


Article

Asymptotic Convergence of Soft-Constrained Neural Networks for Density Estimation

Edmondo Trentin 

Dipartimento di Ingegneria dell'Informazione e Scienze Matematiche, Università di Siena, 53100 Siena, Italy; trentin@dii.unisi.it

Received: 7 March 2020; Accepted: 8 April 2020; Published: 12 April 2020



Abstract: A soft-constrained neural network for density estimation (SC-NN-4pdf) has recently been introduced to tackle the issues arising from the application of neural networks to density estimation problems (in particular, the satisfaction of the second Kolmogorov axiom). Although the SC-NN-4pdf has been shown to outperform parametric and non-parametric approaches (from both the machine learning and the statistics areas) over a variety of univariate and multivariate density estimation tasks, no clear rationale behind its performance has been put forward so far. Neither has there been any analysis of the fundamental theoretical properties of the SC-NN-4pdf. This paper narrows the gaps, delivering a formal statement of the class of density functions that can be modeled to any degree of precision by SC-NN-4pdfs, as well as a proof of asymptotic convergence in probability of the SC-NN-4pdf training algorithm under mild conditions for a popular class of neural architectures. These properties of the SC-NN-4pdf lay the groundwork for understanding the strong estimation capabilities that SC-NN-4pdfs have only exhibited empirically so far.

Keywords: soft-constrained neural network; probabilistic interpretation of neural networks; density estimation; nonparaly density function

1. Introduction

Density estimation has long been a fundamental open issue in statistics and pattern classification. Implicitly or explicitly, it is at the core of statistical pattern recognition and unsupervised learning [1]. Applications embrace data compression and model selection [2], coding [3], and bioinformatics [4]. Density estimation was applied to the modeling of sequences [5,6] and structured data [7,8], as well. Finally, the task of estimating conditional probability distributions is fundamental to the broad area of probabilistic graphical models [9,10]. Statistical parametric and non-parametric techniques are available to practitioners [1], but they suffer from several significant shortcomings [11]. In fact, parametric techniques require a strong assumption on the form of the probability density function (pdf) at hand, while non-parametric approaches are memory-based (i.e., prone to overfitting), overly complex in time and space, and unreliable over small data samples. Consequently, density estimation via artificial neural networks (ANN) has been receiving increasing attention from researchers. Despite the ease of training ANNs for Bayes posterior probability estimation aimed at pattern classification [12,13], learning density functions raises problems entailed by the intrinsically unsupervised task and, above all, to the requirement of satisfying (at least numerically) the axioms of probability [14]. In particular, the integral of the function realized by the ANN shall be equal to one.

Traditional ANN-based attempts to tackle the pdf estimation problem could not overcome some major shortcomings [11]. First of all, when applied to pdf estimation, the so-called “probabilistic neural network” [15] reduces to a neural version of the Parzen window estimator [1], with a hidden neuron per each Parzen kernel, such that it inherits the aforementioned drawbacks of statistical non-parametric techniques. The maximum-likelihood framework [16] does not offer explicit solutions to the numeric

computation of the integral of the neural estimate, and its application has been limited to univariate distributions. The extension of that framework handed out by [17] does not offer solutions to such issues, focusing only on the rate of convergence of the algorithm. A parametric approach is presented in [18], in the form of a self-organizing mixture network. The latter realizes a density estimator based on a self-organizing map (SOM) [19] lattice of units associated with a neuron-specific component density (e.g., a Gaussian distribution), resulting in a mixture of such component pdfs. Although the SOM-like learning procedure is original and effective, the model (being parametric) suffers from the aforementioned limitations of statistical parametric techniques. The approaches based on indirect pdf estimation by differentiation of an estimated cumulative distribution function [20,21] end up violating the axioms of probability (they do not ensure that a proper pdf is obtained) and can hardly be applied to multivariate data, as well. Given these drawbacks of traditional approaches, researchers have recently proposed more robust ANN-based techniques for multivariate density estimation [22–24]. In particular, soft-constrained neural networks for pdf estimation (SC-NN-4pdf) [25] realize a non-parametric neural density estimator that explicitly satisfies the second Kolmogorov axiom. The SC-NN-4pdf was shown to outperform both statistical estimation techniques (parametric and non-parametric) and its neural competitors over difficult multivariate, multimodal pdf estimation tasks [25]. No clear rationale behind such an empirical evidence has been put forward so far, neither has it any analysis of the fundamental theoretical properties of the SC-NN-4pdf. To this end, hereafter we deliver a mathematical study of the major theoretical properties of the SC-NN-4pdf. After introducing the pdf estimation setup (Section 2), followed by a concise review of the fundamentals of the model and its training algorithm (Section 2.1), the formal analysis is devised in Section 3. We first offer a formal statement of the class of pdfs that can be modeled by SC-NN-4pdf to any degree of precision (Section 3.1). Then, in Section 3.2 we hand out a proof of asymptotic convergence in probability under mild conditions of the SC-NN-4pdf training algorithm for a popular class of architectures, namely multilayer perceptrons (MLP) with logistic sigmoid activation functions. Conclusions are drawn in Section 4.

2. Materials and Methods

Let us write $\mathcal{T} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ to represent the sample of n multivariate random vectors, ($\mathbf{x}_k \in \mathbb{R}^d$ for $k = 1, \dots, n$) independently drawn from the unknown pdf $p(\mathbf{x})$. A feed-forward neural network (FFNN) with d input units, one output unit, one or more hidden layers is used to learn $p(\mathbf{x})$ from \mathcal{T} in an unsupervised fashion. We write $f_i(a_i)$ to represent the activation function associated with the generic neuron i in the FFNN. The quantity a_i is computed as $a_i = \sum_j w_{ij} f_j(a_j)$ where w_{ij} is the weight of the connection between neurons j and i . We write \mathbf{w} to represent the weight vector embracing all the adaptive parameters of the FFNN (including the biases of sigmoids). When the FFNN is fed with a generic input $\mathbf{x} = (x_1, \dots, x_d)$, the i -th input neuron realizes the identity function $f_i(a_i) = f_i(x_i) = x_i$. On the contrary, any nonlinearity can be associated with the hidden neurons, but for the intents and purposes of this paper we will assume that logistic sigmoids are used. Finally, the output neuron shall have an activation function whose counterdomain matches the notion of pdf, namely the range $[0, +\infty)$. There are different choices for $f_i(a_i)$ that end up satisfying this requirement, e.g., $f_i(a_i) = \exp(a_i)$. In this paper, we resort to the ReLU activation function, that is $f_i(a_i) = \max(a_i, 0)$, which allows for both the respect of the axioms of probability and the nice modeling properties offered by linear combinations of sigmoids.

In summary, the FFNN computes a function $\varphi(\mathbf{x}, \mathbf{w})$ of its input \mathbf{x} . In [25], without loss of generality, it is assumed that the random samples of interest are limited to a compact $S \subset \mathbb{R}^d$ (therefore, henceforth S can be treated as the domain of $\varphi(\mathbf{x}, \mathbf{w})$). As explained in [25], any data normalization method may be applied to the aim of satisfying this requirement. The training algorithm presented in [25] revolves around a learning rule capable of adapting \mathbf{w} given \mathcal{T} such that a proper estimate of $p(\mathbf{x})$ is achieved via $\varphi(\mathbf{x}, \mathbf{w})$. To this end, two purposes are pursued: (1) capitalizing on the information brought by \mathcal{T} so as to approximate the unknown pdf; (2) holding the FFNN back from reaching degenerate solutions, such that the constraint $\int_S \varphi(\mathbf{x}, \mathbf{w}) d\mathbf{x} = 1$ holds true. Section 2.1

reviews the fundamentals of the training algorithm presented in [25], which actually reaches both of these purposes.

2.1. Review of the Fundamentals of the Training Algorithm

The present Section summarizes the main points of the SC-NN-4pdf training algorithm, as handed out in [25]. In so doing, the notation used in the remainder of the paper is introduced, and a common ground is fixed that allows the reader to make sense of the theoretical analysis presented in Section 3. The estimate $\tilde{p}(\cdot)$ of the unknown pdf $p(\cdot)$ is defined as

$$\tilde{p}(\mathbf{x}, \mathbf{w}) = \frac{\varphi(\mathbf{x}, \mathbf{w})}{\int_S \varphi(\mathbf{x}, \mathbf{w}) d\mathbf{x}} \tag{1}$$

for all \mathbf{x} in S . The FFNN training algorithm is built on non-parametric statistics (see [1]). Given a generic pattern $\hat{\mathbf{x}} \in \mathcal{T}$, let us consider a d -ball $B(\hat{\mathbf{x}}, \mathcal{T})$ centered on $\hat{\mathbf{x}}$ and with the minimum radius $r(\hat{\mathbf{x}}, \mathcal{T})$ s.t. $|B(\hat{\mathbf{x}}, \mathcal{T}) \cap \mathcal{T}| = k_n + 1$ where $k_n = \lfloor k\sqrt{n} \rfloor$ and $k \in \mathbb{N}$ is a hyperparameter. Please note that $\hat{\mathbf{x}}$ is in $B(\hat{\mathbf{x}}, \mathcal{T}) \cap \mathcal{T}$ by construction, therefore a proper estimate of $p(\cdot)$ over $\hat{\mathbf{x}}$ shall not involve the latter in the computation: this is the rationale behind considering $k_n + 1$ patterns within the ball instead of k_n .

We write P to represent the probability that a generic pattern drawn from $p(\mathbf{x})$ is in $B(\hat{\mathbf{x}}, \mathcal{T})$. By definition of pdf, $P = \int_{B(\hat{\mathbf{x}}, \mathcal{T})} p(\mathbf{x}) d\mathbf{x}$. It is seen that $P \simeq k_n/n$ and that $\int_{B(\hat{\mathbf{x}}, \mathcal{T})} p(\mathbf{x}) d\mathbf{x} \simeq p(\hat{\mathbf{x}})V(B(\hat{\mathbf{x}}, \mathcal{T}))$ where $V(\cdot)$ denotes the volume of its argument. As a consequence, $p(\hat{\mathbf{x}}) \simeq \frac{k_n/n}{V(B(\hat{\mathbf{x}}, \mathcal{T}))}$. Therefore, the SC-NN-4pdf training algorithm consists of a gradient-descent minimization of the unsupervised criterion function

$$\mathcal{L}(\mathcal{T}, \mathbf{w}) = \frac{1}{2} \sum_{\mathbf{x}_i \in \mathcal{T}} \left(\frac{k_n/n}{V(B(\mathbf{x}_i, \mathcal{T}))} - \tilde{p}(\mathbf{x}_i, \mathbf{w}) \right)^2 + \frac{\rho}{2} \left(1 - \int_S \varphi(\mathbf{x}, \mathbf{w}) d\mathbf{x} \right)^2 \tag{2}$$

Regarding \mathbf{w} (Equation (2) represents a batch criterion, although a stochastic approach can be taken as well by minimizing on-line a similar pattern-wise criterion upon presentation of individual training patterns one at a time [25]). The first term in the definition of $\mathcal{L}(\mathcal{T}, \mathbf{w})$ aims at normalized FFNN outputs that approximate the aforementioned non-parametric estimate of $p(\cdot)$. The second term results in a “soft” constraint enforcing a unit integral of $\tilde{p}(\mathbf{x}, \mathbf{w})$ on S . The quantity $\rho \in \mathbb{R}^+$ weights the relative contribution from the constraint. In practice, ρ is realized as a vanishing quantity, such that $\rho \rightarrow 0$ as long as training proceeds, i.e., as long as the function realized by the FFNN approaches the pdf sought (a similar asymptotic convergence of ρ to zero is exploited shortly in the following theoretical analysis). The gradient-descent learning rule for the generic parameter w is devised in [25]. The learning rule prescribes a weight modification Δw in the form $\Delta w = -\eta \frac{\partial \mathcal{L}(\cdot)}{\partial w}$, where $\eta \in \mathbb{R}^+$ is the learning rate. To calculate the partial derivative of $\mathcal{L}(\mathcal{T}, \mathbf{w})$ regarding w we need to take the derivatives of both terms on the right of Equation (2). As shown in [25], the derivative of the first term can be written as

$$\begin{aligned} & \frac{\partial}{\partial w} \left\{ \frac{1}{2} \left(\frac{k_n/n}{V(B(\hat{\mathbf{x}}, \mathcal{T}))} - \frac{\varphi(\hat{\mathbf{x}}, \mathbf{w})}{\int_S \varphi(\mathbf{x}, \mathbf{w}) d\mathbf{x}} \right)^2 \right\} = \\ & = -\frac{1}{\int_S \varphi(\mathbf{x}, \mathbf{w}) d\mathbf{x}} \left(\frac{k_n/n}{V(B(\hat{\mathbf{x}}, \mathcal{T}))} - \frac{\varphi(\hat{\mathbf{x}}, \mathbf{w})}{\int_S \varphi(\mathbf{x}, \mathbf{w}) d\mathbf{x}} \right) \cdot \\ & \quad \left\{ \frac{\partial \varphi(\hat{\mathbf{x}}, \mathbf{w})}{\partial w} - \frac{\varphi(\hat{\mathbf{x}}, \mathbf{w})}{\int_S \varphi(\mathbf{x}, \mathbf{w}) d\mathbf{x}} \int_S \frac{\partial \varphi(\mathbf{x}, \mathbf{w})}{\partial w} d\mathbf{x} \right\} \end{aligned} \tag{3}$$

