Quantitative Universal Approximation Bounds for Deep Belief Networks

Julian Sieber [1 2] **Johann Gehringer** [2]

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.

1. Introduction

Deep belief networks (DBNs) are a class of generative probabilistic models obtained by stacking 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 these models 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, let alone a quantitative understanding of the approximation error in terms of the number of hidden neurons. As a measure of proximity between two real-valued probability density functions, one typically considers the L^q-distance or the Kullback-Leibler divergence.

[1]Zalando Ireland Limited, 2WML, Windmill Quarter, Dublin 2, D02 F206, Ireland [2]Department of Mathematics, Imperial College London, London SW7 2AZ, United Kingdom. Correspondence to: Julian Sieber <julian.sieber@zalando.ie>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

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. We also present similar estimates for deep narrow networks. 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 $\varphi \in L^q(\mathbb{R}^d)$ and $f \in W^{1,q}(\mathbb{R}^d)$, where

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

  and

$$W^{1,q}(\mathbb{R}^d) = \big\{ f \in L^q(\mathbb{R}^d) : f \text{ weakly differentiable and } \|\nabla f\|_{L^q} < \infty \big\}.$$

  In addition, we prove that the error admits a bound of order $\mathcal{O}\big(m^{-\left(1-\frac{1}{\min(q,2)}\right)\frac{q}{d(q-1)+q}}\big)$ for each $q \in (1, \infty)$, where m is the number of hidden neurons. A similar estimate is shown for deep narrow networks. In particular, we observe that the expected curse of dimensionality effect gets exacerbated due to a *curse of moments* effect for high values of q.

- 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), where

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

- 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 $\mathcal{O}\big(m^{-\frac{2}{d+2}}\big)$.

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 reducing the required number of layers to $\lceil \frac{2^{d-1}}{d-\log(d)} \rceil$ with d hidden units each, see (Le Roux & Bengio, 2010; Montúfar & Ay, 2011). 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 multilayer 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 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.

Since our methodology involves an approximation by a convex combination of probability densities, we refer the reader to the related works of (Nguyen & McLachlan, 2019; Nguyen et al., 2020) and the references therein for an overview of the wide range of universal approximation results in the context of mixture models. See also (Everitt & Hand, 1981; Titterington et al., 1985; McLachlan & Basford, 1988; McLachlan & Peel, 2000; Robert & Mengersen, 2011; Celeux, 2019) for in-depth treatments of mixture models.

The recent articles (Bailey & Telgarsky, 2018; Perekrestenko et al., 2020) in the context of generative networks show that deep neural networks can transform a one-dimensional uniform distribution to approximate any two-dimensional Lipschitz continuous target density.

Another strand of research related to the questions of this article are works on quantile (or distribution) regression, see (Koenker, 2005) as well as (Dabney et al., 2018; Tagasovska & Lopez-Paz, 2019; Fakoor et al., 2021) for recent methods involving neural networks.

2. 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 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

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

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*. For $m,n\in\mathbb{N}$ we shall denote the set of binary-binary RBMs with these layer sizes by

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

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

Proposition 2.1 ((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

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

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 $p(\cdot \,|\, h_1) \in \mathcal{V}_\varphi^\sigma$ for each $h_1 \in H_1$.

Example 2.2. 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

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

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

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

$$\begin{aligned}\mathsf{DBN}_\varphi(d,m,n) = \Big\{&p \text{ is a DBN with parental density} \\ &\varphi,\ d \text{ visible units},\ m \text{ hidden units on the} \\ &\text{first level, and } n \text{ hidden units on the second level}\Big\},\end{aligned} \tag{7}$$

where $\varphi \in \mathcal{D}(\mathbb{R}^d)$ and $d,m,n \in \mathbb{N}$. We shall not distinguish between the whole DBN and the marginal density of the visible nodes, which is the object we are ultimately interested in, that is, we write

$$p(v) = \sum_{h_1\in H_1}\sum_{h_2\in H_2} p(v,h_1,h_2). \tag{8}$$

In case $p \in \mathsf{DBN}_\varphi(d,m,n)$ with $\varphi \in L^q(\mathbb{R}^{|V|})$ then also the marginal (8) belongs to $L^q(\mathbb{R}^{|V|})$.

