QUANTITATIVE UNIVERSAL APPROXIMATION BOUNDS FOR DEEP BELIEF NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

We show that deep belief networks with binary hidden units can approximate any multivariate probability density under very mild integrability requirements on the parental density of the visible nodes. The approximation is measured in the L^q-norm for $q \in [1, \infty]$ ($q = \infty$ corresponding to the supremum norm) and in Kullback-Leibler divergence. Furthermore, we establish sharp quantitative bounds on the approximation error in terms of the number of hidden units.

Deep belief networks (DBNs) are a class of generative probabilistic models obtained by stacking several restricted Boltzmann machines (RBMs, Smolensky (1986)). For a brief introduction to RBMs and DBNs we refer the reader to the survey articles Fischer & Igel (2012; 2014); Montúfar (2016); Ghojogh et al. (2021). Since their introduction, see Hinton et al. (2006); Hinton & Salakhutdinov (2006), DBNs have been successfully applied to a variety of problems in the domains of natural language processing Hinton (2009); Jiang et al. (2018), bioinformatics Wang & Zeng (2013); Liang et al. (2014); Cao et al. (2016); Luo et al. (2019), financial markets Shen et al. (2015) and computer vision Abdel-Zaher & Eldeib (2016); Kamada & Ichimura (2016; 2019); Huang et al. (2019). However, our theoretical understanding of the class of continuous probability distributions, which can be approximated by them, is limited. The ability to approximate a broad class of probability distributions—usually referred to as *universal approximation property*—is still an open problem for DBNs with real-valued visible units. As a measure of proximity between two real-valued probability density functions, one typically considers the L^q-distance or the Kullback-Leibler divergence.

Contributions. In this article we study the approximation properties of deep belief networks for multivariate continuous probability distributions which have a density with respect to the Lebesgue measure. We show that, as $m \to \infty$, the universal approximation property holds for binary-binary DBNs with two hidden layers of sizes m and $m + 1$, respectively. Furthermore, we provide an explicit quantitative bound on the approximation error in terms of m. More specifically, the main contributions of this article are:

- For each $q \in [1, \infty)$ we show that DBNs with two binary hidden layers and parental density $\varphi : \mathbb{R}^d \to \mathbb{R}_+$ can approximate any probability density $f : \mathbb{R}^d \to \mathbb{R}_+$ in the L^q-norm, solely under the condition that $f, \varphi \in L^q(\mathbb{R}^d)$. In addition, we prove that the error admits a bound of order $\mathcal{O}\big(m^{\frac{1}{\min(q,2)}-1}\big)$ for each $q \in (1, \infty)$, where m is the number of hidden neurons.
- If the target density f is uniformly continuous and the parental density φ is bounded, we provide an approximation result in the L^∞-norm (also known as supremum or uniform norm).
- Finally, we show that continuous target densities supported on a compact subset of $\mathbb{R}^d$ and uniformly bounded away from zero can be approximated by deep belief networks with bounded parental density in Kullback-Leibler divergence. The approximation error in this case is of order $\mathcal{O}\big(m^{-1}\big)$.

Related works. One of the first approximation results for deep belief networks is due to Sutskever & Hinton (2008) and states that any probability distribution on $\{0, 1\}^d$ can be learnt by a DBN with 3×2^d hidden layers of size $d + 1$ each. This result was improved by Le Roux & Bengio (2010); Montúfar & Ay (2011) by reducing the number of layers to $\frac{2^{d-1}}{d-\log(d)}$ with d hidden units each. These

results, however, are limited to discrete probability distributions. Since most applications involve continuous probability distributions, Krause et al. (2013) considered Gaussian-binary DBNs and analyzed their approximation capabilities in Kullback-Leibler divergence, albeit without a rate. In addition, they only allow for target densities that can be written as an infinite mixture of a set of probability densities satisfying certain conditions, which appear to be hard to check in practice.

Similar questions have been studied for a variety of neural network architectures: The famous results of Cybenko (1989); Hornik et al. (1989) state that deterministic multi-layer feed-forward networks are universal approximators for a large class of Borel measurable functions, provided that they have at least one sufficiently large hidden layer. See also the articles Leshno et al. (1993); Chen & Chen (1995); Barron (1993); Burger & Neubauer (2001). Le Roux & Bengio (2008) proved the universal approximation property for RBMs and discrete target distributions. Montúfar & Morton (2015) established the universal approximation property for discrete restricted Boltzmann machines. Montúfar (2014) showed the universal approximation property for deep narrow Boltzmann machines. Montúfar (2015) showed that Markov kernels can be approximated by shallow stochastic feed-forward networks with exponentially many hidden units. Bengio & Delalleau (2011); Pascanu et al. (2014) studied the approximation properties of so-called deep architectures. Merkh & Montúfar (2019) investigated the approximation properties of stochastic feed-forward neural networks.

The recent work Johnson (2018) nicely complements the aforementioned results by obtaining an illustrative negative result: Deep narrow networks with hidden layer width at most equal to the input dimension do not posses the universal approximation property.

1 Deep Belief Networks

A restricted Boltzmann machine (RBM) is a an undirected, probabilistic, graphical model with bipartite vertices that are fully connected with the opposite class. To be more precise, we consider a simple, planar graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ for which the vertex set $\mathcal{V}$ can be partitioned into sets V and H such that the edge set is given by $\mathcal{E} = \big\{\{s,t\} : s \in V,\, t \in H\big\}$. We call vertices in V *visible* units; H contains the *hidden* units. To each of the visible units we associate the state space Ω_V and to the hidden ones we associate Ω_H. We equip $\mathcal{G}$ with a *Gibbs probability measure*

$$\pi(v,h) = \frac{e^{-\mathscr{H}(v,h)}}{\mathcal{Z}}, \qquad v \in (\Omega_V)^V,\, h \in (\Omega_H)^H,$$

where $\mathscr{H} : (\Omega_V)^V \times (\Omega_H)^H \to \mathbb{R}$ is chosen such that $\mathcal{Z} = \iint e^{-\mathscr{H}(v,h)}\, dv\, dh < \infty$. Notice that the integral becomes a sum if Ω_V (resp. Ω_H) is a discrete set. It is customary to identify the RBM with the probability measure π.

An important example are *binary-binary* RBMs. These are obtained by choosing $\Omega_V = \Omega_H = \{0,1\}$ and

$$\mathscr{H} = \langle v, Wh\rangle + \langle v, b\rangle + \langle h, c\rangle\,, \quad v \in \{0,1\}^V,\, h \in \{0,1\}^H, \tag{1}$$

where $b \in \{0,1\}^V$ and $c \in \{0,1\}^H$ are called *biases*, and $W \in \mathbb{R}^{V\times H}$ is called the *weight matrix*. We shall write for $m, n \in \mathbb{N}$,

$$\mathsf{B\text{-}RBM}(m,n) = \{\pi \text{ is a binary-binary RBM with } m \text{ visible and } n \text{ hidden units}\}\,, \tag{2}$$

for the set of binary-binary RBMs with fixed layer sizes.

The following discrete approximation result is well known, see also Montúfar & Ay (2011):

Proposition 1 (Le Roux & Bengio (2008), Theorem 2)**.** *Let $m \in \mathbb{N}$ and μ be a probability distribution on $\{0,1\}^m$. Let*

$$\operatorname{supp}(\mu) = \big\{v \in \{0,1\}^m : \mu(v) > 0\big\}$$

be the support of μ. Set $n = \big|\operatorname{supp}(\mu)\big| + 1$. Then, for each $\varepsilon > 0$, there is a $\pi \in \mathsf{B\text{-}RBM}(m,n)$ such that

$$\left|\mu(v) - \sum_{h\in\{0,1\}^n} \pi(v,h)\right| \leqslant \varepsilon \qquad \forall\, v \in \{0,1\}^m.$$

A deep belief network (DBN) is constructed by stacking two RBMs. To be more precise, we now consider a tripartite graph with hidden layers H_1 and H_2 and visible units V. We assume that the edge set is now given by $\mathcal{E} = \big\{\{s, t_1\}, \{t_1, t_2\} : s \in V, t_1 \in H_1, t_2 \in H_2\big\}$. The state spaces are now $\Omega_V = \mathbb{R}$ and $\Omega_{H_1} = \Omega_{H_2} = \{0, 1\}$. We think of edges in the graph as dependence of the neurons (in the probabilistic sense). The topology of the graph hence shows that the vertices in V and H_2 shall be conditionally independent, that is, we require that the probability distribution of a DBN satisfies

$$p(v, h_1, h_2) = p(v \mid h_1)p(h_1, h_2). \tag{3}$$

The joint density of the hidden units $p(h_1, h_2)$ will be chosen as binary-binary RBM.

Let $\mathcal{D}(\mathbb{R}^d) = \big\{f : \mathbb{R}^d \to \mathbb{R}_+ : \int_{\mathbb{R}^d} f(x)\,dx = 1\big\}$ be the set of probability densities on $\mathbb{R}^d$. For $\varphi \in \mathcal{D}(\mathbb{R}^d)$ and $\sigma > 0$ we set

$$\mathcal{V}_\varphi^\sigma = \left\{\varphi_{\mu,\sigma} = \sigma^{-d}\varphi\left(\frac{x-\mu}{\sigma}\right) : \mu \in \mathbb{R}^d\right\}. \tag{4}$$

Notice that all elements of $\mathcal{V}_\varphi^\sigma$ are themselves probability distributions. We fix a *parental density* $\varphi \in \mathcal{D}(\mathbb{R}^{|V|})$ and choose the conditional density in (3) as $p(\cdot \mid h_1) \in \mathcal{V}_\varphi^\sigma$ for each $h_1 \in H_1$.

Example 2. *The most popular choice of the parental function φ in* (4) *is the d-dimensional standard Gaussian density*

$$\varphi(x) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{|x|^2}{2}\right), \qquad x \in \mathbb{R}^d. \tag{5}$$

Another density considered in previous works is the truncated exponential distribution

$$\varphi(x) = \prod_{i=1}^d \frac{\lambda_i e^{-\lambda_i x_i}}{1 - e^{-b_i\lambda_i}} \mathbb{1}_{[0,b_i]}(x_i), \qquad x = (x_1, \ldots, x_d) \in \mathbb{R}^d, \tag{6}$$

where $b_i, \lambda_i > 0$ for each $i = 1, \ldots, d$.

Similar to (2), we collect all DBNs in the set

$$\mathsf{DBN}_\varphi(d, m, n) = \Big\{p \text{ is a DBN with parental density } \varphi, d \text{ visible units}, m \text{ hidden units on the first level, and } n \text{ hidden units on the second level}\Big\},$$

where $\varphi \in \mathcal{D}(\mathbb{R}^d)$ and $d, m, n \in \mathbb{N}$.

2 Main Results

To state the results of this article, we need to introduce three bits of additional notation:

Let $q \in [1, \infty)$. The space of q-summable functions is denoted by

$$L^q(\mathbb{R}^d) = \left\{f : \mathbb{R}^d \to \mathbb{R} : \|f\|_{L^q} = \left(\int_{\mathbb{R}^d} \big|f(x)\big|^q\,dx\right)^{\frac{1}{q}} < \infty\right\}.$$

We also set

$$L^\infty(\mathbb{R}^d) = \left\{f : \mathbb{R}^d \to \mathbb{R} : \|f\|_{L^\infty} = \sup_{x\in\mathbb{R}^d} \big|f(x)\big| < \infty\right\}.$$

It is convenient to declare $\mathcal{D}_q(\mathbb{R}^d) = \mathcal{D}(\mathbb{R}^d) \cap L^q(\mathbb{R}^d)$. Finally, let us abbreviate the constant

$$\Upsilon_q = \max\left(1, \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |x|^q e^{-\frac{x^2}{2}}\,dx\right)^{\frac{1}{q}} = \begin{cases} 1 & q \leqslant 2, \\ \dfrac{\sqrt{2}}{\pi^{\frac{1}{2q}}}\Gamma\left(\dfrac{q+1}{2}\right), & q > 2, \end{cases} \tag{7}$$

with the Gamma function $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\,dt$, $x > 0$.