where Leibniz rule was applied. Similarly, the derivative of the second term can be written as [25]

$$\begin{aligned} & \frac{\partial}{\partial w} \left\{ \frac{\rho}{2} \left(1 - \int_S \varphi(\mathbf{x}, \mathbf{w}) dx \right)^2 \right\} = \\ & = -\rho \left(1 - \int_S \varphi(\mathbf{x}, \mathbf{w}) dx \right) \int_S \frac{\partial \varphi(\mathbf{x}, \mathbf{w})}{\partial w} dx. \end{aligned} \tag{4}$$

In the equations, the derivative $\frac{\partial \varphi(\mathbf{x}, \mathbf{w})}{\partial w}$ of the FFNN output regarding one of its parameters is obtained via backpropagation (BP), as usual. Let $w = w_{ij}$ represent the weight of the connection holding between the j -th neuron in layer $\ell - 1$ and the i -th neuron in layer ℓ . We can write $\frac{\partial \varphi(\mathbf{x}, \mathbf{w})}{\partial w} = \delta_i f_j'(a_j)$ where $\delta_i = f_i'(a_i)$ if ℓ represents the output layer, or $\delta_i = (\sum_{u \in \ell+1} w_{ui} \delta_u) f_i'(a_i)$ otherwise, provided that the *deltas* for the neurons in layer $\ell + 1$ had already been determined as prescribed by plain BP. Extension of the algorithm to the other adaptive parameters of the FFNN (e.g., the bias values) is straightforward.

To compute the right-hand side of Equations (3) and (4), efficient algorithms for the computation of $\int_S \varphi(\mathbf{x}, \mathbf{w}) dx$ and $\int_S \frac{\partial}{\partial w} \varphi(\mathbf{x}, \mathbf{w}) dx$ are presented in [25] that scale properly with the dimensionality of the data and the size of the network. Eventually, the resulting overall algorithm takes the form of an iterative application of BP-based gradient-descent steps where the *deltas* at the output layer are properly defined, while the backpropagation of the *deltas* downward through the network takes place in a plain BP fashion. An illustrative example of the behavior of the SC-NN-4pdf is given in Figure 1. The figure represents the estimate obtained from a sample of $n = 1000$ random observations drawn from a mixture of three generalized extreme value (GEV) pdfs with different priors and parameters. The SC-NN-4pdf architecture and the hyperparameters are the same we used in [25] for the experiments with univariate data.

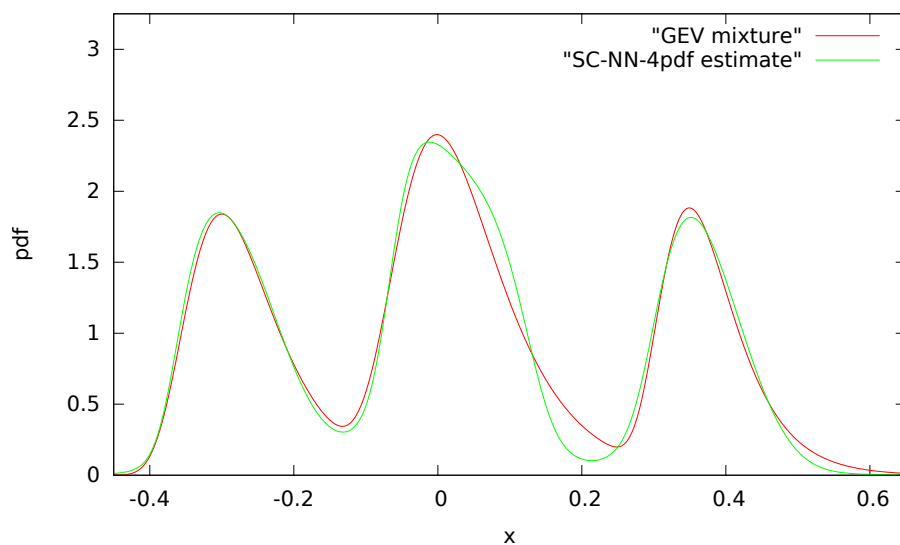


Figure 1. SC-NN-4pdf estimate of a mixture of GEV pdfs ($n = 1000$).

3. Results

We hereafter present the results of the theoretical analysis of SC-NN-4pdfs. Section 3.1 investigates the modeling capabilities of the machine, that is the class of pdfs that can actually be modeled by the machine to any degree of precision (an investigation that turns out to be straightforward in the light of [26,27]). Section 3.2 proves a Theorem of asymptotic convergence in probability of the SC-NN-4pdf training algorithm to the pdf sought under mild conditions. Both Sections rely on established results on the approximation capabilities of ANNs [26–28], as well as on non-parametric statistics [1,29,30].

3.1. Modeling Capabilities

This Section shows that for any nonpaltry pdf $p(\cdot)$ over X [11] and any $\epsilon \in \mathbb{R}^+$ a SC-NN-4pdf exists which computes $\varphi^{(\epsilon)}(\cdot)$ such that the corresponding density estimate $\tilde{p}^{(\epsilon)}(\cdot) = \frac{\varphi^{(\epsilon)}(\cdot)}{\int_X \varphi^{(\epsilon)}(\cdot) dx}$ approximates $p(\cdot)$ over X with precision ϵ . In qualitative terms, nonpaltry pdfs form the class of “interesting” pdfs $p : \mathbb{R}^d \rightarrow \mathbb{R}$ for which there is a compact subset X of I_d (where $I_d = [0, 1]^d$) covering those finite regions of \mathbb{R}^d s.t. the integral of any such $p(\cdot)$ over $\mathbb{R}^d \setminus X$ is $< \epsilon_p$ for an arbitrarily small $\epsilon_p \in \mathbb{R}^+$. In practice, whenever the support of a pdf $p(\cdot)$ of interest is not a subset of I_d , we expect that any suitable data normalization method can be applied so as to restrict the support of $p(\cdot)$ such that $\int_{\mathbb{R}^d \setminus X} p(\mathbf{x}) dx < \epsilon_p$. On the other hand, if the support of $p(\cdot)$ is a subset of I_d then it can be extended to the whole I_d by setting $p(\mathbf{x}) = 0$ for all $\mathbf{x} \in I_d \setminus X$ (note that this does not affect satisfaction of the axioms of probability). These notions are readily formalized in terms of the following, mildly stricter definition [11]:

Definition 1. Let $X \subseteq I_d$ be compact. A continuous pdf $p : \mathbb{R}^d \rightarrow \mathbb{R}$ such that $\int_X p(\mathbf{x}) dx = 1$ is said to be nonpaltry over X .

We write $\mathcal{P}(X)$ to represent the set of nonpaltry pdfs over X . As observed in [11], $\mathcal{P}(X)$ is not a linear space but a linear closure $\hat{\mathcal{P}}(X)$ of $\mathcal{P}(X)$ can be defined as $\hat{\mathcal{P}}(X) = \{ap(\cdot) + bq(\cdot) \mid a, b \in \mathbb{R}, p(\cdot), q(\cdot) \in \mathcal{P}(X)\}$. The universe $\hat{\mathcal{P}}(X)$ is actually a linear subspace of $\mathcal{C}(X)$, that is the set of all continuous, real-valued functions on X . Furthermore, if $f(\cdot) \in \mathcal{C}(X)$, $\alpha = \int_X f(\mathbf{x}) dx$ and $\alpha \neq 0$, then $p_f(\cdot) = f(\cdot)/\alpha$ is in $\hat{\mathcal{P}}(X)$, i.e., any function in $\mathcal{C}(X)$ can be written as a linear combination of functions in $\hat{\mathcal{P}}(X)$. Therefore, $\hat{\mathcal{P}}(X)$ turns out to be a Banach space with Chebyshev norm, as well. Then, since the functions in $\mathcal{P}(X)$ are bounded (in fact, they are continuous over a closed domain), the formal analysis presented in [26] applies, proving that for any nonpaltry pdf $p(\cdot)$ over X at least one SC-NN-4pdf exists that computes $\varphi^{(\epsilon)}(\cdot)$ s.t. $\|\tilde{p}^{(\epsilon)}(\cdot) - p(\cdot)\|_{\infty, X} < \epsilon$ for any $\epsilon \in \mathbb{R}^+$ (from now on we write $\|\cdot\|_{\infty, X}$ to represent the Chebyshev norm over X). In point of fact, a SC-NN-4pdf computing such $\varphi^{(\epsilon)}(\cdot)$ is found in the class of multilayer perceptrons (MLP) with only one hidden layer with logistic sigmoid activation functions.

While Cybenko’s analysis [26] revolves around the uniform norm, the arguments put forward in [27] entail likewise conclusions as pertains the L^l norm on X for $l = 1, 2, \dots$. In the following, we make the dependence of this norm on X explicit by writing $L^l(X)$. Since the functions in $\mathcal{P}(X)$ are bounded, non-negative, and l -integrable on X , Corollary 2.2 in [27] proves the following theorem.

Theorem 1. Given any nonpaltry pdf $p(\cdot)$ over X , a SC-NN-4pdf with one hidden layer of sigmoid activation functions exists that computes $\varphi^{(\epsilon)}(\cdot)$ s.t. $\|\tilde{p}^{(\epsilon)}(\cdot) - p(\cdot)\|_{L^l(X)} < \epsilon$ for any $\epsilon \in \mathbb{R}^+$ and for any $l \in \mathbb{N}$.

Qualitatively speaking, these modeling capabilities of SC-NN-4pdfs (as well as the convergence Theorem presented in the next Section) can be seen as consequences of the fact that since $\|\varphi^{(\epsilon)}(\cdot) - p(\cdot)\|_{L^1(X)}$ is arbitrarily small [27] then $\int_X \varphi^{(\epsilon)}(\cdot) dx$ is arbitrarily close to 1 (see Equation (1)).

3.2. Asymptotic Convergence in Probability

Aside from the theoretical modeling capabilities presented in Section 3.1, the SC-NN-4pdf training process can be shown to actually converge in probability to the true, nonpaltry, and unknown pdf $\hat{p}(\cdot) \in \mathcal{P}(X)$ underlying the distribution of the data at hand, under mild condition, for the popular class of MLPs with a single hidden layer with logistic sigmoid activation functions. The formal proof is given hereafter. It is seen that convergence cannot be guaranteed in general, nonetheless convergence to a pdf that is arbitrarily close to $\hat{p}(\cdot)$ is proven under the mild conditions stated in [31,32].

Let $\varphi(\cdot, A, \mathbf{w})$ be the function on X computed by a SC-NN-4pdf with support X (i.e., $S = X$), architecture A , and parameters \mathbf{w} to be learned from the data. Accordingly, $\varphi(\cdot, A, \mathbf{w}_n)$ represents the

estimate of $\hat{p}(\cdot)$ realize via the SC-NN-4pdf trained over a given training sample $\mathcal{T}_n = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ of n independent random vectors drawn from $\hat{p}(\cdot)$, that is $\phi(\cdot, A, \mathbf{w}_n) = \frac{\varphi(\cdot, A, \mathbf{w}_n)}{\int_X \varphi(\cdot, A, \mathbf{w}_n) dx}$. In fact, $\phi(\cdot, A, \mathbf{w}_n)$ is a random quantity that depends on \mathcal{T}_n , i.e., on the value \mathbf{w}_n of the SC-NN-4pdf parameters upon completion of the training stage on \mathcal{T}_n . Hereafter, we analyze the behavior of SC-NN-4pdfs (that is, of the estimators $\phi(\cdot, A, \mathbf{w}_n)$ they learn to compute) as $n \rightarrow \infty$. First, note that $\frac{k_n/n}{V(B(\mathbf{x}, \mathcal{T}_n))} \in \mathcal{C}(X)$ by construction. Moreover, since $k_n = \lfloor k\sqrt{n} \rfloor$ with $k \in \mathbb{N}$, it is seen that $V(B(\mathbf{x}, \mathcal{T}_n))$ converges asymptotically to zero, that is $\lim_{n \rightarrow \infty} V(B(\mathbf{x}, \mathcal{T}_n)) = 0$, yet slower than $1/n$, that is $\lim_{n \rightarrow \infty} nV(B(\mathbf{x}, \mathcal{T}_n)) = \infty$.

Let $\epsilon \in \mathbb{R}^+$ be any (small) positive real value, which is to say the “degree of precision” of the SC-NN-4pdf solution sought. Since \mathcal{T}_n is a random quantity, we study the convergence of the SC-NN-4pdf estimate $\phi(\cdot, A, \mathbf{w}_n)$ to $\hat{p}(\cdot)$ in mean square (due to Markov’s inequality, this entails convergence in probability). Convergence in mean square prescribes [33] that an integer n_0 exists such that for all $n > n_0$, the following inequalities hold:

$$E \left[\|\phi(\cdot, A, \mathbf{w}_n) - \hat{p}(\cdot)\|_{L^2(X)} \right] < \epsilon \tag{5}$$

and

$$\sigma^2 \left[\|\phi(\cdot, A, \mathbf{w}_n) - \hat{p}(\cdot)\|_{L^2(X)} \right] < \epsilon \tag{6}$$

where $E[\cdot]$ denotes the expectation of a random variable (such an expected value is ideally evaluated over all possible samples $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ drawn from $\hat{p}(\cdot)$) and $\sigma^2[\cdot]$ is the corresponding variance. Let $\epsilon_1 \in \mathbb{R}^+$ be s.t. $\epsilon_1 < \epsilon$. Theorem 1 guarantees that at least one SC-NN-4pdf with a single hidden layer of logistic sigmoid activation functions exists that computes $\varphi^{(\epsilon_1)}(\cdot, A^{(\epsilon_1)}, \mathbf{w}^{(\epsilon_1)})$ s.t. the estimate $\phi^{(\epsilon_1)}(\cdot, A^{(\epsilon_1)}, \mathbf{w}^{(\epsilon_1)}) = \varphi^{(\epsilon_1)}(\cdot, A^{(\epsilon_1)}, \mathbf{w}^{(\epsilon_1)}) / \int_X \varphi(\mathbf{x}, A^{(\epsilon_1)}, \mathbf{w}^{(\epsilon_1)}) dx$ satisfies $\|\hat{p}(\cdot) - \phi^{(\epsilon_1)}(\cdot, A^{(\epsilon_1)}, \mathbf{w}^{(\epsilon_1)})\|_{L^2(X)} < \epsilon_1$. The actual convergence of a given SC-NN-4pdf depends on its architecture. Therefore (and, since ϵ and ϵ_1 have been fixed), the present analysis focuses on the class of SC-NN-4pdfs with architecture $A^{(\epsilon_1)}$ and whose parameters \mathbf{w} are to be learned from the training sample. Of course, these SC-NN-4pdfs compute \mathbf{w} -specific functions $\varphi(\cdot, A^{(\epsilon_1)}, \mathbf{w})$. In the light of the theoretical results found in [32], let us assume that $\ker[\mathcal{X}_n^T] \cap \mathcal{D}_1^{\mathcal{Y}} = \{0\}$ for all the values of n . Herein, the writing \mathcal{X}_n^T denotes the transpose of the $n \times (d + 1)$ matrix yielded by concatenation of the n training vectors $\mathbf{x}_1, \dots, \mathbf{x}_n$ (each of them being d -dimensional), where the “+1” accounts for the presence of the bias terms in the sigmoid activation functions of the hidden layer. The quantity $\mathcal{D}_1^{\mathcal{Y}}$ represents the set of all *delta layer traces* (refer to [31]) $\mathcal{Y}_1 = \left[\frac{\partial C_n(\mathbf{w})}{\partial a_j(m)} \right]$ for the hidden layer of $A^{(\epsilon_1)}$, generated by varying all the parameters \mathbf{w} over their domains, where: $C_n(\mathbf{w}) = \frac{1}{2} \sum_{\mathbf{x}_i \in \mathcal{T}_n} \left(\frac{k_n/n}{V(B(\mathbf{x}_i, \mathcal{T}_n))} - \varphi(\mathbf{x}_i, A^{(\epsilon_1)}, \mathbf{w}) \right)^2$ (compare the latter quantity with the first term in Equation (2)), j ranges over all the hidden neurons in $A^{(\epsilon_1)}$, m ranges over the training patterns ($m = 1, \dots, n$), and $a_j(m)$ is the input fed to j -th hidden neuron when the SC-NN-4pdf is fed with m -th input pattern. The following convergence Theorem can now be proven to hold true:

Theorem 2. Let $\epsilon \in \mathbb{R}^+$, $\epsilon_1 \in \mathbb{R}^+$, and $\epsilon_1 < \epsilon$. Let $\varphi(\cdot, A^{(\epsilon_1)}, \mathbf{w})$ be the function computed by a SC-NN-4pdf with architecture $A^{(\epsilon_1)}$ and parameters \mathbf{w} . Let us assume that for all the values of n , BP training of the SC-NN-4pdf is applied to the minimization of criterion $\mathcal{L}(\mathcal{T}_n, \mathbf{w})$ with no early stopping, and that $\ker[\mathcal{X}_n^T] \cap \mathcal{D}_1^{\mathcal{Y}} = \{0\}$ holds true. Let λ_{\max} be the largest eigenvalue of the Hessian of $\varphi(\cdot, A^{(\epsilon_1)}, \mathbf{w})$ with respect to \mathbf{w} . If the learning rate η satisfies $|1 - \eta\lambda_{\max}| < 1$ at each step of BP, then an integer n_0 exists s.t., for all $n > n_0$, training the SC-NN-4pdf over \mathcal{T}_n converges to parameters \mathbf{w}_n with the result that $\phi(\cdot, A^{(\epsilon_1)}, \mathbf{w}_n)$ satisfies Equations (5) and (6).

Proof. Under the hypotheses, we prove the existence of an integer n_0 s.t. both Equations (5) and (6) hold true for $n > n_0$. First, let us work out the expectation, that is Equation (5). Let $\epsilon_2 \in \mathbb{R}^+$ s.t. $\epsilon = \epsilon_1 + \epsilon_2$. Independently from the specific architecture A of the SC-NN-4pdf, we have:

$$\|\phi(\cdot, A, \mathbf{w}_n) - \hat{p}(\cdot)\|_{L^2(X)} \leq \left\| \phi(\cdot, A, \mathbf{w}_n) - \frac{k_n/n}{V(B(\cdot, \mathcal{T}_n))} \right\|_{L^2(X)} + \left\| \frac{k_n/n}{V(B(\cdot, \mathcal{T}_n))} - \hat{p}(\cdot) \right\|_{L^2(X)} \tag{7}$$

From Equation (7) we can write:

$$E \left[\|\phi(\cdot, A, \mathbf{w}_n) - \hat{p}(\cdot)\|_{L^2(X)} \right] \leq E \left[\left\| \phi(\cdot, A, \mathbf{w}_n) - \frac{k_n/n}{V(B(\cdot, \mathcal{T}_n))} \right\|_{L^2(X)} \right] + E \left[\left\| \frac{k_n/n}{V(B(\cdot, \mathcal{T}_n))} - \hat{p}(\cdot) \right\|_{L^2(X)} \right] \tag{8}$$

Given Equation (8), the existence of an integer n_0 s.t. Equation (5) holds true for $n > n_0$ is guaranteed if the following conditions hold:

$$E \left[\left\| \phi(\cdot, A, \mathbf{w}_n) - \frac{k_n/n}{V(B(\cdot, \mathcal{T}_n))} \right\|_{L^2(X)} \right] < \epsilon_1 \tag{9}$$

and

$$E \left[\left\| \frac{k_n/n}{V(B(\cdot, \mathcal{T}_n))} - \hat{p}(\cdot) \right\|_{L^2(X)} \right] < \epsilon_2 \tag{10}$$

for all $n > n_0$. It is convenient to focus on Equation (10) first. Bearing in mind that $r(\mathbf{x}, \mathcal{T}_n)$ is the radius of $B(\mathbf{x}, \mathcal{T}_n)$, let us define the function $\psi(\cdot)$ of $\mathbf{u} \in X, \mathbf{v} \in X$, and \mathcal{T}_n as follows:

$$\psi(\mathbf{u}, \mathbf{v}, \mathcal{T}_n) = \begin{cases} 1 & \text{if } \mathbf{v} \in B(\mathbf{u}, \mathcal{T}_n) \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

s.t. $\int_X \psi(\mathbf{u}, \mathbf{v}, \mathcal{T}_n) d\mathbf{v} = V(B(\mathbf{u}, \mathcal{T}_n))$. Accordingly, we have

$$\begin{aligned} E \left[\frac{k_n/n}{V(B(\mathbf{x}, \mathcal{T}_n))} \right] &= \frac{1}{n} \sum_{\mathbf{x}_j \in \mathcal{T}_n} E \left[\frac{\psi(\mathbf{x}, \mathbf{x}_j, \mathcal{T}_n)}{V(B(\mathbf{x}, \mathcal{T}_n))} \right] \\ &= \int_X \frac{\psi(\mathbf{x}, \mathbf{v}, \mathcal{T}_n)}{V(B(\mathbf{x}, \mathcal{T}_n))} \hat{p}(\mathbf{v}) d\mathbf{v} \end{aligned} \tag{12}$$

which is a convolution of the unknown density and of the normalized $\psi(\cdot)$ function. Since $\psi(\cdot)$ is well-behaved by construction and since $V(B(\mathbf{x}, \mathcal{T}_n)) \rightarrow 0$ for $n \rightarrow \infty$ s.t. $\int_X \psi(\mathbf{x}, \mathbf{v}, \mathcal{T}_n) d\mathbf{v} \rightarrow 0$ too, then in the infinite sample case $\frac{\psi(\mathbf{x}, \cdot, \mathcal{T}_n)}{V(B(\mathbf{x}, \mathcal{T}_n))}$ converges to a Dirac's delta centered at \mathbf{x} . Therefore, since $\hat{p}(\cdot)$ is continuous over its support (being nonpaltry over X), then $\lim_{n \rightarrow \infty} E \left[\frac{k_n/n}{V(B(\cdot, \mathcal{T}_n))} \right] = \hat{p}(\cdot)$. Consequently, there is an integer n_p such that Equation (10) holds true for all $n > n_p$.

Then, let us consider Equation (9). At any time during training, regardless of n and \mathcal{T}_n , the SC-NN-4pdf computing $\phi(\cdot, A^{(\epsilon_1)}, \mathbf{w})$ can be set and kept in canonical form (Theorem 3 in [32]) without affecting $\phi(\cdot, A^{(\epsilon_1)}, \mathbf{w})$. In so doing, Theorem 4 in [32] ensures that the loss function $C_n(\mathbf{w})$ presents no local minima with respect to \mathbf{w} (since by hypotheses $\ker[\mathcal{X}_n^T] \cap \mathcal{D}_1^y = \{0\}$), therefore neither $\mathcal{L}(\mathcal{T}_n, \mathbf{w})$ does (as we let $\rho \rightarrow 0$). Consequently, for any value of n , BP training (without early stopping) applied to this SC-NN-4pdf converges to the global minimum \mathbf{w}_n^* of the criterion $\mathcal{L}(\mathcal{T}_n, \mathbf{w})$, i.e., $\mathbf{w}_n = \mathbf{w}_n^*$, given the fact that the learning rate η was chosen (by hypotheses) s.t. $|1 - \eta \lambda_{\max}| < 1$ at each step of BP (see [12], Sec. 7.5.1, p. 266).

A paramount consequence of this argument is that if we let $\tilde{\mathcal{L}}_n(\mathbf{w}^{(\epsilon_1)}) = \frac{1}{2} \sum_{\mathbf{x}_i \in \mathcal{T}_n} \left(\phi^{(\epsilon_1)}(\mathbf{x}_i, A^{(\epsilon_1)}, \mathbf{w}^{(\epsilon_1)}) - \frac{k_n/n}{V(B(\mathbf{x}_i, \mathcal{T}_n))} \right)^2 + \frac{\rho}{2} \left(1 - \int_X \varphi^{(\epsilon_1)}(\mathbf{x}, A^{(\epsilon_1)}, \mathbf{w}^{(\epsilon_1)}) d\mathbf{x} \right)^2$ then $\mathcal{L}(\mathcal{T}_n, \mathbf{w}_n) \leq \tilde{\mathcal{L}}_n(\mathbf{w}^{(\epsilon_1)})$ for all the values of n . Ergo, bearing in mind that $\hat{p}(\cdot) = \lim_{n \rightarrow \infty} E \left[\frac{k_n/n}{V(B(\cdot, \mathcal{T}_n))} \right]$ as well as that $A^{(\epsilon_1)}$ and $\mathbf{w}^{(\epsilon_1)}$ were chosen in such a way that $\|\hat{p}(\cdot) - \phi^{(\epsilon_1)}(\cdot, A^{(\epsilon_1)}, \mathbf{w}^{(\epsilon_1)})\|_{L^2(X)} < \epsilon_1$, it is seen that

$$\begin{aligned} \lim_{n \rightarrow \infty} E \left[\|\mathcal{L}(\mathcal{T}_n, \mathbf{w}_n)\|_{L^2(X)} \right] &\leq \lim_{n \rightarrow \infty} E \left[\|\tilde{\mathcal{L}}_n(\mathbf{w}^{(\epsilon_1)})\|_{L^2(X)} \right] \\ &< \epsilon_1^2 \\ &< \epsilon_1 \end{aligned} \tag{13}$$

where we exploited the definition of L^2 norm, as well as the fact that the integral of the function realized by the SC-NN-4pdf converges to 1 as $\varphi(\cdot)$ approaches a pdf. Equation (13) entails the existence of an integer n_ϕ s.t. Equation (9) holds true for all $n > n_\phi$. Since Equation (10) was already proven for all $n > n_p$, and since we fixed ϵ_1 and ϵ_2 s.t. $\epsilon_1 + \epsilon_2 = \epsilon$, if we let $n_0 = \max(n_p, n_\phi)$ then Equation (8) yields $E \left[\|\phi(\cdot, A^{(\epsilon_1)}, \mathbf{w}_n) - \hat{p}(\cdot)\|_{L^2(X)} \right] < \epsilon$ for all $n > n_0$, as sought.

Then, let us work out the convergence of the variance, aiming at satisfying the asymptotic condition (6). Given the fact that the variance is non-negative, and resorting to one of its computational formulae, it is possible to write

$$\begin{aligned} 0 &\leq \sigma^2 \left[\|\phi(\cdot, A^{(\epsilon_1)}, \mathbf{w}_n) - \hat{p}(\cdot)\|_{L^2(X)} \right] \\ &= E \left[\|\phi(\cdot, A^{(\epsilon_1)}, \mathbf{w}_n) - \hat{p}(\cdot)\|_{L^2(X)}^2 \right] - \left\{ E \left[\|\phi(\cdot, A^{(\epsilon_1)}, \mathbf{w}_n) - \hat{p}(\cdot)\|_{L^2(X)} \right] \right\}^2 \\ &\leq E \left[\|\phi(\cdot, A^{(\epsilon_1)}, \mathbf{w}_n) - \hat{p}(\cdot)\|_{L^2(X)}^2 \right] \end{aligned} \tag{14}$$

where we relied on the fact that $\{E[\|\phi(\cdot, A^{(\epsilon_1)}, \mathbf{w}_n) - \hat{p}(\cdot)\|_{L^2(X)}]\}^2$ is non-negative. Following in the footsteps of the analysis we applied earlier to the convergence of the mean, it is seen that $E[\|\phi(\cdot, A^{(\epsilon_1)}, \mathbf{w}_n) - \hat{p}(\cdot)\|_{L^2(X)}^2] < \epsilon$ for all $n > n_0$ which, as a consequence of Equation (14), proves that also $\sigma^2[\|\phi(\cdot, A^{(\epsilon_1)}, \mathbf{w}_n) - \hat{p}(\cdot)\|_{L^2(X)}] < \epsilon$ for all $n > n_0$, as desired. This completes the proof. \square

4. Conclusions

Density estimation is at the core of many practical applications rooted in pattern recognition, unsupervised learning, statistical analysis. and coding. Despite its having long been investigated, it is still an open problem. On the one hand, statistical techniques suffer from severe drawbacks. On the other hand, ANN-based pdf estimation algorithms struggle to break-through due to the difficulties posed by the very nature of the unsupervised estimation task and the requirement of satisfying Kolmogorov’s axioms of probability. The topic has recently been receiving an increasing attention from the Community, and some algorithmic attempts to tackle the issues were presented in the literature. In particular, the SC-NN-4pdf was proposed and successfully applied over several univariate and multivariate density estimation tasks, yielding improvements over the established approaches. Despite the empirical evidence stressing its effectiveness, no theoretical analysis of its properties had been carried out so far. This paper contributed filling such a gap, along two major directions: (1) formal identification of the class of pdfs that can be modeled by SC-NN-4pdfs to any degree ϵ of precision, and (2) proof of the asymptotic convergence in probability of the SC-NN-4pdf training algorithm. In particular, it was shown that under the assumption $\ker[\mathcal{X}_n^T] \cap \mathcal{D}_1^y = \{0\}$, 2-layer SC-NN-4pdfs with proper architecture converge in probability to an expected solution that is close to the true pdf to the desired degree ϵ of precision. These theoretical properties of the SC-NN-4pdf lay the groundwork for understanding the strong estimation capabilities of the present family of ANN-based density estimators. The generality of the class of estimated pdfs (the nonpaltry pdfs)

and the asymptotic convergence to the true pdf underlying a data distribution make the SC-NN-4pdf a promising practical tool especially over large data samples. In particular, in classification tasks involving c classes $\omega_1, \dots, \omega_c$, the corresponding class-conditional pdfs $p(\mathbf{x}|\omega_1), \dots, p(\mathbf{x}|\omega_c)$ can be estimated via c class-specific SC-NN-4pdfs that, as n increases, tend to the true pdfs such that the estimates of the class-posterior probabilities $P(\omega_1|\mathbf{x}), \dots, P(\omega_c|\mathbf{x})$ computed relying on these pdfs estimates will, in turn, converge to the true class-posteriors, and the maximum-a-posteriori decision rule will end up realizing the ideal Bayes decision rule, i.e., the classifier with minimum probability of error. An even more intriguing application can be found in semi-supervised learning, where a limited subsample of the overall training set is actually supervised, while a huge fraction of the training data is unlabeled. Again, SC-NN-4pdf models of the class-conditional pdfs can be estimated from the supervised subsample first, and then applied to the unsupervised data, where a maximum-likelihood criterion based on the SC-NN-4pdf outputs (due to their being proper pdfs) can be applied in order to label a fraction of the unlabeled patterns, enlarging the supervised subset, and so forth. Finally, since the SC-NN-4pdf training algorithm offers a technique for sampling from the neural network (a step required in the computation of the integrals involved), once trained the SC-NN-4pdf can be applied as a generative model in order to draw new data from the pdf learned. Nevertheless, like most statistical and neural density estimators, SC-NN-4pdfs still suffer from the curse of dimensionality [1]. Although the paper did not offer a general solution to such a problem, the asymptotic convergence for $n \rightarrow \infty$ has the obvious practical implication that the curse is less and less dramatic (in probability) as the cardinality of the training set increases. Yet, Theorem 2 proves convergence, but it does not offer insights on the convergence rate, i.e., convergence may be quite slow, especially over high-dimensional feature spaces.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

pdf	Probability density function
ANN	Artificial neural network
MLP	Multilayer perceptron
FFNN	Feed-forward neural network
SC-NN-4pdf	Soft-constrained neural network for pdfs
BP	Backpropagation

References

1. Duda, R.O.; Hart, P.E.; Stork, D.G. *Pattern Classification*, 2nd ed.; Wiley-Interscience: New York, NY, USA, 2000.
2. Liang, F.; Barron, A. Exact Minimax Strategies for Predictive Density Estimation, Data Compression, and Model Selection. *IEEE Trans. Inf. Theory* **2004**, *50*, 2708–2726. [[CrossRef](#)]
3. Beirami, A.; Sardari, M.; Fekri, F. Wireless Network Compression Via Memory-Enabled Overhearing Helpers. *IEEE Trans. Wirel. Commun.* **2016**, *15*, 176–190. [[CrossRef](#)]
4. Yang, Z. *Machine Learning Approaches to Bioinformatics*; World Scientific Publishing Company: Singapore, 2010.
5. Rabiner, L.R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **1989**, *77*, 257–286. [[CrossRef](#)]
6. Trentin, E.; Scherer, S.; Schwenker, F. Emotion recognition from speech signals via a probabilistic echo-state network. *Pattern Recognit. Lett.* **2015**, *66*, 4–12. [[CrossRef](#)]
7. Bongini, M.; Rigutini, L.; Trentin, E. Recursive Neural Networks for Density Estimation Over Generalized Random Graphs. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *29*, 5441–5458. [[CrossRef](#)] [[PubMed](#)]
8. Trentin, E.; Di Iorio, E. Nonparametric small random networks for graph-structured pattern recognition. *Neurocomputing* **2018**, *313*, 14–24. [[CrossRef](#)]

9. Bairoletti, M.; Milani, A.; Santucci, V. Learning Bayesian Networks with Algebraic Differential Evolution. In Proceedings of the 15th International Conference on Parallel Problem Solving from Nature (PPSN XV), Coimbra, Portugal, 8–12 September 2018; Part II; Lecture Notes in Computer Science; Auger A., Fonseca C., Lourenço N., Machado P., Paquete L., Whitley, D., Eds.; Springer: Cham, Switzerland, 2018; Volume 11102; pp. 436–448.
10. Wang, C.; Xu, C.; Yao, X.; Tao, D. Evolutionary Generative Adversarial Networks. *IEEE Trans. Evol. Comput.* **2019**, *23*, 921–934. [[CrossRef](#)]
11. Trentin, E.; Lusnig, L.; Cavalli, F. Parzen neural networks: Fundamentals, properties, and an application to forensic anthropology. *Neural Netw.* **2018**, *97*, 137–151. [[CrossRef](#)] [[PubMed](#)]
12. Bishop, C.M. *Neural Networks for Pattern Recognition*; Oxford University Press: Oxford, UK, 1995.
13. Trentin, E.; Freno, A. Probabilistic Interpretation of Neural Networks for the Classification of Vectors, Sequences and Graphs. In *Innovations in Neural Information Paradigms and Applications*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 155–182.
14. Trentin, E.; Bongini, M. Probabilistically Grounded Unsupervised Training of Neural Networks. In *Unsupervised Learning Algorithms*; Celebi, M., Aydin, K., Eds.; Springer: Cham, Switzerland, 2016; pp. 533–558.
15. Specht, D. Probabilistic Neural Networks. *Neural Netw.* **1990**, *3*, 109–118. [[CrossRef](#)]
16. Modha, D.S.; Fainman, Y. A learning law for density estimation. *IEEE Trans. Neural Netw.* **1994**, *5*, 519–23. [[CrossRef](#)] [[PubMed](#)]
17. Modha, D.S.; Masry, E. Rate of convergence in density estimation using neural networks. *Neural Comput.* **1996**, *8*, 1107–1122. [[CrossRef](#)]
18. Yin, H.; Allinson, N.M. Self-organizing mixture networks for probability density estimation. *IEEE Trans. Neural Netw.* **2001**, *12*, 405–411. [[CrossRef](#)] [[PubMed](#)]
19. Kohonen, T. Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **1982**, *43*, 59–69. [[CrossRef](#)]
20. Vapnik, V.N.; Mukherjee, S. Support Vector Method for Multivariate Density Estimation. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2000; pp. 659–665.
21. Magdon-Ismael, M.; Atiya, A. Density estimation and random variate generation using multilayer networks. *IEEE Trans. Neural Netw.* **2002**, *13*, 497–520. [[CrossRef](#)] [[PubMed](#)]
22. Trentin, E. Soft-Constrained Nonparametric Density Estimation with Artificial Neural Networks. In Proceedings of the 7th Workshop on Artificial Neural Networks in Pattern Recognition (ANNPR), Ulm, Germany, 28–30 September 2016; Springer: Cham, Switzerland, 2016; pp. 68–79.
23. Chilinski, P.; Silva, R. Neural Likelihoods via Cumulative Distribution Functions. *arXiv* **2018**, arXiv:1811.00974.
24. Trentin, E. Maximum-Likelihood Estimation of Neural Mixture Densities: Model, Algorithm, and Preliminary Experimental Evaluation. In Proceedings of the 8th IAPR TC3 Workshop on Artificial Neural Networks in Pattern Recognition, Siena, Italy, 19–21 September 2018; pp. 178–189.
25. Trentin, E. Soft-Constrained Neural Networks for Nonparametric Density Estimation. *Neural Process. Lett.* **2018**, *48*, 915–932. [[CrossRef](#)]
26. Cybenko, G. Approximation by superposition of sigmoidal functions. *Math. Control. Signal Syst.* **1989**, *2*, 303–314. [[CrossRef](#)]
27. Hornik, K.; Stinchcombe, M.; White, H. Multilayer feedforward networks are universal approximators. *Neural Netw.* **1989**, *2*, 359–366. [[CrossRef](#)]
28. Kolmogorov, A.; Fomin, S. *Elementy Teorii Funktsii I Funktsional'nogo Analiza*; Nauka (MIR): Moscow, Russia, 1980.
29. Fukunaga, K. *Introduction to Statistical Pattern Recognition*, 2nd ed.; Academic Press: San Diego, CA, USA, 1990.
30. Dekking, F.; Kraaikamp, C.; Lopuhaä, H. *A Modern Introduction to Probability and Statistics: Understanding Why and How*; Springer: London, UK, 2005.
31. Gori, M.; Tesi, A. On the Problem of Local Minima in Backpropagation. *IEEE Trans. Pattern Anal. Mach. Intell.* **1992**, *14*, 76–86. [[CrossRef](#)]

32. Gori, M.; Tsoi, A.C. Comments on local minima free conditions in multilayer perceptrons. *IEEE Trans. Neural Netw.* **1998**, *9*, 1051–1053. [[CrossRef](#)] [[PubMed](#)]
33. Parzen, E. *Modern Probability Theory and its Applications*; John Wiley & Sons: Hoboken, NJ, USA, 1962.



© 2020 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).