (i_1, \ldots, i_q)'$ ranging through some set $I \subseteq \mathbb{N}^q$. We assume that those random elements attain values in some normed space E with norm $|\cdot|$.

For example, a two-dimensional random field of random variables indexed by $\boldsymbol{i} \in \{1, \ldots, n_1\} \times \{1, \ldots, n_2\}$, i.e., a matrix of dimension $n_1 \times n_2$ with random entries, is a natural model for an image of resolution $n_1 \times n_2$.

For two random fields $\{X_{\boldsymbol{n}} : \boldsymbol{n} \geq 1\}$ and $\{X_{\boldsymbol{n}} : \boldsymbol{n} \geq \mathbf{1}\}$, where $\mathbf{1} = (1, \ldots, 1)'$, the convergence $|X_{\boldsymbol{n}} - Y_{\boldsymbol{n}}| \xrightarrow{P} 0$, as $\boldsymbol{n} \to \infty$, is defined as follows: For every $\varepsilon > 0$ and every $\delta > 0$ there exits $N \in \mathbb{N}$ such that for all $n_j \geq N$, $j = 1, \ldots, q$, it holds $P(|X_{(n_1, \ldots, n_q)} - Y_{(n_1, \ldots, n_q)}| > \varepsilon) < \delta$. Limits such as $|E|X_{\boldsymbol{n}}| - E|Y_{\boldsymbol{n}}|| \to 0$, as $\boldsymbol{n} \to \infty$, are defined analogously.

3.2. Convergence of moments

Let us assume that we are given a random field of stochastic processes,

$$X_{\boldsymbol{n}}(\lambda), \quad \lambda \in \Lambda, \boldsymbol{n} \in \mathbb{N}^q,$$

which can be approximated by another random field

$$Y_{\boldsymbol{n}}(\lambda), \quad \lambda \in \Lambda, \boldsymbol{n} \in \mathbb{N}^q.$$

The following theorem shows that the moments of $Y_{\boldsymbol{n}}(\lambda)$ are uniformly close to the moments of $X_{\boldsymbol{n}}(\lambda)$ under weak assumptions, which only concern the (joint) marginal distribution and avoid to assume that $\sup_{\lambda \in \Lambda} |X_{\boldsymbol{n}}|$ and $\sup_{\lambda \in \Lambda} |Y_{\boldsymbol{n}}|$ are uniformly integrable.

Theorem 3.1. *Fix* $0 < r < \infty$. *Let* $\{X_n(\lambda) : \boldsymbol{n} \geq \mathbf{1}, \lambda \in \Lambda\}$ *and* $\{Y_{\boldsymbol{n}}(\lambda) : \boldsymbol{n} \geq \mathbf{1}, \lambda \in \Lambda\}$ *be parameterized random fields satisfying the strict marginal stationarity condition*

$$(X_{\boldsymbol{n}}(\lambda), Y_{\boldsymbol{n}}(\lambda)) \overset{d}{=} (X_{\boldsymbol{n}}(\lambda'), Y_{\boldsymbol{n}}(\lambda')), \tag{3.1}$$

for all $\lambda, \lambda' \in \Lambda$ *and* $\boldsymbol{n} \geq \mathbf{1}$. *Suppose that* $\{|X_{\boldsymbol{n}}(\lambda)|^r : \boldsymbol{n} \geq \mathbf{1}\}$ *and* $\{|Y_{\boldsymbol{n}}(\lambda)|^r : \boldsymbol{n} \geq \mathbf{1}\}$ *are uniformly integrable*, $\lambda \in \Lambda$, *with*

$$\sup_{\lambda \in \Lambda} |X_{\boldsymbol{n}}(\lambda) - Y_{\boldsymbol{n}}(\lambda)| \xrightarrow{P} 0,$$

as $\boldsymbol{n} \to \infty$. *Then the following assertions hold*:

(i) $\sup_{\lambda \in \Lambda} E|X_{\boldsymbol{n}} - Y_{\boldsymbol{n}}|^r \to 0$, *as* $\boldsymbol{n} \to \infty$, *for* $0 < r \leq 1$.

(ii) $\sup_{\lambda \in \Lambda} |E|X_{\boldsymbol{n}}|^r - E|Y_{\boldsymbol{n}}|^r| \to 0$, *as* $\boldsymbol{n} \to \infty$, *for* $0 < r \leq 1$.

(iii) $\sup_{\lambda \in \Lambda} |E(X_{\boldsymbol{n}}) - E(Y_{\boldsymbol{n}})| \to 0$, *as* $\boldsymbol{n} \to \infty$.

(iv) $\sup_{\lambda \in \Lambda} |\|X_{\boldsymbol{n}}\|_r - \|Y_{\boldsymbol{n}}\|_r| \to 0$, *as* $\boldsymbol{n} \to \infty$, *for* $0 < r < \infty$.

The condition (3.1) is automatically satisfied for a large class of cases: Suppose that $\Lambda = \mathbb{Z}$ and for some sequence of i.i.d. random elements $\{\xi_\lambda : \lambda \in \mathbb{Z}\}$ taking values in some measurable space $(E, \mathcal{E})$ we have

$$X_{\boldsymbol{n}}(\lambda) = \Psi_{\boldsymbol{n}}(\xi_\lambda, \xi_{\lambda-1}, \cdots),$$

and

$$Y_{\boldsymbol{n}}(\lambda) = \Phi_{\boldsymbol{n}}(\xi_\lambda, \xi_{\lambda-1}, \cdots),$$

for all $\lambda \in \mathbb{Z}$ and $\boldsymbol{n} \geq \mathbf{1}$. Then, for arbitrary $\lambda, \lambda' \in \Lambda$ and all $\boldsymbol{n}$,

$$\begin{aligned}(X_{\boldsymbol{n}}(\lambda), Y_{\boldsymbol{n}}(\lambda)) &= (\Psi_{\boldsymbol{n}}(\xi_\lambda, \xi_{\lambda-1}, \cdots), \Phi_{\boldsymbol{n}}(\xi_\lambda, \xi_{\lambda-1}, \cdots)) \\ &\overset{d}{=} (\Psi_{\boldsymbol{n}}(\xi_{\lambda'}, \xi_{\lambda'-1}, \cdots), \Phi_{\boldsymbol{n}}(\xi_{\lambda'}, \xi_{\lambda'-1}, \cdots)) \\ &= (X_{\boldsymbol{n}}(\lambda'), Y_{\boldsymbol{n}}(\lambda')),\end{aligned}$$

which verifies (3.1). Observe that the ξ_λ's may be random variables, random vectors or general random elements such as random functions taking values in an infinite-dimensional space.

Let us now consider parameterized models where

$$X_{\boldsymbol{n}}(\lambda) = X_{\boldsymbol{n}}(\lambda; \vartheta),$$

for some parameter vector $\vartheta \in \Theta$. It is assumed that Θ is a subset of a normed space equipped with a norm $\|\cdot\|$. Partition $\vartheta = (\eta', \zeta')'$ and assume that the surrogate model is obtained by estimating, say, ζ, such that

$$Y_{\boldsymbol{n}}(\lambda) = X_{\boldsymbol{n}}(\lambda; (\eta', \widehat{\zeta}'_{\boldsymbol{n}})'), \tag{3.2}$$

where $\widehat{\zeta}_{\boldsymbol{n}}$ is a statistical estimator of ζ, obtained by some statistical method of estimation from a random sample (also called calibration to the sample), satisfying

$$\|\widehat{\zeta}_{\boldsymbol{n}} - \zeta\| \xrightarrow{P} 0,$$

as $\boldsymbol{n} \to \infty$. If the mapping $X_{\boldsymbol{n}}(\lambda; \vartheta)$ is Lipschitz continuous in η with a uniform Lipschitz constant L such that

$$\sup_{\lambda \in \Lambda} | X_{\boldsymbol{n}}(\lambda; (\eta', \zeta_1')') - X_{\boldsymbol{n}}(\lambda; (\eta', \zeta_2')')| \leq L \| \zeta_1 - \zeta_2 \|,$$

for all η, ζ_1, ζ_2 and all $\boldsymbol{n} \in \mathbb{N}^q$, then using (3.2)

$$\sup_{\lambda \in \Lambda} | X_{\boldsymbol{n}}(\lambda; (\eta', \zeta')') - Y_{\boldsymbol{n}}(\lambda)| \leq L \| \widehat{\zeta}_{\boldsymbol{n}} - \zeta \| \xrightarrow{P} 0$$

as $\boldsymbol{n} \to \infty$, follows.

Putting things together and noting that the L above can be random without affecting the convergence, we obtain the following result.

Theorem 3.2. *Assume that $X_{\boldsymbol{n}}(\lambda)$ and $Y_{\boldsymbol{n}}(\lambda)$ are parametrized by some parameter $\vartheta = (\eta', \zeta')' \in \Theta$, for some set Θ, and are of the form*

$$X_{\boldsymbol{n}}(\lambda) = \Psi_{\boldsymbol{n}}(\xi_\lambda, \xi_{\lambda-1}, \cdots; (\eta', \zeta')'),$$

and

$$Y_{\boldsymbol{n}}(\lambda) = \Psi_{\boldsymbol{n}}(\xi_\lambda, \xi_{\lambda-1}, \cdots; (\eta', \widehat{\zeta}'_{\boldsymbol{n}})'),$$

for all $\lambda \in \mathbb{Z}$ and $\boldsymbol{n} \in \mathbb{N}^q$, for some sequene $(\xi_\lambda : \lambda \in \Lambda\}$, where $\Lambda \subset \mathbb{Z}$ and ξ_λ are i.i.d. and attain values in some measurable space $(E, \mathcal{E})$. Further suppose that the mapping $\Psi_{\boldsymbol{n}}$ is Lipschitz continuous in ζ, such that for some random variable L

$$\sup_{\lambda \in \Lambda} | X_{\boldsymbol{n}}(\lambda; (\eta', \zeta')') - Y_{\boldsymbol{n}}(\lambda)| \leq L \| \widehat{\zeta}_{\boldsymbol{n}} - \zeta \|,$$

$\boldsymbol{n} \in \mathbb{N}^q$. If $\widehat{\zeta}_{\boldsymbol{n}}$ is a consistent estimator of ζ, then the assumptions of Theorem 3.1 *are satisfied.*

3.3. Proof. We give the details of the proof of Theorem 3.1.

Proof of Theorem 3.1. By the c_r-inequality

$$|X_{\boldsymbol{n}}(\lambda) - Y_{\boldsymbol{n}}(\lambda)|^r \leq 2^r(|X_{\boldsymbol{n}}(\lambda)|^r + |Y_{\boldsymbol{n}}(\lambda)|^r),$$

we may conclude that $\{|X_{\boldsymbol{n}}(\lambda) - Y_{\boldsymbol{n}}(\lambda)|^r : \boldsymbol{n} \geq 1\}$ is uniformly integrable. Let $\varepsilon > 0$. We have

$$E|X_{\boldsymbol{n}}(\lambda) - Y_{\boldsymbol{n}}(\lambda)|^r \leq \varepsilon^r + E[|X_{\boldsymbol{n}}(\lambda) - Y_{\boldsymbol{n}}(\lambda)|^r 1(|X_{\boldsymbol{n}}(\lambda) - Y_{\boldsymbol{n}}(\lambda)| > \varepsilon)].$$

Fix $\lambda_0 \in \Lambda$. By uniform integrability, there exists $\eta = \eta(\lambda_0) > 0$ such that for all events A with $P(A) < \eta$ we have $E[|X_{\boldsymbol{n}}(\lambda_0) - Y_{\boldsymbol{n}}(\lambda_0)|^r 1_A] < \varepsilon.$

Since $\sup_{\lambda \in \Lambda}|X_{\boldsymbol{n}}(\lambda) - Y_{\boldsymbol{n}}(\lambda)| \xrightarrow{P} 0,$ as $\boldsymbol{n} \to \infty$, there exists $\boldsymbol{n}_0$ such that for all $\boldsymbol{n} \geq \boldsymbol{n}_0$

$$P(|X_{\boldsymbol{n}}(\lambda_0) - Y_{\boldsymbol{n}}(\lambda_0)| > \varepsilon) \leq P\left(\sup_{\lambda \in \Lambda}|X_{\boldsymbol{n}}(\lambda) - Y_{\boldsymbol{n}}(\lambda)| > \varepsilon\right) < \eta.$$

It follows that

$$\begin{aligned}
&\sup_{\lambda \in \Lambda} E[|X_{\boldsymbol{n}}(\lambda) - Y_{\boldsymbol{n}}(\lambda)|^r 1(|X_{\boldsymbol{n}}(\lambda) - Y_{\boldsymbol{n}}(\lambda)| > \varepsilon)] \\
&\quad = E[|X_{\boldsymbol{n}}(\lambda_0) - Y_{\boldsymbol{n}}(\lambda_0)|^r 1(|X_{\boldsymbol{n}}(\lambda_0) - Y_{\boldsymbol{n}}(\lambda_0)| > \varepsilon)] \\
&\quad < \varepsilon,
\end{aligned}$$

where the equality is a consequence of (3.1), leading to

$$\sup_{\lambda \in \Lambda} E|X_{\boldsymbol{n}}(\lambda) - Y_{\boldsymbol{n}}(\lambda)|^r \leq \varepsilon^r + \varepsilon, \quad \boldsymbol{n} \geq \boldsymbol{n}_0,$$

which shows (i), since ε is arbitrary. To show (ii), apply the inequality $|x + y|^r \leq |x|^r + |y|^r$ to obtain $||x|^r - |y|^r| \leq |x - y|^r,$ such that

$$|E|X_{\boldsymbol{n}}(\lambda)|^r - E|Y_{\boldsymbol{n}}(\lambda)|^r| \le E|X_{\boldsymbol{n}}(\lambda) - Y_{\boldsymbol{n}}(\lambda)|^r \le \sup_{\lambda \in \Lambda} E|X_{\boldsymbol{n}}(\lambda) - Y_{\boldsymbol{n}}(\lambda)|^r,$$

which yields which establishes (ii). (iii) follows by linearity, $\sup_{\lambda \in \Lambda} |E(X_{\boldsymbol{n}}) - E(Y_{\boldsymbol{n}})| \le \sup_{\lambda \in \Lambda} E|X_{\boldsymbol{n}} - Y_{\boldsymbol{n}}|$, and (iv) is shown as in the proof of Theorem 2.1. □

4. Applications to Surrogate Models: Deep Learning and Gaussian Processes

As a surrogate model is used to generate cheap artificial data sets (e.g., to enrich real observed data), classes of models with convincing approximation properties are preferable. Deep learning neural networks as well as Gaussian processes are two widespread frameworks for surrogate modelling, as they satisfy this requirement. Deep learners are a popular approach for nonlinear relationships going beyond classical statistical regression models. Gaussian processes have gained interest to model curves as random trajectories. Typically, one calibrates such a model to a (relatively small) data set of real data and then simulates from the fitted model to obtain simulated data samples which should be close to real samples. Applications are widespread and diverse: Signal processing and image analysis as briefly discussed above, classification and pattern matching, analysis of data such as asset returns and option prices from financial markets [7], as well as applications to quality control and reliability analysis [3].

Deep learning networks

A deep learning artificial neural network, see, e.g., [4], is a mapping $f : \mathcal{X} \to \mathbb{R}^q$, which maps an input vector x of the input space $\mathcal{X} \subset \mathbb{R}^p$, assumed to be a compact set, to a q-dimensional output vector $\boldsymbol{y}$, $p, q \in \mathbb{N}$, and is given by the composition of H layers in the form

$$\boldsymbol{y} = f(\boldsymbol{x}) = f_H(\cdots f_2(\boldsymbol{W}_2 f_1(\boldsymbol{W}_1 x + \boldsymbol{b}_1) + \boldsymbol{b}_2)\cdots), \quad \boldsymbol{x} \in \mathcal{X},$$

where $\boldsymbol{W}_l$ are weighting matrices, $\boldsymbol{b}_l$ intercept terms, and f_l are activation functions, $l = 1, \dots, H$. The input can be a regressor of explanatory variables for a regression problem, log returns or prices of assets, options and futures contracts as in finance, a vectorized (sequence) of digitized images or audio signals as in classification problems, or even a discretized trajectory of a stochastic process. The parameter vector of the net is $\vartheta = (\mathrm{vec}\boldsymbol{W}_1, \dots, \mathrm{vec}\boldsymbol{W}_H, b_1', \dots, b_H')'$, where vec$\boldsymbol{A}$ denotes the vectorized version of a matrix $\boldsymbol{A}$ obtained by stacking columns, such that

$$f(x) = f(x; \vartheta).$$

The activation functions are typically nonlinear and always chosen as Lipschitz continuous functions. Clearly, the sum of two Lipschitz functions with Lipschitz constants L_1 and L_2 is again Lipschitz with Lipschitz constant $L_1 + L_1$, and the composition $f \circ g$ of two Lipschitz functions f and g with constants L_1 and L_2 is again Lipschitz with Lipschitz constant $L_1 L_2$, because $|f(g(x)) - f(g(y))| \le L_1 |g(x) - g(y)| \le L_1 L_2 |x - y|$. Therefore, such deep learning networks are Lipschitz continuous in the parameters. Indeed, current efforts focus on calculating the Lipschitz constants. It is not restrictive to assume that $E|f(\boldsymbol{X}; \vartheta)|^{1+\delta} < \infty$, for some $\delta > 0$, where $\boldsymbol{X}$ is a random input. Alternatively, assume that $E|\boldsymbol{X}|^{1+\delta} < \infty$ holds and the existence of some $x_0 \in \boldsymbol{\mathcal{X}}$ such that $f(x_0; \vartheta) = 0$. Then

$$\begin{aligned} E|f(\boldsymbol{X}; \vartheta)|^{1+\delta} &= E|f(\boldsymbol{X}; \vartheta) - f(x_0; \vartheta)|^{1+\delta} \\ &\le LE|\boldsymbol{X} - x_0|^{1+\delta} \\ &\le L(\|\boldsymbol{X}\|_{1+\delta} + \|x_0\|_{1+\delta})^{1+\delta} \\ &< \infty, \end{aligned}$$

where L denotes the Lipschitz constant of the net.