After their introduction in (Hinton & Salakhutdinov, 2006), deep belief networks rose to prominence due to a training algorithm developed in (Hinton et al., 2006) which addressed the vanishing gradient problem by pre-training deep networks. Instead of naïvely stacking two RBMs the authors considered several such stacked layers and greedily pre-trained the weights over the layers on a contrastive divergence loss. To be more precise, first, the visible and the first hidden layer are considered as a classical RBM and the weights of the first hidden layer are learnt. In the next step, the weights of the second hidden layer are learnt based on the first hidden layer using Gibbs sampling. This procedure repeats iteratively until all hidden layers are trained. For more details we refer to (Fischer & Igel, 2014; Ghojogh et al., 2021).

3. Main Results

To state the results of this article, we need to introduce two bits of additional notation: Let $q \in [1,\infty]$. We declare $\mathcal{D}_q(\mathbb{R}^d) = \mathcal{D}(\mathbb{R}^d) \cap L^q(\mathbb{R}^d)$. Finally, for $q \in [1,\infty)$, let us abbreviate the constant

$$\begin{aligned}\Upsilon_q &= \max\left(1, \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} |x|^q e^{-\frac{x^2}{2}}\,dx\right)^{\frac{1}{q}} \\ &= \begin{cases} 1, & q \leqslant 2, \\ \dfrac{\sqrt{2}}{\pi^{\frac{1}{2q}}}\Gamma\left(\dfrac{q+1}{2}\right)^{\frac{1}{q}}, & q > 2, \end{cases}\end{aligned} \tag{9}$$

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 3.1. *Let $q \in [1,\infty)$ and $f, \varphi \in \mathcal{D}_q(\mathbb{R}^d)$. Suppose that $f \in W^{1,q}(\mathbb{R}^d)$ and $\mathcal{M}_\varphi = \int_{\mathbb{R}^d} |x|\varphi(x)\,dx < \infty$. Then, for each $m \in \mathbb{N}$, the following quantitative bound holds:*

$$\begin{aligned}&\inf_{p\in\mathsf{DBN}_\varphi(d,m,m+1)} \|f - p\|_{L^q} \\ &\leqslant \left(d - \frac{d}{q} + 1\right)\left(\frac{\mathcal{M}_\varphi \|\nabla f\|_{L^q}}{d\big(1-\frac{1}{q}\big)}\right)^{\frac{d(q-1)}{d(q-1)+q}} \\ &\quad\times \left(\frac{2\Upsilon_q\|\varphi\|_{L^q}}{m^{1-\frac{1}{\min(q,2)}}}\right)^{\frac{1}{d-\frac{d}{q}+1}},\end{aligned} \tag{10}$$

where the constant Υ_q is defined in (9).

While this bound becomes trivial if $q = 1$, the following qualitative approximation result still holds in that case: 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

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

Remark 3.2. Returning to Example 2.2, 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 (10) 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 3.5 below. For a detailed review of the exponential family's properties we refer to (Brown, 1986).

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

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

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 3.3. *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 3.4. The uniform continuity requirement on f in Theorem 3.3 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 3.1 and 3.3 we can deduce the following quantitative approximation bound in the Kullback-Leibler divergence:

Corollary 3.5. *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*

$$\begin{aligned} &\inf_{p \in \mathsf{DBN}_\varphi(d,m,m+1)} \mathrm{KL}(f\|p) \\ &\leqslant \frac{M^{\frac{2}{d+2}}}{\eta m^{\frac{2}{d+2}}} \Bigg(2(d+2)^2 \left(\frac{\mathcal{M}_\varphi \|\nabla f\|_{L^2}}{d}\right)^{\frac{2d}{d+2}} \\ &\qquad \times \left(\Upsilon_2 \|\varphi\|_{L^2}\right)^{\frac{4}{d+2}} + \|f - \varphi\|^2_{L^2(\Omega)} \Bigg), \end{aligned} \tag{11}$$

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

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 (11) is actually finite. This follows from the *interpolation inequality*

$$\|\varphi\|_{L^2} \leqslant \sqrt{\|\varphi\|_{L^1}\|\varphi\|_{L^\infty}} = \sqrt{\|\varphi\|_{L^\infty}}, \tag{12}$$

see (Brezis, 2011), Exercise 4.4.

Corollary 3.5 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 4.9 below for details.

Remark 3.6. The results of Theorems 3.1 and 3.3 and Corollary 3.5 also hold for narrow deep belief networks with multiple hidden layers. In fact, our proofs below only use that the second hidden layer of size $m+1$ can approximate a probability distribution on $\{1, 2, \ldots, m\}$ by appealing to Proposition 2.1. Applying the results of (Montúfar & Ay, 2011), this transfers to narrow multi-layer deep belief networks as we detail in Section 5.

4. Proofs

This section presents the proofs of Theorems 3.1, 3.3 and Corollary 3.5. As a first step, we shall establish a couple of preliminary results in the next two subsections.

4.1. 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

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

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

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

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 4.1. *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 $(\alpha_1, \dots, \alpha_m) \in \triangle_m$ and $(\mu_1, \dots, \mu_m) \in \big(\mathbb{R}^d\big)^m$ 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 i^{th} unit vector.

Let us fix $q \in [1,\infty]$ and $\sigma > 0$. By Proposition 2.1 there is a $\pi \in \mathsf{B\text{-}RBM}(m, m+1)$ such that for every , $h_1 \in \{0,1\}^m$

$$\left| \tilde{\alpha}(h_1) - \sum_{h_2 \in \{0,1\}^{m+1}} \pi(h_1, h_2) \right| \leqslant \frac{\varepsilon}{m\sigma^{-d+\frac{d}{p}} \|\varphi\|_{L^q}}. \tag{13}$$

We set

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

and

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

This is the desired approximation since

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

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

4.2. 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 $\varphi_\sigma = \varphi_{0,\sigma}$ 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 that $f \star \varphi_\sigma \in L^q(\mathbb{R}^d)$. In addition, the following approximation result holds, see Appendix A.1 for the proof:

Proposition 4.2. *Let $\varphi \in \mathcal{D}(\mathbb{R}^d)$. Then all of the following hold true:*

1. *For each $q \in [1,\infty)$ and each $f \in W^{1,q}(\mathbb{R}^d)$, we have*

$$\|f - \varphi_\sigma \star f\|_{L^q} \leqslant \mathcal{M}_\varphi \|\nabla f\|_{L^q} \sigma \quad \forall \sigma > 0.$$

2. *If $f \in L^\infty(\mathbb{R}^d) \cap \mathcal{C}_u(\mathbb{R}^d)$, then*

$$\lim_{\sigma \downarrow 0} \|f - f \star \varphi_\sigma\|_{L^\infty} = 0.$$

4.3. Approximation Theory in Banach Spaces

The second ingredient needed in the proof of Theorem 3.1 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 $f_1, \dots, f_k \in \mathcal{X}$,

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

holds, where $\epsilon_1, \dots, \epsilon_k$ 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 4.3. 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 be important below:

Proposition 4.4 ((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 $f \in \overline{\operatorname{conv}}(A)$. 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 $h \in \operatorname{conv}_m(A)$ satisfying*

$$\|f - h\|_\mathcal{X} \leqslant \frac{C\xi}{m^{1-\frac{1}{\mathfrak{t}}}}. \tag{14}$$

Notice that the bound (14) is of course trivial for $\mathfrak{t} = 1$. Moreover, in Appendix A.2 we provide an example which shows that the convergence rate $m^{\frac{1}{\mathfrak{t}}-1}$ is optimal.

Corollary 4.5. *Let $A \subset L^q(\mathbb{R}^d)$, $1 \leqslant q < \infty$, and suppose that $f \in \overline{\operatorname{conv}}(A)$. If $\xi = \sup_{g\in A} \|f - g\|_\mathcal{X} < \infty$, then for all $m \in \mathbb{N}$, there is a $h \in \operatorname{conv}_m(A)$ 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 (9)*.*

Proof. Owing to Example 4.3, we are in the regime of Proposition 4.4. The sharp constant $C = \Upsilon_q$ was derived in (Haagerup, 1981). □

4.4. Proof of Theorems 3.1 and 3.3

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

1. By Proposition 4.2 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 $f \star \varphi_\sigma \in \overline{\operatorname{conv}}(\mathcal{V}_\varphi^\sigma)$, Proposition 4.5 allows us to show that for each $\varepsilon > 0$ and each $m \in \mathbb{N}$, we can pick $\sigma > 0$ such that
$$\inf_{g\in\operatorname{conv}_m(\mathcal{V}_\varphi^\sigma)} \|f - g\|_{L^q} \leqslant \varepsilon + \frac{2\Upsilon_q \|\varphi\|_{L^q}}{m^{1-\frac{1}{\min(q,2)}}}.$$
3. Finally, we employ Lemma 4.1 to conclude the desired estimate (10).

Lemma 4.6. *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*

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

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 $g \notin \overline{\operatorname{conv}}(\mathcal{V}_\varphi^\sigma)$. As a consequence of the Hahn-Banach theorem, g is separated from $\overline{\operatorname{conv}}(\mathcal{V}_\varphi^\sigma)$ 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 $h \in \overline{\operatorname{conv}}(\mathcal{V}_\varphi^\sigma)$, see (Brezis, 2011), Theorem 1.7. On the other hand, we however have

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

which is the desired contradiction. □

We also need the following estimate, which can be shown by another application of the Hahn-Banach theorem:

Lemma 4.7. *Let $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. Then for any measure ν on $\mathbb{R}^d$ it holds that*

$$\left\|\int_{\mathbb{R}^d} g(\cdot, y)\, d\nu(y)\right\|_{L^q} \leqslant \int_{\mathbb{R}^d} \big\|g(\cdot, y)\big\|_{L^q}\, d\nu(y), \tag{15}$$

where we understand that the L^q-norm is taken with respect to the first argument of g.

We can now establish the main results of this article:

Proof of Theorems 3.1 and 3.3. Let us first assume that $q \in (1, \infty)$ and prove the quantitative bound (10). To this end fix $\varepsilon > 0$ and $m \in \mathbb{N}$. We first observe that, by Proposition 4.2, we have

$$\big\|f - f \star \varphi_\sigma\big\|_{L^q} \leqslant \mathcal{M}_\varphi \|\nabla f\|_{L^q}\sigma.$$

Employing Lemma 4.6 and Corollary 4.5 with $A = \mathcal{V}_\varphi^\sigma$, we can find a $g_m \in \operatorname{conv}_m(\mathcal{V}_\varphi^\sigma)$ such that

$$\begin{aligned}\|f - g_m\|_{L^q} &\leqslant \big\|f - f \star \varphi_\sigma\big\|_{L^q} + \big\|f \star \varphi_\sigma - g_m\big\|_{L^q}\\ &\leqslant \mathcal{M}_\varphi \|\nabla f\|_{L^q}\sigma\\ &\quad + \frac{\Upsilon_q}{m^{1-\frac{1}{\min(q,2)}}} \sup_{\mu\in\mathbb{R}^d} \big\|f \star \varphi_\sigma - \varphi_{\mu,\sigma}\big\|_{L^q}.\end{aligned} \tag{16}$$

For the last term we use Lemma 4.7 to estimate

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

where the final step follows by a change of variables. Inserting this back into (16), we arrive at

$$\|f-g_m\|_{L^q} \leqslant \mathcal{M}_\varphi \|\nabla f\|_{L^q}\sigma + \frac{2\Upsilon_q\|\varphi\|_{L^q}}{\sigma^{d(1-\frac{1}{q})}m^{1-\frac{1}{\min(q,2)}}}.$$

The right-hand side is minimized for

$$\sigma_\star = \left(\frac{2\big(d-\frac{d}{q}\big)\Upsilon_q\|\varphi\|_{L^q}}{\mathcal{M}_\varphi\|\nabla f\|_{L^q}m^{1-\frac{1}{\min(q,2)}}}\right)^{\frac{q}{d(q-1)+q}}, \qquad (17)$$

which leads to the bound

$$\|f-g_m\|_{L^q} \leqslant \left(d-\frac{d}{q}+1\right)\left(\frac{\mathcal{M}_\varphi\|\nabla f\|_{L^q}}{d\big(1-\frac{1}{q}\big)}\right)^{\frac{d(q-1)}{d(q-1)+q}} \times \left(\frac{2\Upsilon_q\|\varphi\|_{L^q}}{m^{1-\frac{1}{\min(q,2)}}}\right)^{\frac{1}{d-\frac{d}{q}+1}}. \qquad (18)$$

We present the details of this computation in Appendix A.3. Finally, Lemma 4.1 allows us to choose $p \in \mathsf{DBN}_\varphi(d,m,m+1)$ such that $\|g_m - p\|_{L^q} \leqslant \varepsilon$. Therefore, we conclude

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

Since $\varepsilon > 0$ was arbitrary, the bound (10) follows.

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

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

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 $g_m \in \mathrm{conv}_m(\mathcal{V}_\varphi^\sigma)$ 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 4.2 to find a $\sigma > 0$ for which $\big\|f - f\star\varphi_\sigma\big\|_{L^q} \leqslant \frac{\varepsilon}{3}$. Finally employing Lemma 4.1 as above, there is a $p \in \mathsf{DBN}_\varphi(d,m,m+1)$ such that

$$\begin{aligned}&\|f-p\|_{L^q}\\ &\leqslant \big\|f - f\star\varphi_\sigma\big\|_{L^q} + \big\|f\star\varphi_\sigma - g_m\big\|_{L^q} + \big\|g_m - p\big\|_{L^q}\\ &\leqslant \varepsilon.\end{aligned}$$

□

4.5. Kullback-Leibler Approximation on Compacts

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

Lemma 4.8 ((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*

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

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

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

which is the required estimate. □

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

Proof of Corollary 3.5. Extending the target density f by zero on $\mathbb{R}^d \setminus \Omega$, the corollary follows from Theorem 3.1 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 3.3 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 4.8 we obtain

$$\begin{aligned}&\inf_{p\in\mathsf{DBN}_\varphi(d,m,m+1)} \mathrm{KL}(f\|p)\\ &\leqslant 2(d+2)^2\left(\frac{\mathcal{M}_\varphi\|\nabla f\|_{L^2}}{d}\right)^{\frac{2d}{d+2}}\big(\Upsilon_2\|\varphi\|_{L^2}\big)^{\frac{4}{d+2}}m^{-\frac{2}{d+2}}\end{aligned} \qquad (19)$$

for all $m \geqslant M$. An 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 $p(v\,|\,h_1) = \varphi$ for each $h_1 \in \{0,1\}^m$ in (3). Hence, the visible units of the DBN have density φ. Applying Lemma 4.8 once more gives

$$\inf_{p\in\mathsf{DBN}_\varphi(d,m,m+1)} \mathrm{KL}(f\|p) \leqslant \mathrm{KL}(f\|\varphi) \leqslant \frac{1}{\eta}\|f-\varphi\|^2_{L^2(\Omega)} \qquad (20)$$

for all $m = 1,\dots,M-1$. Finally, combining (19) and (20) we get the required estimate:

$$\begin{aligned}&\inf_{p\in\mathsf{DBN}_\varphi(d,m,m+1)} \mathrm{KL}(f\|p)\\ &\leqslant \frac{M^{\frac{2}{d+2}}}{\eta m^{\frac{2}{d+2}}}\Bigg(2(d+2)^2\left(\frac{\mathcal{M}_\varphi\|\nabla f\|_{L^2}}{d}\right)^{\frac{2d}{d+2}}\big(\Upsilon_2\|\varphi\|_{L^2}\big)^{\frac{4}{d+2}}\\ &\qquad + \|f-\varphi\|^2_{L^2(\Omega)}\Bigg).\end{aligned}$$

This concludes the proof. □

Remark 4.9. Our strategy of the proof of the Kullback-Leibler approximation in Corollary 3.5 through Lemma 4.8 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 a simple counterexample in Appendix A.4.

5. Extension to Narrow DBNs With Multiple Hidden Layers

As we have noted in Remark 3.6, the results of Theorems 3.1 and 3.3 and Corollary 3.5 naturally extend to narrow DBNs with multiple hidden layers. More specifically, a DBN with $L \geqslant 2$ hidden layers is defined by generalizing (3) to a graphical model with probability distribution

$$\begin{aligned} &p(v, h_1, \dots, h_L) \\ =&p(v \,|\, h_1)p(h_2 \,|\, h_3) \cdots p(h_{L-2} \,|\, h_{L-1})p(h_{L-1}, h_L), \end{aligned} \tag{21}$$

where $p(h_{L-1}, h_L)$ is a binary-binary RBM.

For these models there exists a discrete approximation result similar to Proposition 2.1 proven by (Montúfar & Ay, 2011) building on ideas of (Le Roux & Bengio, 2010):

Proposition 5.1 ((Montúfar & Ay, 2011), Theorem 3)**.** *Let $m \in \mathbb{N}$ and μ be a probability distribution on $\{0,1\}^m$. Let $b = \min\{k \in \mathbb{N} : m \leqslant 2^{k-1} + k\}$. