Journal of Multivariate Analysis 102 (2011) 789–800

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

A numerical method for minimum distance estimation problems

C. Cervellera*, D. Macciò

Istituto di Studi sui Sistemi Intelligenti per l'Automazione, Consiglio Nazionale delle Ricerche, Via de Marini 6, 16149 Genova, Italy

ARTICLE INFO

Article history:
Received 19 July 2010
Available online 5 January 2011

AMS subject classifications:
62F10
65D15

Keywords:
Minimum distance estimation
Point estimation
Sampling
Functional optimization
Approximation

ABSTRACT

This paper introduces a general method for the numerical derivation of a minimum distance (MD) estimator for the parameters of an unknown distribution. The approach is based on an active sampling of the space in which the random sample takes values and on the optimization of the parameters of a suitable approximating model. This allows us to derive the MD estimator function for any given distribution, by which we can immediately obtain the MD estimate of the unknown parameters in correspondence to any observed random sample. Convergence of the method is proved when mild conditions on the sampling process and on the involved functions are satisfied, and it is shown that favorable rates can be obtained when suitable deterministic sequences are employed. Finally, simulation results are provided to show the effectiveness of the proposed algorithm on two case studies.

1. Introduction

A fundamental topic in the wide field of inferential statistics is the analysis of a sample of i.i.d. realizations of a random variable, with the aim of gaining information about the probability distribution according to which it has been generated. This problem, which is generally referred to as statistical inference and arises in many different fields such as engineering, physics, biology, and health care, can assume different forms related to the specific context. Several techniques specifically tailored to these instances have been developed in the literature. Maximum likelihood estimation [6,10], minimum chi-square [11], the method of moments [12], and the Kolmogorov–Smirnov test [11] are all examples of popular inference methods.

The *minimum distance* method [16] is a very general technique that formalizes the inference problem as the search for a distribution function that is as close as possible to the empirical distribution given by the observed data.

Formally, the minimum distance (MD) problem can be stated as follows: consider a sample $\boldsymbol{z}$ of n i.i.d. realizations $\boldsymbol{z} = (x_1, \dots, x_n) \in X^n \subseteq \mathbb{R}^n$ of a real random variable $x \in X \subseteq \mathbb{R}$, drawn from a cumulative distribution function (CDF) which is known to belong to

$$\mathscr{F} = \{F(\cdot, \boldsymbol{\theta}) | \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k\},$$

where $\boldsymbol{\theta}$ is a set of unknown parameters that we want to estimate from the random sample.

Denote by $d[\cdot, \cdot]$ any proper distance function defined on $\mathscr{F} \times \mathscr{F}$ and denote by $F_n(\cdot, \boldsymbol{z})$ the empirical distribution based on the sample $\boldsymbol{z} = (x_1, \dots, x_n)$, defined as

$$F_n(t, \boldsymbol{z}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[x_i, \infty)}(t),$$

where $\mathbb{1}_A$ is the characteristic function of A.

* Corresponding author.
E-mail addresses: cervellera@ge.issia.cnr.it (C. Cervellera), ddmach@ge.issia.cnr.it (D. Macciò).

0047-259X/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2010.12.006

Furthermore, define $\gamma(n)$ as a positive function that tends to 0 as $n \to \infty$.

The following definition is due to [16].

Definition 1.1. If there exists a $\hat{\boldsymbol{\theta}}$ in Θ such that

$$d[F(\cdot, \hat{\boldsymbol{\theta}}), F_n(\cdot, \boldsymbol{z})] < \inf_{\boldsymbol{\theta} \in \Theta} d[F(\cdot, \boldsymbol{\theta}), F_n(\cdot, \boldsymbol{z})] + \gamma(n), \tag{1}$$

then $\hat{\boldsymbol{\theta}}$ is called the *minimum distance estimate* of $\boldsymbol{\theta}$, given $\boldsymbol{z}$.

A function $\hat{\boldsymbol{\theta}}$ that solves (1) for any $\boldsymbol{z} \in X^n$ is a *minimum distance estimator* for the family of distributions $\mathscr{F}$.

As regards the distance function, in this paper we consider the usual $\mathscr{L}^p$-norm. Then, we have

$$d[F(\cdot, \boldsymbol{\theta}), F_n(\cdot, \boldsymbol{z})] = \left(\int_X \left(F(t, \boldsymbol{\theta}) - F_n(t, \boldsymbol{z}) \right)^p \mathrm{d}t \right)^{\frac{1}{p}}. \tag{2}$$

The main issue with this method is that we are not able, in general, to derive an expression for the MD estimator as a function of $\boldsymbol{z}$. In fact, this would require the solution of a functional optimization problem (i.e., a problem where the solution is a function of the random sample) that, in a general case, cannot be solved analytically. Thus, in general, we need to use a numerical optimization technique to find the minimum in correspondence to each observed random sample $\boldsymbol{z}$, which can be computationally demanding, if not impossible, in a real-time context.

This is probably the main reason why, in th literature, MD estimation has been addressed mostly from a theoretical point of view and as a methodological basis for other techniques, while, to the best of the authors' knowledge, no actual methods for deriving MD estimators in general cases have been proposed. However, a method to obtain MD estimators is worth investigation, not only due to the good properties of such estimators like, for example, consistency [16], but also due to the fact that all the inference approaches mentioned above can be seen as special cases of the minimum distance estimation framework (see, e.g., [1] for a discussion).

In this paper, we introduce a new method based on active sampling, capable of yielding an MD estimate of the parameters of the unknown distribution in correspondence to any random sample of fixed size drawn according to the distribution given by any actual value of the parameters. In other words, the output of the method is a function of $\boldsymbol{z}$ that provides an MD estimate in any point of X^n, i.e., the MD estimator. For this reason, the proposed method will be called global approximate minimum distance (GAMD) estimation.

The GAMD solution is obtained by selecting the best element within a suitable class of approximating functions through an empirical risk minimization principle, and a uniform sampling of the space X^n where the random sample takes values. It is proved that the obtained estimator converges to the true MD estimator that minimizes (1), provided that the sampling of the random sample space X^n is sufficiently uniform (according to a notion of *discrepancy*, as will be detailed in the following) and the involved functions satisfy some mild regularity assumptions.

Notice that we are dealing with two different concepts of a sample. One is the n-dimensional random sample $\boldsymbol{z}$, which is the data we observe online and is generated by an external source according to the distribution F. The other is a uniform sampling by which we discretize the space X^n where such a random sample takes values, i.e., a set of L points $\{\boldsymbol{z}_1, \dots, \boldsymbol{z}_L\}$ where each $\boldsymbol{z}_l \in X^n$ for $l = 1, \dots, L$, and it is chosen offline in order to derive the estimator, according to the procedure that will be described in the following. This means that we can build the MD estimator for F without actually having to generate samples according to F, which may be problematic in some cases. All we need is a sample of L points of X^n that are well uniformly distributed. In particular, it will be proved that, by employing sets coming from i.i.d. sequences with uniform distribution or *low-discrepancy* sequences, $\mathscr{L}^1$ convergence can be achieved with an almost quadratic or linear rate, respectively, as the number of observations grows. This makes the proposed approach particularly suited to problems with a high-dimensional input space.

The method turns out to be simple and computationally manageable. In fact, the computational effort is reduced to a pointwise minimization in the space of the parameters that characterize the chosen approximating architectures. Furthermore, once the GAMD estimator has been obtained, the output (i.e., the MD estimate) can be evaluated instantaneously for any given observed random sample, without the need for performing any minimization in real time.

The paper is organized as follows. In Section 2, the basic algorithm for the solution of MD problem (1) is described. Section 3 contains an analysis of the convergence properties of the proposed method, whereas Section 4 is devoted to experimental results regarding the application of the proposed approach to two case studies. Section 5 draws some conclusions. Finally, the Appendices contain definitions and proofs.