If a deep learning network is trained at time n, say from a data stream, using the most recent n data points $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ with associated outputs $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$, by estimating the parameters ϑ, the trained network is given by

$$f(x; \widehat{\vartheta}_n),$$

where

$$\widehat{\vartheta}_n = \widehat{\vartheta}_n(\xi_1, \dots, \xi_n),$$

with $\xi_i = (\boldsymbol{X}_i', \boldsymbol{Y}_i')'$, $i = 1, \dots, n$.

The next step is to simulate a random input $\boldsymbol{X} \sim G$, say for some G with $\int |x|^{1+\delta} dG(x) < \infty$. Then the associated output,

$$Y_n = f(\boldsymbol{X}; \widehat{\vartheta}_n(\xi_1, \dots, \xi_n)),$$

is a surrogate for $X_n = f(\boldsymbol{X}; \vartheta)$. Obviously, Theorem 3.2 applies and yields the convergence of moments. The simulation of inputs $\boldsymbol{X} \sim G$ is usually based on random number generators. For algorithms and background, we refer to [2].

Gaussian processes kriging

Let us consider the following example studied in some depth in [3] dealing with reliability analysis. Let X denote a d-dimensional random vector with density f_X and support D. Given a performance function $g : D \to \mathbb{R}$ a *failure*, e.g., of a system, can be modeled by the event $\{g(X) \leq 0\}$. Conducting such experiments in practice is, however, sometimes expensive, whereas simulations from an appropriate (surrogate) model are usually cheap. Since g is unknown, a surrogate model for g is used, which allows to estimate (or predict) g and quantify the involved uncertainty. The Gaussian process kriging approach assumes that g is a sample path of an underlying Gaussian process $\mathcal{G}$,

$$\mathcal{G}(x) = f(x)^{\prime}\beta + Z(x), \quad x \in D.$$

Here $f(x)^{\prime}\beta$ is the linear predictor with respect to given functions $f_1(x), \dots, f_p(x)$ from a basis of, say, the function space L_2, and a parameter vector $\beta \in \mathbb{R}^p$, and $Z(x)$ is a mean zero stationary Gaussian process with a stationary correlation function, often chosen in practice as

$$R(x - x', \ell_1, \dots, \ell_d) = \exp\left(-\sum_{k=1}^{d}[(x_k - x_k')/\ell_l]^2\right),$$

for scaling parameters $\ell_1, \dots, \ell_p > 0$.

Given a random sample $X_1, \dots, X_n$ of size n, the best linear unbiased (kriging) estimator of $\mathcal{G}(x)$ at x is Gaussian and interpolates the observations $g(X_i) = f(X_i)^{\prime}\beta$, if $g \in \text{span}\{f_1, \dots, f_p\}$, i.e., there is no residual uncertainty (at the observed data points). By increasing p as n gets larger, any L_2-function can be estimated in this way. Note that the predictor depends on n. An observation $Y_n = Y_n(x)$ simulated from the surrogate model for some $x \notin \{X_1, \dots, X_n\}$ is regarded as an approximation of an unobserved $X_n(x)$ (obtained in a Gedanken experiment which is too expensive to be carried out).

References

[1] K. L. Chung, A Course in Probability Theory, 3rd Edition, Academic Press, San Diego, CA, 2001.

[2] L. Devroye, Non-Uniform Random Variate Generation, Springer, New York, 1986.

[3] V. Dubourg, B. Sudret and F. Deheeger, Metamodel-based importance sampling for structural reliability analysis, Probabilistic Engineering Mechanics 33 (2013), 47-57.

DOI: https://doi.org/10.1016/j.probengmech.2013.02.002

[4] I. Goodfellow, Y. Bengio and A. Courville, Deep Learning, MIT Press, 2016.

http://www.deeplearningbook.org

[5] L. Leskelä and M. Vihola, Stochastic order characterizations of uniform integrability and tightness, Statistics and Probability Letters 83(1) (2013), 382-389.

DOI: https://doi.org/10.1016/j.spl.2012.09.023

[6] A. Müller and D. Stoyan, Comparison Methods for Stochastic Models and Risks, Wiley, New York, 2002.

[7] A. Steland, Financial Statistics and Mathematical Finance: Methods, Models and Applications, Wiley, Chichester, 2012.

■