The main results of this paper are stated in the following two theorems:

Theorem 3. *Let $q \in [1, \infty)$ and $f, \varphi \in \mathcal{D}_q(\mathbb{R}^d)$. Then, for any $\varepsilon > 0$, there is an $M \in \mathbb{N}$ such that, for each $m \geqslant M$, we can find a $p \in \mathsf{DBN}_\varphi(d, m, m+1)$ satisfying*

$$\|f - p\|_{L^q} \leqslant \varepsilon.$$

If $q \in (1, \infty)$, then, for each $m \in \mathbb{N}$, the following quantitative bound holds:

$$\inf_{p \in \mathsf{DBN}_\varphi(d,m,m+1)} \|f - p\|_{L^q} \leqslant \frac{2\Upsilon_q \|\varphi\|_{L^q}}{m^{1 - \frac{1}{\min(q,2)}}}, \tag{8}$$

where the constant Υ_q is defined in (7)*.*

Remark 4. *Returning to Example 2, we find that $\|\varphi\|_{L^q} = q^{-\frac{d}{2q}}$ for the d-dimensional standard normal distribution* (5) *and*

$$\|\varphi\|_{L^q} = \prod_{i=1}^{d} \frac{\lambda_i^{1-\frac{1}{q}}}{q^{\frac{1}{q}}\left(1 - e^{-b_i \lambda_i}\right)} \left(1 - e^{-q\lambda_i b_i}\right)^{\frac{1}{q}}$$

for the truncated exponential distribution (6)*. Our bound* (8) *thus shows that deep belief networks with truncated exponential parental density (for suitable choice of the parameters b and λ) better approximate the target density f. This is especially prevalent for small q, which is the primary case of interest, see Corollary 7 below.*

To state the approximation in the L^∞-norm, we need to introduce the space of bounded and uniformly continuous functions:

$$\mathcal{C}_u(\mathbb{R}^d) = \left\{ f \in L^\infty(\mathbb{R}^d) : \lim_{\delta \downarrow 0} \sup_{|x-y| \leqslant \delta} \big|f(x) - f(y)\big| = 0 \right\}.$$

Notice that any probability density $f \in \mathcal{D}(\mathbb{R}^d)$, which is differentiable and has a bounded derivative, belongs to $\mathcal{C}_u(\mathbb{R}^d)$ since any uniformly continuous and integrable function is bounded.

Theorem 5. *Let $f \in \mathcal{D}(\mathbb{R}^d) \cap \mathcal{C}_u(\mathbb{R}^d)$ and $\varphi \in \mathcal{D}_\infty(\mathbb{R}^d)$. Then, for any $\varepsilon > 0$, there is an $M \in \mathbb{N}$ such that, for each $m \geqslant M$, we can find a $p \in \mathsf{DBN}_\varphi(d, m, m+1)$ satisfying*

$$\|f - p\|_{L^\infty} \leqslant \varepsilon.$$

Remark 6. *The uniform continuity requirement on the target density in Theorem 5 can actually be relaxed to essential uniform continuity, that is, f is uniformly continuous except on a set with zero Lebesgue measure. The most notable example of such a function is the uniform distribution $f = \mathbb{1}_{[0,1]}$.*

Another important metric between between probability densities $p, q : \mathbb{R}^d \to \mathbb{R}_+$ is the *Kullback-Leibler divergence* (or *relative entropy*) defined by

$$\mathrm{KL}(f \| g) = \int_{\mathbb{R}^d} f(x) \log\left(\frac{f(x)}{g(x)}\right) dx$$

if $\{x \in \mathbb{R}^d : g(x) = 0\} \subset \{x \in \mathbb{R}^d : f(x) = 0\}$ and $\mathrm{KL}(f\|g) = \infty$ otherwise. From Theorems 3 and 5 we can deduce the following quantitative approximation bound in the Kullback-Leibler divergence:

Corollary 7. *Let $\varphi \in \mathcal{D}_\infty(\mathbb{R}^d)$. Let $\Omega \subset \mathbb{R}^d$ be a compact set and $f : \Omega \to \mathbb{R}_+$ be a continuous probability density. Suppose that there is an $\eta > 0$ such that both $f \geqslant \eta$ and $\varphi \geqslant \eta$ on Ω. Then there is a constant $M > 0$ such that, for each $m \in \mathbb{N}$, it holds that*

$$\inf_{p \in \mathsf{DBN}_\varphi(d,m,m+1)} \mathrm{KL}(f\|p) \leqslant \frac{M}{\eta m}\left(8\|\varphi\|_{L^2}^2 + \|f - \varphi\|_{L^2(\Omega)}^2\right), \tag{9}$$

where $\|f - \varphi\|_{L^2(\Omega)}^2 = \int_\Omega \big|f(x) - \varphi(x)\big|^2 dx$.

Let us note that any $\varphi \in \mathcal{D}_\infty(\mathbb{R}^d)$ is square-integrable so that the right-hand side of the bound (9) is actually finite. This follows from the *interpolation inequality*

$$\|\varphi\|_{L^2} \leqslant \sqrt{\|\varphi\|_{L^1}\|\varphi\|_{L^\infty}} = \sqrt{\|\varphi\|_{L^\infty}}, \tag{10}$$

see (Brezis, 2011, Exercise 4.4).

Corollary 7 considerably generalizes the results of (Krause et al., 2013, Theorem 7): There, the authors only prove that deep belief networks can approximate any density in the closure of the convex hull of a set of probability densities satisfying certain conditions, which appear to be difficult to check in practice. That work also does not contain a convergence rate. In comparison, our results directly describe the class of admissible target densities and do not rely on the indirect description through the convex hull. Finally, there is an unjustified step in the argument of Krause et al., which appears hard to reconcile, see Remark 15 below for details.

3 PROOFS

This section presents the proofs of Theorems 3, 5 and Corollary 7. As a first step, we shall establish a couple of preliminary results in the next two subsections.

3.1 L^q-APPROXIMATION OF FINITE MIXTURES

Given a set $A \subset L^q(\mathbb{R}^d)$, the *convex hull* of A is by definition the smallest convex set containing A; in symbols $\operatorname{conv}(A)$. It can be shown that

$$\operatorname{conv}(A) = \left\{ \sum_{i=1}^{n} \alpha_i a_i : \alpha = (\alpha_1, \dots, \alpha_n) \in \triangle_n, a_1, \dots, a_n \in A, n \in \mathbb{N} \right\}$$

with $\triangle_n = \left\{ x \in [0,1]^n : \sum_{i=1}^n x_i = 1 \right\}$, the n-dimensional standard simplex. It is also convenient to introduce the *truncated convex hull*

$$\operatorname{conv}_m(A) = \left\{ \sum_{i=1}^{m} \alpha_i a_i : \alpha = (\alpha_1, \dots, \alpha_m) \in \triangle_m, a_1, \dots, a_m \in A \right\}$$

for $m \in \mathbb{N}$ so that $\operatorname{conv}(A) = \bigcup_{m \in \mathbb{N}} \operatorname{conv}_m(A)$. The *closed convex hull* $\overline{\operatorname{conv}}(A)$ is the smallest closed convex set containing A and it is straight-forward to check that it coincides with the closure of $\operatorname{conv}(A)$ in the topology of $L^q(\mathbb{R}^d)$.

The next result shows that we can approximate any probability density in the truncated convex hull of the set (4) arbitrarily well by a DBN with a fixed number of hidden units:

Lemma 8. *Let $q \in [1, \infty]$, $\varphi \in \mathcal{D}_q(\mathbb{R}^d)$, $\sigma > 0$, and $m \in \mathbb{N}$. Then, for every $f \in \operatorname{conv}_m(\mathcal{V}_\varphi^\sigma)$ and every $\varepsilon > 0$, there is a deep belief network $p \in \mathsf{DBN}_\varphi(d, m, m+1)$ such that*

$$\|f - p\|_{L^q} \leqslant \varepsilon.$$

Proof. Since $f \in \operatorname{conv}_m(\mathcal{V}_\varphi^\sigma)$, there are by definition of $\triangle_m$ $(\alpha_1, \dots, \alpha_m) \in \triangle_m$ and $(\mu_1, \dots, \mu_m) \in \left(\mathbb{R}^d\right)^m$ such that

$$f = \sum_{i=1}^{m} \alpha_i \varphi_{\mu_i, \sigma}.$$

We can think of $\alpha = (\alpha_1, \dots, \alpha_m)$ as a probability distribution $\tilde{\alpha}$ on $\{0,1\}^m$ by declaring

$$\tilde{\alpha}(h_1) = \begin{cases} \alpha_i, & \text{if } h_1 = e_i, \\ 0, & \text{else,} \end{cases} \qquad h_1 \in \{0,1\}^m,$$

where $(e_i)_j = \delta_{i,j}$, $j = 1, \dots, m$, is the ith unit vector.

Let us fix $q \in [1, \infty]$ and $\sigma > 0$. By Proposition 1 there is a $\pi \in \mathsf{B\text{-}RBM}(m, m+1)$ such that

$$\left| \tilde{\alpha}(h_1) - \sum_{h_2 \in \{0,1\}^{m+1}} \pi(h_1, h_2) \right| \leqslant \frac{\varepsilon}{m\sigma \|\varphi\|_{L^q}} \qquad \forall\, h_1 \in \{0,1\}^m. \tag{11}$$

We set

$$p(v \,|\, h_1) = \begin{cases} \varphi_{\mu_i, \sigma}(v), & h_1 = e_i, \\ 0, & \text{else,} \end{cases}$$

and

$$p(v, h_1, h_2) = p(v \mid h_1)\pi(h_1, h_2) \in \mathsf{DBN}_\varphi(d, m, m+1).$$

This is the desired approximation since

$$\|f - p\|_{L^q} \leqslant \sum_{i=1}^{m} \left| \alpha_i - \sum_{h_2 \in \{0,1\}^{m+1}} \pi(e_i, h_2) \right| \left\| \varphi_{\mu_i,\sigma} \right\|_{L^q} \leqslant \varepsilon,$$

where we used that $\|\varphi_{\mu,\sigma}\|_{L^q} = \sigma \|\varphi\|_{L^q}$ for each $\mu \in \mathbb{R}^d$ and each $\sigma > 0$. □

3.2 Approximation by Convolution

Let $f \in L^q(\mathbb{R}^d)$, $q \in [1, \infty]$, and $\varphi \in \mathcal{D}(\mathbb{R}^d)$. We denote the *convolution* of f and $\varphi_\sigma(\cdot) = \sigma^{-d}\varphi\big(\sigma^{-1}\,\cdot\big)$ by

$$\big(f \star \varphi_\sigma\big)(x) = \int_{\mathbb{R}^d} f(\mu)\varphi_\sigma(x - \mu)\, d\mu = \int_{\mathbb{R}^d} f(\mu)\varphi_{\mu,\sigma}(x)\, d\mu.$$

Young's convolution inequality, Young (1912), implies $f \star \varphi_\sigma \in L^q(\mathbb{R}^d)$:

$$\big\| f \star \varphi_\sigma \big\|_{L^q} \leqslant \|f\|_{L^q} \|\varphi_\sigma\|_{L^1} = \|f\|_{L^q}.$$

In addition, the following approximation result holds:

Proposition 9. *Let $\varphi \in \mathcal{D}(\mathbb{R}^d)$. Then all of the following hold true:*

1. *For each $q \in [1, \infty)$ and each $f \in L^q(\mathbb{R}^d)$, we have*
$$\lim_{\sigma \downarrow 0} \big\| f - f \star \varphi_\sigma \big\|_{L^q} = 0.$$

2. *If $f \in L^\infty(\mathbb{R}^d) \cap \mathcal{C}_u(\mathbb{R}^d)$, then*
$$\lim_{\sigma \downarrow 0} \big\| f - f \star \varphi_\sigma \big\|_{L^\infty} = 0.$$

Proof. Item 1 is well known, see e.g. (Folland, 1999, Theorem 8.14). For 2 fix $\varepsilon > 0$. By uniform continuity of f, we can find a $\delta > 0$ such that

$$\sup_{|\mu| \leqslant \delta} \big| f(x) - f(x - \mu) \big| \leqslant \frac{\varepsilon}{2} \qquad \forall\, x \in \mathbb{R}^d.$$

In particular,

$$\Big| f(x) - \big(f \star \varphi_\sigma\big)(x) \Big| \leqslant \int_{\mathbb{R}^d} \varphi_\sigma(\mu) \big| f(x) - f(x - \mu) \big|\, d\mu \leqslant 2\|f\|_{L^\infty} \int_{\{|\mu| > \delta\}} \varphi_\sigma(\mu)\, d\mu + \frac{\varepsilon}{2}.$$

Since

$$\int_{\{|\mu| > \delta\}} \varphi_\sigma(\mu)\, d\mu = \int_{\left\{|\mu| > \frac{\delta}{\sigma}\right\}} \varphi(\mu)\, d\mu \to 0 \qquad \text{as } \sigma \downarrow 0,$$

we can choose $\sigma_0 > 0$ such that $\big\| f - (f \star \varphi_\sigma) \big\|_{L^\infty} \leqslant \varepsilon$ for all $\sigma \in (0, \sigma_0)$. This completes the proof. □

3.3 Approximation Theory in Banach Spaces

The second ingredient needed in the proof of Theorem 3 is an abstract result from the geometric theory of Banach spaces. To formulate it, we need to introduce the following notion: The Rademacher type of a Banach space $\big(\mathcal{X}, \|\cdot\|_{\mathcal{X}}\big)$ the largest number $\mathfrak{t} \geqslant 1$ for which there is a constant $C > 0$ such that, for each $k \in \mathbb{N}$ and each $f_1, \ldots, f_k \in \mathcal{X}$,

$$\mathbb{E}\left[\left\| \sum_{i=1}^{k} \epsilon_i f_i \right\|_{\mathcal{X}}^{\mathfrak{t}} \right] \leqslant C \sum_{i=1}^{k} \|f_i\|_{\mathcal{X}}^{\mathfrak{t}}$$

holds, where $\epsilon_1, \ldots, \epsilon_k$ are i.i.d. Rademacher random variables, that is, $\mathbb{P}\big(\epsilon_1 = \pm 1\big) = \frac{1}{2}$. It can be shown that $\mathfrak{t} \leqslant 2$ for every Banach space.

Example 10. *The space $L^q(\mathbb{R}^d)$ has Rademacher type $\mathfrak{t} = \min(q, 2)$ for $q \in [1, \infty)$. The space $L^\infty(\mathbb{R}^d)$ on the other hand has only trivial type $\mathfrak{t} = 1$.*

A good reference for the above results on the Rademacher type is (Ledoux & Talagrand, 1991, Section 9.2). The next approximation result and its application to $L^q(\mathbb{R}^d)$ will become important in the sequel:

Proposition 11 (Donahue et al. (1997), Theorem 2.5)**.** *Let $\big(\mathcal{X}, \|\cdot\|_{\mathcal{X}}\big)$ be a Banach space of Rademacher type $\mathfrak{t} \in [1, 2]$. Let $A \subset \mathcal{X}$ and $f \in \overline{\mathrm{conv}}(A)$. Suppose that $\xi = \sup_{g \in A} \|f - g\|_{\mathcal{X}} < \infty$. Then there is a constant $C > 0$ only depending on the Banach space $\big(\mathcal{X}, \|\cdot\|_{\mathcal{X}}\big)$ such that, for each $m \in \mathbb{N}$, we can find an element $g \in \mathrm{conv}_m(A)$ satisfying*

$$\|f - g\|_{\mathcal{X}} \leqslant \frac{C\xi}{m^{1-\frac{1}{\mathfrak{t}}}}. \tag{12}$$

Notice that the bound (12) is of course useless for $\mathfrak{t} = 1$. In addition, it can be shown that the convergence rate $m^{\frac{1}{\mathfrak{t}}-1}$ is sharp, see (Donahue et al., 1997, Section 2.3).

Corollary 12. *Let $A \subset L^q(\mathbb{R}^d)$, $1 \leqslant q < \infty$, and suppose that $f \in \overline{\mathrm{conv}}(A)$. If $\xi = \sup_{g \in A} \|f - g\|_{\mathcal{X}} < \infty$, then for all $m \in \mathbb{N}$, there is a $g \in \mathrm{conv}_m(A)$ such that*

$$\|f - h\|_{L^q} \leqslant \frac{\Upsilon_q \xi}{m^{1-\frac{1}{\min(q,2)}}},$$

where Υ_q is the constant defined in (7)*.*

Proof. Owing to Example 10 we are in the regime of Proposition 11. The sharp constant $C = \Upsilon_q$ was derived in Haagerup (1981). □

3.4 Proof of Theorems 3 and 5

Before giving the technical details of the proofs, let us provide an overview of the strategy:

1. By Proposition 9 we can approximate the density $f \in \mathcal{D}_q(\mathbb{R}^d)$ with $f \star \varphi_\sigma$ up to an error which vanishes as $\sigma \downarrow 0$.
2. Upon showing that $f \star \varphi_\sigma \in \overline{\mathrm{conv}}(\mathcal{V}_\varphi^\sigma)$, Proposition 12 allows us to show that for each $\varepsilon > 0$ and each $m \in \mathbb{N}$, we can pick $\sigma > 0$ such that

$$\inf_{g \in \mathrm{conv}_m(\mathcal{V}_\varphi^\sigma)} \|f - g\|_{L^q} \leqslant \varepsilon + \frac{2\Upsilon_q \|\varphi\|_{L^q}}{m^{1-\frac{1}{\min(q,2)}}}.$$

3. Finally, we employ Lemma 8 to conclude the desired estimate (8).

Lemma 13. *Let $q \in [1, \infty]$, $f \in \mathcal{D}_q(\mathbb{R}^d)$, and $\varphi \in \mathcal{D}(\mathbb{R}^d)$. Then, for each $\sigma > 0$, we have*

$$f \star \varphi_\sigma \in \overline{\mathrm{conv}}(\mathcal{V}_\varphi^\sigma),$$

with the closure understood with respect to the norm $\|\cdot\|_{L^q}$.

Proof. Let us abbreviate $g = f \star \varphi_\sigma$. We argue by contradiction. Suppose that $g \notin \overline{\mathrm{conv}}(\mathcal{V}_\varphi^\sigma)$. As a consequence of the Hahn-Banach theorem, g is separated from $\overline{\mathrm{conv}}(\mathcal{V}_\varphi^\sigma)$ by a hyperplane. More precisely, there is a continuous linear function $\rho : L^q(\mathbb{R}^d) \to \mathbb{R}$ such that $\rho(h) < \rho(g)$ for all $h \in \overline{\mathrm{conv}}(\mathcal{V}_\varphi^\sigma)$, see (Brezis, 2011, Theorem. 1.7). On the other hand, we however have

$$\rho(g) = \rho\left(\int_{\mathbb{R}^d} f(\mu)\varphi_{\mu,\sigma}\, d\mu\right) = \int_{\mathbb{R}^d} f(\mu)\rho\big(\varphi_{\mu,\sigma}\big)\, d\mu < \rho(g) \int_{\mathbb{R}^d} f(\mu)\, d\mu = \rho(g),$$

which is the desired contradiction. □

We can now establish the main results of this article:

Theorems 3 and 5. Let us first assume that $q \in (1, \infty)$ and prove the quantitative bound (8). To this end fix $\varepsilon > 0$ and $m \in \mathbb{N}$. We first observe that, by Proposition 9, we can choose $\sigma > 0$ sufficiently small such that $\big\|f - f \star \varphi_\sigma\big\|_{L^q} \leqslant \frac{\varepsilon}{2}$. Employing Lemma 13 and Corollary 12 with $A = \mathcal{V}_\varphi^\sigma$, we can find a $g_m \in \operatorname{conv}_m(\mathcal{V}_\varphi^\sigma)$ such that

$$\|f - g_m\|_{L^q} \leqslant \big\|f - f \star \varphi_\sigma\big\|_{L^q} + \big\|f \star \varphi_\sigma - g_m\big\|_{L^q} \leqslant \frac{\varepsilon}{2} + \frac{\Upsilon_q}{m^{1-\frac{1}{\min(q,2)}}} \sup_{\mu \in \mathbb{R}^d} \big\|f \star \varphi_\sigma - \varphi_{\mu,\sigma}\big\|_{L^q}.$$

For the last term we bound

$$\begin{aligned}
\sup_{\mu \in \mathbb{R}^d} \big\|f \star \varphi_\sigma - \varphi_{\mu,\sigma}\big\|_{L^q} &= \sup_{\mu \in \mathbb{R}^d} \left(\int_{\mathbb{R}^d} \left| \int_{\mathbb{R}^d} f(x) \big(\varphi_\sigma(y-x) - \varphi_\sigma(y-\mu)\big)\, dx \right|^q dy \right)^{\frac{1}{q}} \\
&\leqslant \int_{\mathbb{R}^d} f(x) \sup_{\mu \in \mathbb{R}^d} \left(\int_{\mathbb{R}^d} \big|\varphi_\sigma(y-x) - \varphi_\sigma(y-\mu)\big|^q dy \right)^{\frac{1}{q}} \\
&= \sup_{\mu \in \mathbb{R}^d} \big\|\varphi - \varphi_{\mu,1}\big\|_{L^q} \leqslant 2\|\varphi\|_{L^q},
\end{aligned}$$

whence

$$\|f - g_m\|_{L^q} \leqslant \frac{\varepsilon}{2} + \frac{2\Upsilon_p \|\varphi\|_{L^q}}{m^{1-\frac{1}{\min(q,2)}}}.$$

Finally, Lemma 8 allows us to choose $p \in \mathsf{DBN}_\varphi(d, m, m{+}1)$ such that $\|g_m - p\|_{L^q} \leqslant \frac{\varepsilon}{2}$. Therefore, we conclude

$$\|f - p\|_{L^q} \leqslant \varepsilon + \frac{2\Upsilon_q \|\varphi\|_{L^q}}{m^{1-\frac{1}{\min(q,2)}}}.$$

Since $\varepsilon > 0$ was arbitrary, the bound (8) follows.

If $q = 1$ or $q = \infty$, we use the fact that

$$\overline{\operatorname{conv}}(A) = \overline{\bigcup_{m \in \mathbb{N}} \operatorname{conv}_m(A)}$$

for any subset A of either $L^1(\mathbb{R}^d)$ or $L^\infty(\mathbb{R}^d)$, respectively. This implies that, for each $\varepsilon > 0$, we can find $m \in \mathbb{N}$ and $g_m \in \operatorname{conv}_m(\mathcal{V}_\varphi^\sigma)$ such that $\big\|f \star \varphi_\sigma - g_m\big\|_{L^q} \leqslant \frac{\varepsilon}{3}$. If $q = \infty$, we note that a uniformly continuous and integrable function is always bounded. Hence, in any case we can apply Proposition 9 to find a $\sigma > 0$ for which $\big\|f - f \star \varphi_\sigma\big\|_{L^q} \leqslant \frac{\varepsilon}{3}$. Finally employing Lemma 8 as above, there is a $p \in \mathsf{DBN}_\varphi(d, m, m+1)$ such that

$$\big\|f - p\big\|_{L^q} \leqslant \big\|f - f \star \varphi_\sigma\big\|_{L^q} + \big\|f \star \varphi_\sigma - g_m\big\|_{L^q} + \big\|g_m - p\big\|_{L^q} \leqslant \varepsilon,$$

as required. □

3.5 Kullback-Leibler Approximation on Compacts

Let us begin by bounding the Kullback-Leibler divergence in terms of the L^2-norm:

Lemma 14 (Zeevi & Meir (1997), Lemma 3.3)**.** *Let $\Omega \subset \mathbb{R}^d$, $f : \Omega \to \mathbb{R}_+$, and $g : \mathbb{R}^d \to \mathbb{R}_+$ be probability densities. If there is an $\eta > 0$ such that both $f, g \geqslant \eta$ on Ω, then*

$$\mathrm{KL}(f\|g) \leqslant \frac{1}{\eta} \|f - g\|_{L^2(\Omega)}^2.$$

Proof. We use Jensen's inequality and the elementary fact $\log x \leqslant x - 1$, $x > 0$, to obtain

$$\begin{aligned}
\mathrm{KL}(f\|g) &= \int_\Omega \log\left(\frac{f(x)}{g(x)}\right) f(x)\, dx \leqslant \log\left(\int_\Omega \frac{f(x)^2}{g(x)}\, dx\right) \\
&\leqslant \int_\Omega \frac{f(x)^2}{g(x)}\, dx - 1 = \int_\Omega \frac{(f(x) - g(x))^2}{g(x)}\, dx \leqslant \frac{1}{\eta} \|f - g\|_{L^2}^2,
\end{aligned}$$

as required. □

Finally, we can prove the approximation bound in Kullback-Leibler divergence:

Corollary 7. Extending the target density f by zero on $\mathbb{R}^d \setminus \Omega$, the corollary follows from Theorem 3 upon showing that, for each $m \in \mathbb{N}$, we can choose the approximation $p \in \mathsf{DBN}_\varphi(d, m, m+1)$ in such a way that $p \geqslant \frac{\eta}{2}$ on Ω.

To see this, we notice that f is uniformly continuous since Ω is compact. Hence, Theorem 5 allows us to pick an $M \in \mathbb{N}$ such that, for each $m \geqslant M$, there is a $p_m \in \mathsf{DBN}_\varphi(d, m, m+1)$ with $\|f - p_m\|_{L^\infty} \leqslant \frac{\eta}{2}$. In particular, each of these DBNs satisfies $p_m \geqslant \frac{\eta}{2}$ on Ω. Consequently, by Lemma 14 we obtain

$$\inf_{p \in \mathsf{DBN}_\varphi(d,m,m+1)} \mathrm{KL}(f \| p) \leqslant \frac{8\|\varphi\|_{L^2}^2}{\eta m} \qquad \forall m \geqslant M. \tag{13}$$

A crude upper bound on $\inf_{p \in \mathsf{DBN}_\varphi(d,m,m+1)} \mathrm{KL}(f \| p)$ for $m < M$ can be obtained choosing both zero weights and biases in (1) as well as $p(v \mid h_1) = \varphi$ for each $h_1 \in \{0,1\}^m$ in (3). Hence, the visible units of the DBN have density φ. This gives

$$\inf_{p \in \mathsf{DBN}_\varphi(d,m,m+1)} \mathrm{KL}(f \| p) \leqslant \mathrm{KL}(f \| \varphi) \leqslant \frac{1}{\eta} \|f - \varphi\|_{L^2(\Omega)}^2 \qquad \forall\, m = 1, \dots, M-1, \tag{14}$$

again by Lemma 14. Finally, combining (13) and (14) we get

$$\inf_{p \in \mathsf{DBN}_\varphi(d,m,m+1)} \mathrm{KL}(f \| p) \leqslant \frac{M}{\eta m} \left(8\|\varphi\|_{L^2}^2 + \|f - \varphi\|_{L^2(\Omega)}^2 \right),$$

as required. □

Remark 15. *Our strategy of the proof of the Kullback-Leibler approximation in Corollary 7 through Lemma 14 differs from the one employed in (Krause et al., 2013, Theorem 7). There, the authors built on the results of Li & Barron (1999) and in the course of their argument claim that the following statement holds true:*

Let $f_m, f : \Omega \to \mathbb{R}_+$, $m \in \mathbb{N}$, be probability densities on a compact set $\Omega \subset \mathbb{R}^d$ with $f_m, f \geqslant \eta > 0$. If $\mathrm{KL}(f \| f_m) \to 0$ as $m \to \infty$, then $f_m \to f$ in the norm $\| \cdot \|_{L^\infty}$.

This however does not hold as we illustrate by the following simple counterexample: Let $\Omega = [0,1]$ and consider the sequence of probability densities given by

$$f_m(x) = C_m \left(1 \wedge \left(mx + \frac{1}{2} \right) \right), \qquad m \in \mathbb{N},$$

where $C_m = (1 - 1/(8m))^{-1}$ is chosen such that $\int_0^1 f_m(x)\,dx = 1$. Then we have $f_m(x) \to \mathbb{1}_{[0,1]}(x) = f(x)$ pointwise on $(0,1]$ but certainly not uniformly. It is straight-forward to check that $\left\| f_m - \mathbb{1}_{[0,1]} \right\|_{L^2} \to 0$ and since $f_m, f \geqslant 1/2$ on Ω, we have $\mathrm{KL}(f_m \| f) \to 0$ as $m \to \infty$ by Lemma 14.

4 Conclusion

We investigated the approximation capabilities of deep belief networks with two binary hidden layers of sizes m and $m+1$, respectively, and real-valued visible units. We showed that, under minimal regularity requirements on the parental density φ as well as the target density f, these networks are universal approximators in the strong L^q and Kullback-Leibler distances as $m \to \infty$. Moreover, we gave sharp quantitative bounds on the approximation error. We emphasize that the convergence rate in the number of hidden units is independent of the choice of the parental density.

Our results apply to virtually all practically relevant examples thereby theoretically underpinning the tremendous empirical success of DBN architectures we have seen over the last couple of years. As we alluded to in Remark 4, the frequently made choice of a Gaussian parental density does not provide the theoretically optimal DBN approximation of a given target density. Since, in practice, the choice of parental density cannot solely be determined from an approximation standpoint, but also the difficulty of the training of the resulting networks needs to be considered, it is interesting to further empirically study the choice of parental density on both artificial and real-world datasets.

REFERENCES

A. M. Abdel-Zaher and A. M. Eldeib. Breast cancer classification using deep belief networks. *Expert Systems with Applications*, 46:139–144, 2016.

A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

Y. Bengio and O. Delalleau. On the expressive power of deep architectures. In *International conference on algorithmic learning theory*, pp. 18–36. Springer, 2011.

H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011.

M. Burger and A. Neubauer. Error bounds for approximation with neural networks. *Journal of Approximation Theory*, 112:235–250, 10 2001.

R. Cao, D. Bhattacharya, J. Hou, and J. Cheng. Deepqa: improving the estimation of single protein model quality with deep belief networks. *BMC Bioinformatics*, 17(1):1–9, 2016.

T. Chen and H. Chen. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its applications to dynamic systems. *IEEE Transactions on Neural Networks*, pp. 911–917, 1995.

G. V. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2:303–314, 1989.

M. J. Donahue, C. Darken, L. Gurvits, and E. Sontag. Rates of Convex Approximation in Non-Hilbert Spaces. *Constructive Approximation*, 13(2):187–220, 1997.

A. Fischer and C. Igel. Training restricted Boltzmann machines: An introduction. *Pattern Recognition*, 47(1):25–39, 2014.

A. Fischer and Ch. Igel. An introduction to restricted Boltzmann machines. In *Iberoamerican Congress on Pattern Recognition*, pp. 14–36. Springer, 2012.

G. B. Folland. *Real analysis*. Pure and Applied Mathematics (New York). John Wiley & Sons, Inc., New York, second edition, 1999.

B. Ghojogh, A. Ghodsi, F. Karray, and M. Crowley. Restricted Boltzmann machine and deep belief network: Tutorial and survey. *arXiv preprint arXiv:2107.12521*, 2021.

U. Haagerup. The best constants in the Khintchine inequality. *Studia Math.*, 70(3):231–283, 1981.

G. E. Hinton. Deep belief networks. *Scholarpedia*, 4(5):5947, 2009.

G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313:504–507, 2006.

G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Y. Huang, A. Panahi, H. Krim, Y. Yu, and S. L. Smith. Deep adversarial belief networks. *arXiv preprint arXiv:1909.06134*, 2019.

M. Jiang, Y. Liang, X. Feng, X. Fan, Z. Pei, Y. Xue, and R. Guan. Text classification based on deep belief network and softmax regression. *Neural Computing and Applications*, 29(1):61–70, 2018.

J. Johnson. Deep, skinny neural networks are not universal approximators. In *International Conference on Learning Representations*, 2018.

S. Kamada and T. Ichimura. An adaptive learning method of deep belief network by layer generation algorithm. In *2016 IEEE Region 10 Conference (TENCON)*, pp. 2967–2970. IEEE, 2016.

S. Kamada and T. Ichimura. An object detection by using adaptive structural learning of deep belief network. In *2019 International joint conference on neural networks (IJCNN)*, pp. 1–8. IEEE, 2019.

O. Krause, A. Fischer, T. Glasmachers, and C. Igel. Approximation properties of dbns with binary hidden units and real-valued visible units. In *International Conference on Machine Learning*, pp. 419–426, 2013.

N. Le Roux and Y. Bengio. Representational power of restricted Boltzmann machines and deep belief networks. *Neural Computation*, 20(6):1631–1649, 2008.

N. Le Roux and Y. Bengio. Deep belief networks are compact universal approximators. *Neural computation*, 22:2192–2207, 2010.

M. Ledoux and M. Talagrand. *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3)*. Springer-Verlag, Berlin, 1991.

M Leshno, V. Y. Lin, A. Pinkus, and S. Schocken. Multilayer feedforward networks with a non-polynomial activation function can approximate any function. *Neural networks*, 6(6):861–867, 1993.

J. Q. Li and A. R. Barron. Mixture density estimation. In *NIPS*, volume 12, pp. 279–285, 1999.

M. Liang, Z. Li, T. Chen, and J. Zeng. Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. *IEEE/ACM transactions on computational biology and bioinformatics*, 12(4):928–937, 2014.

P. Luo, Y. Li, L.-P. Tian, and F.-X. Wu. Enhancing the prediction of disease–gene associations with multimodal deep learning. *Bioinformatics*, 35(19):3735–3742, 2019.

T. Merkh and G. Montúfar. Stochastic feedforward neural networks: Universal approximation. *arXiv preprint arXiv:1910.09763*, 2019.

G. Montúfar. Deep narrow Boltzmann machines are universal approximators. 2014.

G. Montúfar. Universal approximation of markov kernels by shallow stochastic feedforward networks. *arXiv preprint arXiv:1503.07211*, 2015.

G. Montúfar. Restricted Boltzmann machines: Introduction and review. In *Information Geometry and Its Applications IV*, pp. 75–115. Springer, 2016.

G. Montúfar and N. Ay. Refinements of universal approximation results for deep belief networks and restricted Boltzmann machines. *Neural computation*, 23(5):1306–1319, 2011.

G. Montúfar and J. Morton. Discrete restricted Boltzmann machines. *J. Mach. Learn. Res.*, 16(1): 653–672, 2015.

R. Pascanu, G. Montúfar, and Y. Bengio. On the number of response regions of deep feed forward networks with piece-wise linear activations. In *International Conference on Learning Representations*, 2014.

F. Shen, J. Chao, and J. Zhao. Forecasting exchange rate using deep belief networks and conjugate gradient method. *Neurocomputing*, 167:243–253, 2015.

P. Smolensky. Information processing in dynamical systems: foundations of harmony theory. In *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations*, pp. 194–281. 1986.

I. Sutskever and G. E. Hinton. Deep, narrow sigmoid belief networks are universal approximators. *Neural computation*, 20(11):2629–2636, 2008.

Y. Wang and J. Zeng. Predicting drug-target interactions using restricted Boltzmann machines. *Bioinformatics*, 29(13):i126–i134, 2013.

W. H. Young. On the multiplication of successions of Fourier constants. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 87 (596):331–339, 1912.

A. J. Zeevi and R. Meir. Density estimation through convex combinations of densities: Approximation and estimation bounds. *Neural Networks*, 10(1):99–109, 1997.