2. The global approximate minimum distance estimation approach

In this section, we introduce the proposed numerical procedure for the approximate solution of the minimum distance estimation problem. In particular, we show how the extension of the integral defining the distance to the sample space allows us to derive a solution, through the minimization of an empirical estimate based on a sample of observations.

First of all, we consider the minimum distance problem through the minimization of the distance of the true distribution from a *smoothed empirical distribution*. In particular, we employ the definition of smoothed empirical distribution introduced

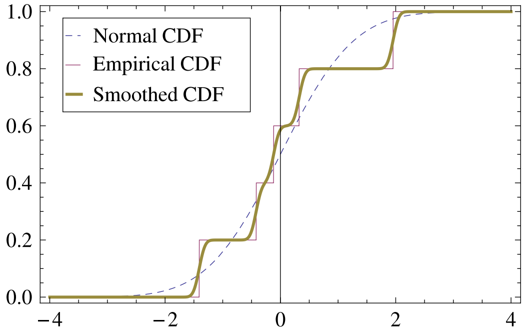


Fig. 1. CDFs of a normal random variable.

in [5], which is given for a sample $\boldsymbol{z} = (x_1, \dots, x_n)$ by

$$\tilde{F}_n(t, \boldsymbol{z}) = \frac{1}{n} \sum_{i=1}^{n} K_n(t - x_i),$$

where K_n is a sequence of CDFs converging to the unit-step function as $n \to \infty$. A common way to obtain such a K_n is by integrating a kernel function:

$$K_n(t) = \int_{-\infty}^{t} a_n^{-1} k(\tau/a_n) \mathrm{d}\tau, \tag{3}$$

where $k(t) \geq 0$, $\int_{-\infty}^{\infty} k(t)\,\mathrm{d}t = 1$ and $a_n \geq 0$ is monotonically decreasing to 0.

As an example, Fig. 1 depicts the CDF of a normal distributed random variable with zero mean and variance equal to 1, together with its empirical and smoothed empirical versions.

With this choice, we can define

$$J(\boldsymbol{z}, \boldsymbol{\theta}) = d[F(\cdot, \boldsymbol{\theta}), \tilde{F}_n(\cdot, \boldsymbol{z})], \tag{4}$$

and take the function $\boldsymbol{\theta}^\circ$ as the one that minimizes J for each $\boldsymbol{z} \in X^n$:

$$\boldsymbol{\theta}^\circ(\boldsymbol{z}) = \arg\min_{\boldsymbol{\theta} \in \Theta} J(\boldsymbol{z}, \boldsymbol{\theta}),$$

which in general may not be unique.

Notice that in defining $\boldsymbol{\theta}^\circ$ we have assumed the existence of the minimum of J over Θ; if this is not the case, we can consider a $\boldsymbol{\theta}^\circ \in \Theta$ such that $J(\boldsymbol{z}, \boldsymbol{\theta}^\circ) < \inf_{\boldsymbol{\theta} \in \Theta} J(\boldsymbol{z}, \boldsymbol{\theta}) + \delta$ for some arbitrarily small $\delta > 0$, without changing the validity of the following analysis.

Proposition 2.1. *The function $\boldsymbol{\theta}^\circ$ that minimizes the distance when the smoothed distribution is employed is an MD estimator according to Definition* 1.1.

The proof can be found in Appendix B.

Then, if we consider the $\mathscr{L}^p$-norm as the distance d, we are now looking for a function $\boldsymbol{\theta}^\circ$ of the random sample $\boldsymbol{z}$ which minimizes the functional

$$J(\boldsymbol{z}, \boldsymbol{\theta}) = \int_X \left(F(t, \boldsymbol{\theta}) - \tilde{F}_n(t, \boldsymbol{z})\right)^p \mathrm{d}t. \tag{5}$$

Notice that, with respect to (2), we have dropped the annoying exponent $1/p$, since it does not affect the point of minimum.

For the purpose of the present analysis, we assume that the set X^n in which the random sample $\boldsymbol{z}$ takes values is a compact subset of $\mathbb{R}^n$, for every $\boldsymbol{\theta}$ in Θ. This is not a serious practical limitation since, if X^n is not compact, due to Ulam's theorem [4, p. 225] we know that we can find a compact X_c^n such that the probability measure of the set $X^n \setminus X_c^n$ is smaller than any fixed positive value ϵ.

Let us consider the set of MD estimators for F, defined as

$$W_{\min} = \{\boldsymbol{\theta}^\circ | \boldsymbol{\theta}^\circ(\boldsymbol{z}) = \arg\min_{\boldsymbol{\theta} \in \Theta} J(\boldsymbol{z}, \theta) \quad \text{for all } \boldsymbol{z} \in X^n\}.$$

Next, define the extension over X^n of the integral defining J as

$$\Phi(\theta) = \int_{X^n} J(\boldsymbol{z}, \boldsymbol{\theta}(\boldsymbol{z}))\,\mathrm{d}\boldsymbol{z},$$

and denote by $\boldsymbol{\theta}^*$ the minimizer of $\Phi(\boldsymbol{\theta})$ over the set W, where W is such that $W_{\min} \subset W$:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in W} \Phi(\boldsymbol{\theta}).$$

Then, $\boldsymbol{\theta}^*$ is defined as the function belonging to W that minimizes the extension of the integral J over X^n. Again, $\boldsymbol{\theta}^*$ is possibly not unique; let W^* be the collection of possible $\boldsymbol{\theta}^*$:

$$W^* = \{\boldsymbol{\theta}^* \in W | \boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in W} \Phi(\boldsymbol{\theta})\}.$$

Notice that we have $W_{\min} \subset W^*$.

In the following, we show that the MD estimation problem can be solved by considering the set W^* instead of $W_{\min}$. In particular, we can prove the following lemma.

Lemma 2.1. *For every $\boldsymbol{\theta}^* \in W^*$ there exists $\boldsymbol{\theta}^\circ \in W_{\min}$ such that $J(\boldsymbol{z}, \boldsymbol{\theta}^\circ(\boldsymbol{z})) = J(\boldsymbol{z}, \boldsymbol{\theta}^*(\boldsymbol{z}))$ almost everywhere.*

This result can be extended to obtain an actual equivalence between W^* and $W_{\min}$, when the function J satisfies some further regularity properties.

Lemma 2.2. *Assume that J is such that*

$$W = \left\{\boldsymbol{\theta} |\, \textit{for each } \boldsymbol{z} \in X^n \textit{ there exists } \Omega(\boldsymbol{z}) \subset X^n \textit{ with } \lambda(\Omega(\boldsymbol{z})) \neq 0 \textit{ such that } \boldsymbol{\theta} \textit{ is continuous on } \Omega(\boldsymbol{z})\right\},$$

where λ denotes the Lebesgue measure. Then, we have $W^* = W_{\min}$.

Proofs of Lemmas 2.1 and 2.2 can be found in Appendix B.

Summing up, Lemma 2.1 (and 2.2, when the function J is sufficiently well behaved) asserts that we can focus our attention on the minimizer $\boldsymbol{\theta}^*$ of the integral J extended over the sample space X^n. This allows us to derive a numerical procedure based on sampling to obtain a solution, as will be detailed in the following.

To this purpose, we first need to define a proper class of functions Γ where we look for $\boldsymbol{\theta}^*$ or, in general, an ϵ-close approximation of $\boldsymbol{\theta}^*$. In particular, we employ parameterized functions for the class Γ, so that the minimization of $\Phi(\boldsymbol{\theta})$ reduces to an optimization procedure in a finite-dimensional space of parameters. Then, Γ has the form $\{\boldsymbol{\psi}(\boldsymbol{z}, \boldsymbol{\alpha}) | \boldsymbol{\alpha} \in \Lambda \subseteq \mathbb{R}^q\}$, where $\boldsymbol{\alpha}$ is a vector of parameters that have to be optimized.

The procedure for obtaining the MD estimator is based now on the approximation of $\Phi(\boldsymbol{\theta})$ by means of an empirical version of the integral, obtained through a finite sample of points in $X \times X^n$.

In particular, let us consider again the expression of $\Phi(\theta)$. Then, we can write

$$\Phi(\boldsymbol{\theta}) = \int_{X^n} J(z, \boldsymbol{\theta}(\boldsymbol{z}))\,\mathrm{d}\boldsymbol{z} \tag{6a}$$

$$= \int_{X^n} \left(\int_X \left(F(t, \boldsymbol{\theta}(\boldsymbol{z})) - \tilde{F}_n(t, \boldsymbol{z})\right)^p \mathrm{d}t \right) \mathrm{d}\boldsymbol{z} \tag{6b}$$

$$= \int_{X^{n+1}} \left(F(t, \boldsymbol{\theta}(\boldsymbol{z})) - \tilde{F}_n(t, \boldsymbol{z}) \right)^p \mathrm{d}(t, \boldsymbol{z}), \tag{6c}$$

where the equality between (6b) and (6c) is ensured by the Fubini–Tonelli theorem [4, p. 137]. This version of Φ, written as an integral over the extended space X^{n+1}, can now be evaluated through an empirical approximation.

In particular, choose $\boldsymbol{w}^L = \{(t_1, \boldsymbol{z}_1), \ldots, (t_L, \boldsymbol{z}_L)\}$ as a sample of discretization points of X^{n+1}, and an approximating function $\boldsymbol{\psi}(\boldsymbol{z}, \boldsymbol{\alpha}) \in \Gamma$. Then, the empirical version of $\Phi(\boldsymbol{\theta})$ is given by

$$\Phi_{\mathrm{emp}}(\boldsymbol{w}^L, \boldsymbol{\alpha}) = \frac{1}{L} \sum_{l=1}^{L} \left(F(t_l, \boldsymbol{\psi}(\boldsymbol{z}_l, \boldsymbol{\alpha})) - \tilde{F}_n(t_l, \boldsymbol{z}_l)\right)^p. \tag{7}$$

The problem of obtaining an approximate MD estimator is thus reduced to the following: find $\boldsymbol{\alpha}_L^*$ such that

$$\boldsymbol{\alpha}_L^* = \arg\min_{\boldsymbol{\alpha}\in\Lambda} \Phi_{\mathrm{emp}}(\boldsymbol{w}^L, \boldsymbol{\alpha}).$$

Once the value of $\boldsymbol{\alpha}_L^*$ has been determined, the MD estimator provided by GAMD, given $\boldsymbol{z}$, is then $\boldsymbol{\psi}(\boldsymbol{z}, \boldsymbol{\alpha}_L^*)$.

Notice that, as said, due to the extension of the integration over X^n, the solution turns out to be "globally good" for the whole considered space in which the random sample takes values.

Summing up, the proposed GAMD relies on two key elements: (i) the selection of a suitable class Γ of parameterized approximating functions and (ii) the choice of a sampling scheme for the space X^{n+1}. In the following section, we provide conditions on the two aforementioned elements in order to guarantee convergence of the obtained estimator to the true MD estimator.

3. Convergence results

In this section, we show that the approach based on the minimization of the empirical risk $\Phi_{\text{emp}}(\boldsymbol{w}^L, \boldsymbol{\alpha})$ is asymptotically consistent, i.e., it leads to the true estimator $\boldsymbol{\theta}^\circ$ for $L \to \infty$. In particular, conditions on the class of models Γ and the random sampling process for the generation of $\boldsymbol{w}^L$ are provided.

For the sake of the convergence analysis, in the following we assume that $X^n = [0, 1]^n$, i.e., the n-dimensional unitary cube, without loss of generality. In fact, the proposed approach can be extended to problems where $X^n = \prod_{i=1}^n [a, b]$ by simple scaling.

Define, for notational convenience, the function

$$U(t, \boldsymbol{z}, \boldsymbol{\theta}) = \left(F(t, \boldsymbol{\theta}) - \tilde{F}_n(t, \boldsymbol{z})\right)^p.$$

We can prove the consistency of the method if the function U is sufficiently well behaved; for instance, if it satisfies a Lipschitz condition.

Assumption 3.1. The function $U(t, \boldsymbol{z}, \boldsymbol{\theta})$ is Lipschitz with respect to $\boldsymbol{\theta}$ for every $(t, \boldsymbol{z}) \in X^{n+1}$, i.e., there exists a finite C such that

$$|U(t, \boldsymbol{z}, \boldsymbol{\theta}_1) - U(t, \boldsymbol{z}, \boldsymbol{\theta}_2)| \le C\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$$

for every $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$.

Notice that the fulfillment of Assumption 3.1 is strictly related to the behavior of $F(t, \boldsymbol{\theta})$. In particular, most of the distributions used in practice are differentiable with respect to $\boldsymbol{\theta}$; then Assumption 3.1 is naturally satisfied.

Recall from the previous section that the MD estimation problem has been reduced to finding a good approximation over the whole space X^n of the function $\boldsymbol{\theta}^\circ$ that minimizes $J(\boldsymbol{z}, \boldsymbol{\theta}) = \int_X U(t, \boldsymbol{z}, \boldsymbol{\theta})\mathrm{d}t$ for every $\boldsymbol{z}$. This has been obtained through the minimization of the empirical cost defined in (7), yielding the approximate estimator $\boldsymbol{\psi}(\cdot, \boldsymbol{\alpha}_L^*)$.

Thus, in the rest of the paper we measure the quality of performance of the GAMD in terms of an $\mathscr{L}^1$ error, defined by

$$e_1(L) = \int_{X^n} \left(J(\boldsymbol{z}, \boldsymbol{\psi}(\boldsymbol{z}, \boldsymbol{\alpha}_L^*)) - J(\boldsymbol{z}, \boldsymbol{\theta}^\circ(\boldsymbol{z}))\right) \mathrm{d}\boldsymbol{z}.$$

The expression of $e_1(L)$ measures the distance between the functional J evaluated in the optimal $\boldsymbol{\theta}^\circ$ and the one given by employing $\boldsymbol{\psi}(\cdot, \boldsymbol{\alpha}_L^*)$, i.e., it measures how well the true MD estimator $\boldsymbol{\theta}^\circ$ is approximated.

By considering Lemma 2.1 and Eqs. (6a)–(6c), we can eventually write the error in this form, which is the one we use for convergence analysis in the following:

$$e_1(L) = \int_{X^{n+1}} \left(U(t, \boldsymbol{z}, \boldsymbol{\psi}(\boldsymbol{z}, \boldsymbol{\alpha}_L^*)) - U(t, \boldsymbol{z}, \boldsymbol{\theta}^*(\boldsymbol{z}))\right)\mathrm{d}(t, \boldsymbol{z}).$$

Consider the quantity $\xi(\boldsymbol{z}) = \min_{\boldsymbol{\alpha} \in \Lambda} \|\boldsymbol{\theta}^*(\boldsymbol{z}) - \boldsymbol{\psi}(\boldsymbol{z}, \boldsymbol{\alpha})\|$ and $\boldsymbol{\alpha}^*$ as the argument that attains the minimum, i.e., $\boldsymbol{\psi}(\cdot, \boldsymbol{\alpha}^*)$ is the element in Γ that is closest to $\boldsymbol{\theta}^*$. Notice that $\boldsymbol{\alpha}^*$ can be equivalently defined as the minimum of $\Phi(\boldsymbol{\psi}(\boldsymbol{z}, \boldsymbol{\alpha}))$.

From Assumption 3.1 and the definition of $\xi(\boldsymbol{z})$, the next inequality follows.

$$\begin{aligned} e_1(L) &= \int_{X^{n+1}} \left(U(t, \boldsymbol{z}, \boldsymbol{\psi}(z, \alpha_L^*)) - U(t, \boldsymbol{z}, \boldsymbol{\psi}(\boldsymbol{z}, \boldsymbol{\alpha}^*))\right)\mathrm{d}(t, z) + \int_{X^{n+1}} \left(U(t, \boldsymbol{z}, \boldsymbol{\psi}(\boldsymbol{z}, \boldsymbol{\alpha}^*)) - U(t, \boldsymbol{z}, \boldsymbol{\theta}^*(\boldsymbol{z}))\right)\mathrm{d}(t, \boldsymbol{z}) \\ &\le e_1^L(L) + C\int_{X^n} \xi(\boldsymbol{z})\mathrm{d}\boldsymbol{z}. \end{aligned} \tag{8}$$

Inequality (8) states that the error $e_1(L)$ can be seen as the sum of two different contributions. The term containing $\xi(\boldsymbol{z})$ depends only on the approximating capabilities of the class of models Γ. For this reason, we refer to this term as the *approximation error*. Conversely, $e_1^L(L)$ is related to how close we can get to the best element within Γ by minimizing Φ_{emp}, i.e., by the element of Γ corresponding to $\boldsymbol{\alpha}_L^*$. We denote this quantity as the *estimation error*.

Concerning the approximation error, we introduce the following assumption.

Assumption 3.2. The class of models Γ is endowed with a *universal approximation property*, i.e., for any $\varepsilon > 0$, we have

$$\min_{\boldsymbol{\alpha} \in \Lambda} \|\boldsymbol{\theta}^*(z) - \boldsymbol{\psi}(\boldsymbol{z}, \boldsymbol{\alpha})\| < \varepsilon.$$

This property is usually obtained by considering classes of growing complexity, so that any function can be approximated with arbitrary accuracy by increasing the parameter that rules the richness of the elements of the class, such as, for example, the number of radial basis functions in radial basis function networks, the number of neural units in feedforward neural networks, etc.

As is known, many commonly employed classes of functions, both linear and nonlinear, fulfill Assumption 3.2. To name a few: sigmoidal neural networks, expansions of radial basis functions, splines, and orthogonal polynomials (such as Legendre or Chebyshev). In Section 4, containing experimental results, we test the use of one-hidden-layer neural networks and kernel approximators.

As regards the estimation error, this quantity measures the performance of the approximation $\boldsymbol{\psi}(\cdot, \boldsymbol{\alpha}_L^*)$ obtained through the GAMD when used in place of $\boldsymbol{\psi}(\cdot, \boldsymbol{\alpha}^*)$. This term depends on the algorithm employed to generate the sample of discretization points $\boldsymbol{w}^L = \{(t_1, \boldsymbol{z}_1), \ldots, (t_L, \boldsymbol{z}_L)\}$ in X^{n+1}, used to build the empirical distance Φ_{emp}. In the following, we show how convergence of the error can be strictly related to (i) the regularity of the functions involved and (ii) the uniformity of the sampling of the space X^{n+1}.

To this purpose, define $V_{\text{HK}}^U(\boldsymbol{\alpha})$ as the variation in the sense of Hardy and Krause over X^{n+1} of $U(\cdot, \cdot, \boldsymbol{\psi}(\cdot, \boldsymbol{\alpha}))$, and $\mathscr{D}^*(\boldsymbol{w}^L)$ as the star discrepancy of the sample of points $\boldsymbol{w}^L$. Definitions of variation in the sense of Hardy and Krause and of star discrepancy, measures of the regularity of a function and of the uniformity of a set of points, respectively, commonly employed in number-theoretic and quasi-Monte Carlo methods [13], can be found in Appendix A.

We are now ready to prove the convergence of the GAMD. First, we make the following assumptions.

Assumption 3.3.

(i) The functions U and $\boldsymbol{\psi}$ are such that

$$\sup_{\boldsymbol{\alpha}\in\Lambda} V_{\text{HK}}^U(\boldsymbol{\alpha}) < \infty. \tag{9}$$

(ii) The sequence of points $\boldsymbol{w}^L$ satisfies

$$\lim_{L\to\infty} \mathscr{D}^*(\boldsymbol{w}^L). \tag{10}$$

Then, we can prove the following theorem.

Theorem 3.1. *If Assumptions 3.2 and 3.3 hold, then*

$$\lim_{L\to\infty} e_1(L) = 0.$$

Furthermore, the estimation error $e_1^L(L)$ has the same rate of convergence as the star discrepancy $\mathscr{D}^*(\boldsymbol{w}^L)$ in (10).

The proof of the theorem can be found in Appendix A.

Concerning point (i) in Assumption 3.3, it can be proved, for instance, that the composition of functions with bounded partial derivatives has bounded variation in the sense of Hardy and Krause. Consequently, a sufficient condition for $V_{\text{HK}}^U(\boldsymbol{\alpha})$ to be finite is that $\boldsymbol{\psi}$ and U have bounded partial derivatives. Notice that this is easily verified for the most common approximating architectures.

As to the discrepancy of the sample $\boldsymbol{w}^L$, since the rate of convergence of $\mathscr{D}^*(\boldsymbol{w}^L)$ controls that of $e_1^L(L)$, we need to choose sampling schemes that present favorable discrepancy rates. To this purpose, a good choice is the use of the low-discrepancy sequences [13, Ch. 3], a family of deterministic sequences originally developed in the numerical integration framework. In this case they gave rise to the so-called quasi-Monte Carlo methods, a set of techniques introduced to outperform the performances of the well-known Monte Carlo methods. In the context of the present paper, we propose the use of a particular type of low-discrepancy sequences, namely (t, n)-*sequences* [13, Ch. 4], that attain an almost linear rate of convergence for the discrepancy, i.e.,

$$\mathscr{D}^*(\boldsymbol{w}^L) \le \mathscr{O}\left(\frac{\log(L)^{d-1}}{L}\right).$$

With such a choice, the rate of convergence of the estimation error is given by $e_1^L(L) \simeq \mathscr{O}(1/L)$.

Notice that i.i.d. sequences with uniform distribution, typical of classic Monte Carlo methods, can also be employed. In this case, it can be proved that condition (10) is still satisfied, but the rate of convergence of the discrepancy is now quadratic [3], i.e., $e_1^L(L) \simeq \mathscr{O}(1/\sqrt{L})$. In spite of this slower theoretical asymptotic rate, i.i.d. sequences have the advantage of being simpler to obtain with respect to (t, n)-sequences, and they are already available for most software platforms (even if implementations of low-discrepancy sequences are spreading quickly in mathematical software packages).

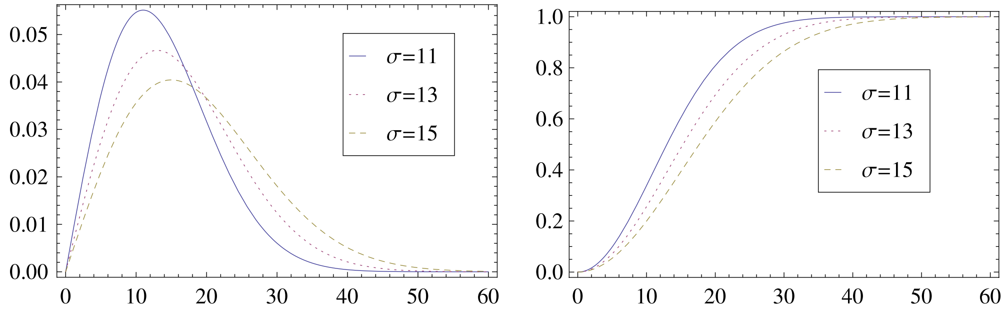
4. Simulation results

In this section, we address two case studies to show how the true MD estimator $\boldsymbol{\theta}^\circ$ can be approximated well by the solution obtained through the GAMD method described in the previous sections.

In particular, we consider random samples coming from (i) a Rayleigh distribution and (ii) a mixture of Gaussians.

Concerning the smoothed empirical distribution, in these examples we obtain the CDF K_n by integrating a kernel density estimator:

$$K_n(t) = \int_{-\infty}^{t} a_n^{-1} k(\tau/a_n)\mathrm{d}\tau, \tag{11}$$



(a) PDF. (b) CDF.

Fig. 2. Rayleigh distribution.

Table 1
Mean and standard deviation of the AEs.

ν, L	σ				
	11	12	13	14	15
10, 1000	0.22, 0.17	0.20, 0.17	0.18, 0.16	0.15, 0.14	0.15, 0.11
10, 2000	0.20, 0.16	0.19, 0.14	0.18, 0.14	0.15, 0.13	0.13, 0.11
20, 3000	0.11, 0.11	0.11, 0.09	0.08, 0.09	0.08, 0.08	0.07, 0.07
20, 4000	0.11, 0.11	0.09, 0.09	0.07, 0.09	0.07, 0.08	0.08, 0.07

where $k(t) \geq 0$, $\int_{-\infty}^{\infty} k(t)\,dt = 1$ and $a_n \geq 0$ is such that $a_n \to 0$ as $n \to \infty$, monotonically. Here we take $k(t) = \exp(-\pi t^2)$ and $a_n = 1/n$.

As far as the functional J in (5) is concerned, in these simulations we use the $\mathscr{L}^2$-norm, i.e., $p = 2$.

4.1. Rayleigh distribution

The Rayleigh distribution is a popular probability distribution, describing the distance from the origin of a point (X, Y) when X and Y are independent and normally distributed with equal variance. It is often employed in engineering applications to model radial errors in a plane. More specifically, the CDF of a Rayleigh distribution with parameter $\sigma > 0$ is given by

$$F(t, \sigma) = 1 - \exp(-t^2/2\sigma^2),$$

for $t > 0$.

Fig. 2 shows the Rayleigh distribution for different values of σ. In the left part of the figure, the probability density function (PDF) is shown, while the right part depicts the cumulative distribution function.

The goal of the test is to approximate the MD estimator, $\hat{\sigma}_{\mathrm{MD}}(\boldsymbol{z})$, of σ for 10-dimensional samples $\boldsymbol{z} \in X^{10}$ in the range [0, 50], by employing the GAMD procedure using a one-hidden-layer neural network with ν sigmoidal neural units as the class Γ.

We recall that a one-hidden-layer neural network is a map of the form

$$\psi(\boldsymbol{z}, \boldsymbol{\alpha}) = \sum_{i=1}^{\nu} c_i h\left(\sum_{j=1}^{n} a_{ij} x_j + b_i\right) + c_0,$$

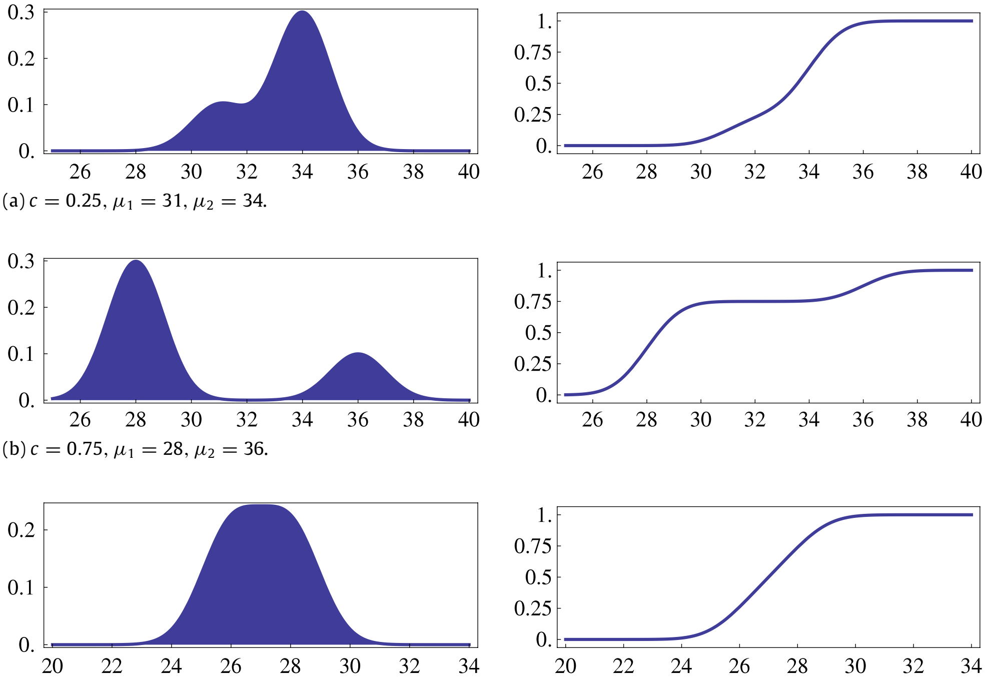
where

$$h(\zeta) = \frac{\exp(\zeta) - \exp(-\zeta)}{\exp(\zeta) + \exp(-\zeta)}$$

is the chosen activation function, $c_0, c_i, b_i \in \mathbb{R}$, and $\boldsymbol{a}_i := [a_{i1}, \ldots, a_{id}]^{\mathrm{T}} \in \mathbb{R}^n$ are the weights. In this case, $\boldsymbol{\alpha} = [a_1^{\mathrm{T}}, \ldots, a_\nu^{\mathrm{T}}, b_1, \ldots, b_\nu, c_0, \ldots, c_\nu]^{\mathrm{T}}$.

The optimal weights of the network have been determined by minimizing the empirical risk Φ_{emp} on a sufficiently rich set $\boldsymbol{w}^L$ of sampling points in the 11-dimensional hypercube $X^{n+1} = [0, 50]^{11}$. In particular, we tested four discretizations of $[0, 50]^{11}$ based on a low discrepancy sequence (specifically, a Sobol' sequence [15]), having size $L = 1000, 2000, 3000$ and 4000, respectively. For the cases $L = 1000$ and 2000, we employed neural networks with $\nu = 10$ while, for $L = 3000$ and 4000, $\nu = 20$ was chosen. Increasing values of L have been employed in order to show the improvements expected from the theory given by having more sampling points at our disposal.

In order to test the performance of the GAMD, for a given value σ and a given $\boldsymbol{z}$ we used as reference the true MD estimate $\boldsymbol{\theta}^\circ(\boldsymbol{z})$ obtained by minimizing $J(\boldsymbol{z}, \sigma)$ as in (5) through a nonlinear programming routine. Table 1 reports the results, evaluated by generating 100 10-dimensional Rayleigh distributed samples for different values of σ, and by computing the



(a) $c = 0.25$, $\mu_1 = 31$, $\mu_2 = 34$.

(b) $c = 0.75$, $\mu_1 = 28$, $\mu_2 = 36$.

(c) $c = 0.5$, $\mu_1 = 26$, $\mu_2 = 28$.

Fig. 3. PDFs (left panel) and CDFs (right panel) of three mixtures of Gaussians.

absolute error (AE) between the pointwise reference solution defined above and the one provided by the obtained network. In the table, both the mean and the standard deviation of the AE over the 100 points are reported.

Looking at the table, it can be seen that the performance of the approximate MD estimator turns out to be very satisfactory. It is also worth noting that the performance of the neural estimator generally improves as the number of discretization points increases, in accordance with the theory presented in the paper.

4.2. Mixtures of Gaussians

Mixtures of distributions are defined as convex combinations of different probability distributions. When the sum is made of Gaussian densities, the model is referred to as a mixture of Gaussians. More formally, a mixture of distributions takes the form

$$F(t, \boldsymbol{\theta}) = \sum_{i=1}^{M} c_i P_i(t, \boldsymbol{\gamma}_i),$$

where $\sum_{i=1}^{M} c_i = 1$ and every P_i is a CDF. The parameters to be estimated are the real coefficients c_i and possibly the unknown parameters $\boldsymbol{\gamma}_i$ of the functions P_i.

The importance of the mixtures of Gaussians is related to the possibility of approximating any suitably regular distribution function by properly increasing M [14], making these types of model useful even in the context of nonparametric estimation.

In the present section, we consider the mixture

$$F(t, \boldsymbol{\theta}) = c\mathcal{N}_{\mu_1,\sigma_1}(t) + (1 - c)\mathcal{N}_{\mu_2,\sigma_2}(t) \tag{12}$$

in the domain $X = [20, 40]$, where $c \in [0, 1]$, and the functions $\mathcal{N}$ are normal distributions with means μ_1 and μ_2 and variances σ_1 and σ_2, respectively. We assume that the variances are known and both equal to 1, so that the MD problem consists in estimating $\boldsymbol{\theta} = [c, \mu_1, \mu_2]$. The sample size has been taken equal to $n = 10$. Fig. 3 shows three examples of mixtures of Gaussians of the form of (12) with different values of the triple $[c, \mu_1, \mu_2]$.

In this case, we tested the GAMD approach using local kernel approximator schemes [8] for the class of parameterized functions Γ. This kind of model assumes that the function g to be estimated is known in a finite set of points $\Sigma_K = \{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K\}$ and defines the approximating function $\hat{g}$ in a point $\boldsymbol{u}$ not belonging to Σ_K as the average of the values $g(\boldsymbol{\xi}_k)$, $k = 1, \ldots, K$, weighted according to a measure of distance of $\boldsymbol{u}$ from each $\boldsymbol{\xi}_k$. In particular, the weight is given by a decreasing function, called the *kernel*, that depends only on the distance from the center points $\boldsymbol{\xi}_k$ and is parameterized by a scalar variable $r > 0$, which defines its range of influence. Notice that the kernels used to get the class of models Γ must not be confused with the ones employed to generate the smoothed version of the empirical distribution in (3).

Table 2
Mean and standard deviation of the AEs.

(a) $c = 0.25$, $\mu_1 = 31$, $\mu_2 = 34$			
K, L	θ		
	c	μ_1	μ_2
500, 1000	0.022, 0.13	0.95, 0.88	0.82, 0.45
1000, 2000	0.017, 0.14	0.85, 0.99	0.82, 0.47
1500, 3000	0.015, 0.11	0.78, 0.45	0.71, 0.45
2000, 4000	0.014, 0.13	0.65, 0.59	0.70, 0.42
(b) $c = 0.75$, $\mu_1 = 28$, $\mu_2 = 36$			
K, L	θ		
	c	μ_1	μ_2
500, 1000	0.020, 0.12	1.01, 1.77	0.72, 1.14
1000, 2000	0.017, 0.11	0.95, 1.82	0.69, 1.07
1500, 3000	0.014, 0.11	0.83, 1.76	0.65, 1.04
2000, 4000	0.015, 0.12	0.75, 1.78	0.65, 1.25
(c) $c = 0.5$, $\mu_1 = 26$, $\mu_2 = 28$			
K, L	θ		
	c	μ_1	μ_2
500, 1000	0.033, 0.16	1.05, 0.76	0.74, 0.55
1000, 2000	0.027, 0.17	0.94, 0.70	0.70, 0.57
1500, 3000	0.028, 0.16	0.86, 0.68	0.69, 0.61
2000, 4000	0.022, 0.19	0.71, 0.71	0.63, 0.57

More formally, we are considering structures of the form

$$\hat{g}(\boldsymbol{u}) = \frac{\sum_{k=1}^{K} \mathcal{K}_r(\boldsymbol{u}, \boldsymbol{\xi}_k) g(\boldsymbol{\xi}_k)}{\sum_{k=1}^{K} \mathcal{K}_r(\boldsymbol{u}, \boldsymbol{\xi}_k)},$$

where $\mathcal{K}_r(\boldsymbol{u}, \boldsymbol{\xi}_k)$ is the instance of the kernel, typically defined by $\mathcal{G}(\|\boldsymbol{u} - \boldsymbol{\xi}_k\|/r)$, where $\mathcal{G}(s)$ is a non-increasing function for $s > 0$ having a maximum at $s = 0$. A typical example of $\mathcal{G}(s)$ is the Gaussian function $\exp(-\pi s^2)$.

As in the previous case, we tested four samples $\boldsymbol{w}^L$ of points in $X^{n+1} = [20, 40]^{11}$ by employing a Sobol' sequence, with size $L = 1000, 2000, 3000$ and 4000, respectively, for the purpose of evaluating the performances of the approximation given by the GAMD as L increases.

According to the size of L, the size of the sets Σ_K used to define the class Γ has been taken equal to $K = 500, 1000, 1500$ and 2000, respectively, with the points again coming from a Sobol' sequence.

The MD estimates given by the GAMD approach have been compared with the MD estimates computed, for a given $\boldsymbol{z}$ and given values of c, μ_1, μ_2, again solving (5). In particular, the AE has been computed in 100 10-dimensional test points generated according to mixtures of Gaussians with three different combinations of c, μ_1, μ_2, depicted in Fig. 3. In this case, given that the minimization problem is three-dimensional, we have implemented the *sequential algorithm for optimization with NT-nets* (SNTO) routine to find the point of the minimum. SNTO is an efficient global minimization routine based on number-theoretic methods, introduced in [7], where it was also successfully applied to the pointwise solution of a maximum-likelihood estimation problem. Here again we employed Sobol' sequences to obtain the SNTO solution for each of the 100 combinations of c, μ_1, μ_2, taking such a solution as the reference MD estimate. Notice that the value of the MD estimates to be used as coefficients of the kernel basis functions of Γ (i.e., the points where the function to be approximated is known) has also been obtained by solving (5) pointwise through SNTO.

Table 2 shows the mean and the standard deviation of the AE over the 100 points in the three cases, confirming again the excellent performances of the GAMD in approximating the MD estimator.

As a last remark, we point out the importance of having obtained, through the GAMD, the MD estimator as a function of the random sample. In fact, evaluating the SNTO solution for all the 100 test samples requires about 20 min of computation (on a 1.8 GHz Intel Core2 CPU with 1 Gb of RAM), whereas obtaining the 100 GAMD estimates, due to the fact that they are simply the output of the estimator, is almost instantaneous. This makes the GAMD approach particularly suited to situations in which time constraints are restrictive.

5. Conclusions and future work

In this paper, a numerical method for the solution of the minimum distance estimation problem has been proposed. The method is able to provide an approximate MD estimate in correspondence to any given random sample generated according to a given distribution function with unknown parameters, i.e., it provides the MD estimator function. The approach, called global approximate minimum distance, relies on the discretization of the space in which the random sample takes values by means of uniformly scattered sets of points, and it is based on the minimization of an empirical version of the distance through the optimization of the parameters of a suitable approximating function. The convergence analysis has proved that the use of particular sequences of points commonly employed in quasi-Monte Carlo integration, namely, low-discrepancy sequences, can lead an to almost linear asymptotic rate of convergence of the estimator given by the GAMD method to the best approximation of the true MD estimator within the chosen class of approximating models. Results on application of the proposed approach to two case studies, namely a Rayleigh distribution and a mixture of Gaussian distributions, indicate the method as promising and computationally efficient.

Concerning future work, more detailed research on the mathematical framework and the analysis of other case studies on particular minimum distance instances will be the subject of further investigations.

Appendix A. Variation and discrepancy

For each vertex of a given subinterval $B = \prod_{i=1}^{d}[a_i, b_i]$ of I^d, it is possible to define a binary label by assigning '0' to every a_i and '1' to every b_i. For every function $f: I^d \to \mathbb{R}$, we define $\Delta(f, B)$ as the alternating sum of f computed at the vertices of B, i.e.,

$$\Delta(f, B) = \sum_{\boldsymbol{y}\in e_B} f(\boldsymbol{y}) - \sum_{\boldsymbol{y}\in o_B} f(\boldsymbol{y}),$$

where e_B is the set of vertices with an even number of '1's in their label, and o_B is the set of vertices with an odd number of '1's.

Definition A.1. The *variation in the sense of Vitali* of a real-valued function f on I^d is defined by

$$V^{(d)}(f) = \sup_{P\in\mathcal{P}} |\Delta(f, P)|,$$

where $\mathcal{P}$ is any partition of I^d into subintervals.

If the partial derivatives of f are continuous on I^d, it is possible to write $V^{(d)}(f)$ in an easier way:

$$V^{(d)}(f) = \int_0^1 \cdots \int_0^1 \left| \frac{\partial^d f}{\partial y_1 \cdots \partial y_d} \right| \mathrm{d}y_1 \cdots \mathrm{d}y_d,$$

where y_i is the i-th component of $\boldsymbol{y}$.

For $1 \le k \le d$ and $1 \le i_1 < i_2 < \cdots < i_k \le d$, let $V^{(k)}(f; i_1 \ldots i_k)$ be the variation in the sense of Vitali of the restriction of f to the k-dimensional face $\{(y_1, \ldots, y_d) \in I^d | y_i = 1 \text{ for } i \neq i_1, \ldots, i_k\}$.

Definition A.2. Let f be a real-valued function defined on I^d; then

$$V_{\mathrm{HK}}(f) = \sum_{k=1}^{d} \sum_{1\le i_1<i_2<\cdots<i_k\le d} V^{(k)}(f; i_1 \ldots i_k)$$

is called the *variation in the sense of Hardy and Krause*, and f is of bounded variation in this sense if $V_{\mathrm{HK}}(f)$ is finite.

Consider a sample $\boldsymbol{v}^L$ consisting of L points $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_L\}$ in the d-dimensional unit cube I^d. For an arbitrary subset $B \in I^d$, let us define by $\mathcal{C}(B, v^L)$ the number of points of $\boldsymbol{v}^L$ that belong to B, i.e.,

$$\mathcal{C}(B, \boldsymbol{v}^L) = \sum_{l=1}^{L} \mathbb{1}_B(\boldsymbol{v}_l).$$

Definition A.3. If $\mathcal{J}^*$ is the family of all the closed subintervals of I^d that can be written as $\prod_{i=1}^{d}[0, b_i]$, the *star discrepancy* $\mathcal{D}^*(\boldsymbol{v}^L)$ is defined as

$$\mathcal{D}^*(\boldsymbol{v}^L) = \sup_{B\in\mathcal{J}^*} \left| \frac{\mathcal{C}(B, \boldsymbol{v}^L)}{L} - \lambda(B) \right|,$$

where $\lambda(B)$ is the Lebesgue measure of B.

Loosely speaking, the star discrepancy is a quantitative measure of the spread of points in a region of interest; the more uniformly distributed the sequence of points in the space is, the smaller the star discrepancy is.

Appendix B. Proofs

Proof of Proposition 2.1. Consider a $\boldsymbol{\theta}_{\min}(\boldsymbol{z})$ such that

$$d[F(\cdot, \boldsymbol{\theta}_{\min}(\boldsymbol{z})), F_n(\cdot, \boldsymbol{z})] < \inf_{\boldsymbol{\theta}\in\Theta} d[F(\cdot, \boldsymbol{\theta}), F_n(\cdot, \boldsymbol{z})] + \epsilon(n),$$

where $\epsilon(n) \to 0$ as $n \to 0$.

Now, we can write

$$\begin{aligned} d[F(\cdot, \boldsymbol{\theta}^\circ(\boldsymbol{z})), F_n(\cdot, \boldsymbol{z})] &\le d[F(\cdot, \boldsymbol{\theta}^\circ(\boldsymbol{z})), \tilde{F}_n(\cdot, \boldsymbol{z})] + d[\tilde{F}_n(\cdot, \boldsymbol{z}), F_n(\cdot, \boldsymbol{z})] \\ &\le d[F(\cdot, \boldsymbol{\theta}_{\min}(\boldsymbol{z})), \tilde{F}_n(\cdot, \boldsymbol{z})] + d[\tilde{F}_n(\cdot, \boldsymbol{z}), F_n(\cdot, \boldsymbol{z})] \\ &\le d[F(\cdot, \boldsymbol{\theta}_{\min}(\boldsymbol{z})), F_n(\cdot, \boldsymbol{z})] + 2d[\tilde{F}_n(\cdot, \boldsymbol{z}), F_n(\cdot, \boldsymbol{z})] \\ &< \inf_{\boldsymbol{\theta}\in\Theta} d[F(\cdot, \boldsymbol{\theta}), F_n(\cdot, \boldsymbol{z})] + 2d[\tilde{F}_n(\cdot, \boldsymbol{z}), F_n(\cdot, \boldsymbol{z})] + \epsilon(n). \end{aligned}$$

For any $\boldsymbol{z}$, we have that the term $d[F_n(\cdot, \boldsymbol{z}), \tilde{F}_n(\cdot, \boldsymbol{z})]$ tends to 0 as $n \to \infty$ by construction. Thus the proposition follows, taking into account the fact that also $\epsilon(n) \to 0$ as $n \to 0$. □

Proof of Lemma 2.1. Suppose that there exists $\theta^* \in W^*$ such that, for every $\boldsymbol{\theta}^\circ \in U_{\min}$, a neighborhood $\Omega(\boldsymbol{\theta}^\circ) \in X^n$ with $\lambda(\Omega(\boldsymbol{\theta}^\circ)) > 0$ (λ denoting Lebesgue measure) can be found, where $J(\boldsymbol{z}, \boldsymbol{\theta}^\circ(\boldsymbol{z})) < J(\boldsymbol{z}, \boldsymbol{\theta}^*(\boldsymbol{z}))$ for every $\boldsymbol{z} \in X^n$. Then, we have $\Phi(\boldsymbol{\theta}^\circ) < \Phi(\boldsymbol{\theta}^*)$ for every $\boldsymbol{\theta}^\circ \in W_{\min}$, which contradicts the definition of $\boldsymbol{\theta}^*$. □

Proof of Lemma 2.2. $\Phi(\boldsymbol{\theta}^\circ)$ is obviously minimum for every $\boldsymbol{\theta}^\circ \in W_{\min}$, which implies that $W_{\min} \subset W^*$. Now, suppose that $\Phi(\boldsymbol{\theta}^*)$ is minimum, but $\boldsymbol{\theta}^* \notin W_{\min}$. Then, for every $\boldsymbol{\theta}^\circ \in W_{\min}$, there exists $\tilde{\boldsymbol{z}} \in X^n$ such that $\boldsymbol{\theta}^\circ(\tilde{\boldsymbol{z}}) \neq \boldsymbol{\theta}^*(\tilde{\boldsymbol{z}})$ and $J(\tilde{\boldsymbol{z}}, \boldsymbol{\theta}^\circ(\tilde{\boldsymbol{z}})) < J(\tilde{\boldsymbol{z}}, \boldsymbol{\theta}^*(\tilde{\boldsymbol{z}}))$. The assumption of the theorem implies that there is $\Omega(\boldsymbol{z}, \boldsymbol{\theta}^\circ)$ with $\lambda(\Omega(\boldsymbol{z}, \boldsymbol{\theta}^\circ)) \neq 0$ such that $J(\tilde{\boldsymbol{z}}, \boldsymbol{\theta}^*(\tilde{\boldsymbol{z}})) > J(\tilde{\boldsymbol{z}}, \boldsymbol{\theta}^\circ(\tilde{\boldsymbol{z}}))$ for every $\boldsymbol{\theta}^\circ \in \Omega(\boldsymbol{z}, \boldsymbol{\theta}^\circ)$. Since $\boldsymbol{\theta} \in U$, this implies that $\Phi(\boldsymbol{\theta}^*) > \Phi(\boldsymbol{\theta}^\circ)$, which contradicts the hypothesis that $\Phi(\boldsymbol{\theta}^*)$ is minimum. Thus $W^* \subset W_{\min}$. □

Proof of Theorem 3.1. From (8), we recall that

$$e_1(L) \le e_1^L(L) + C\int_{X^n} \xi(\boldsymbol{z})\mathrm{d}\boldsymbol{z},$$

where the first term is the estimation error and the second one is the approximation error.

The latter can be trivially annihilated by considering Assumption 3.2 and the definition of $\xi(\boldsymbol{z}) = \min_{\boldsymbol{\alpha}\in\Lambda} \|\boldsymbol{\theta}^*(\boldsymbol{z}) - \boldsymbol{\psi}(\boldsymbol{z}, \boldsymbol{\alpha})\|$.

Then all we need is to prove that the estimation error $e_1^L(L)$ converges to zero as $L \to \infty$.

First, recall that $\boldsymbol{\alpha}^*$ and $\boldsymbol{\alpha}_L^*$ are defined as

$$\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}\in\Lambda} \Phi(\boldsymbol{\psi}(\boldsymbol{z}, \boldsymbol{\alpha})), \qquad \boldsymbol{\alpha}_L^* = \arg\min_{\boldsymbol{\alpha}\in\Lambda} \Phi_{\text{emp}}(\boldsymbol{w}^L, \boldsymbol{\alpha}).$$

Following the proof of Theorem 4 in [2], based on the *Koksma–Hlawka inequality* [9], we can prove that

$$\lim_{L\to\infty} \sup_{\boldsymbol{\alpha}} \left| \frac{1}{L}\sum_{l=1}^{L} U(t_l, \boldsymbol{z}_l, \boldsymbol{\psi}(\boldsymbol{z}_l, \boldsymbol{\alpha})) - \int_{X^{n+1}} U(t, \boldsymbol{z}, \boldsymbol{\psi}(\boldsymbol{z}, \boldsymbol{\alpha}))\mathrm{d}(t, \boldsymbol{z}) \right| = 0$$

with the same rate of convergence as $\mathscr{D}^*(\boldsymbol{w}^L)$ in (10).

Therefore, for any $\epsilon > 0$, we can choose $\bar{L} = \bar{L}(\epsilon)$ such that, for every $L \ge \bar{L}$,

$$\int_{X^{n+1}} U(t, \boldsymbol{z}, \boldsymbol{\psi}(\boldsymbol{z}, \boldsymbol{\alpha}_L^*))\mathrm{d}(t, \boldsymbol{z}) \le \frac{1}{L}\sum_{l=1}^{L} U(t_l, \boldsymbol{z}_l, \boldsymbol{\psi}(z_l, \boldsymbol{\alpha}_L^*)) + \frac{\epsilon}{2} \tag{13}$$

and

$$\frac{1}{L}\sum_{l=1}^{L} U(t_l, \boldsymbol{z}_l, \boldsymbol{\psi}(z_l, \boldsymbol{\alpha}^*)) \le \int_{X^{n+1}} U(t, \boldsymbol{z}, \boldsymbol{\psi}(\boldsymbol{z}, \boldsymbol{\alpha}^*))\mathrm{d}(t, \boldsymbol{z}) + \frac{\epsilon}{2}. \tag{14}$$

Thus, by combining (13) and (14) with the fact that, by definition of $\boldsymbol{\alpha}_L^*$,

$$\sum_{l=1}^{L} U(t_l, \boldsymbol{z}_l, \boldsymbol{\psi}(\boldsymbol{z}_l, \boldsymbol{\alpha}_L^*)) \le \sum_{l=1}^{L} U(t_l, \boldsymbol{z}_l, \boldsymbol{\psi}(\boldsymbol{z}_l, \boldsymbol{\alpha}^*)),$$

we have

$$e_1^L(L) = \int_{X^{n+1}} U(t, \boldsymbol{z}, \boldsymbol{\psi}(\boldsymbol{z}, \boldsymbol{\alpha}_L^*))\mathrm{d}(t, \boldsymbol{z}) - \int_{X^{n+1}} U(t, \boldsymbol{z}, \boldsymbol{\psi}(\boldsymbol{z}, \boldsymbol{\alpha}^*))\mathrm{d}(t, \boldsymbol{z}) < \epsilon. \quad \square$$

References

[1] C.R. Blyth, On the inference and decision models of statistics, Ann. Math. Statist. 41 (3) (1970) 1034–1058.

[2] C. Cervellera, M. Muselli, Deterministic design for neural network learning: an approach based on discrepancy, IEEE Trans. Neural Netw. 15 (2004) 533–543.

[3] K.L. Chung, An estimate concerning the Kolmogoroff limit distribution, Trans. Amer. Math. Soc. 67 (1949) 36–50.

[4] R.M. Dudley, Real Analysis and Probability, Cambridge University Press, 2002.

[5] M. Falk, Relative efficiency and deficiency of kernel type estimators of smooth distribution functions, Statist. Neerlandica 37 (1983) 73–83.

[6] J. Fan, Q. Yao, Nonlinear Time Series: Nonparametric and Parametric Methods, Springer-Verlag, Berlin, 2005.

[7] K.-T. Fang, Y. Wang, Number-Theoretic Methods in Statistics, Chapman & Hall, London, 1994.

[8] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer-Verlag, 2001.

[9] E. Hlawka, Funktionen von Beschränkter variation in der theorie der Gleichverteilung, Ann. Mat. Pura Appl. 54 (1961) 325–333.

[10] E. Kuhn, M. Lavielle, Maximum likelihood estimation in nonlinear mixed effects models, Comput. Statist. Data Anal. 49 (4) (2005) 1020–1038.

[11] E.L. Lehmann, J.P. Romano, Testing Statistical Hypotheses, 3rd ed., Springer, New York, 2005.

[12] A.M. Mood, F.A. Graybill, D.C. Boes, Introduction to the Theory of Statistics, McGraw-Hill Companies, 1974.

[13] H. Niederreiter, Random Number Generation and Quasi-Monte Carlo Methods, SIAM, Philadelphia, 1992.

[14] T. Poggio, F. Girosi, Networks for approximation and learning, Proc. IEEE 78 (9) (1990) 1481–1497.

[15] I.M. Sobol', The distribution of points in a cube and the approximate evaluation of integrals, Zh. Vychisl. Mat. Mat. Fiz. 7 (1967) 784–802.

[16] J. Wolfowitz, The minimum distance method, Ann. Math. Statist. 28 (1957) 75–87.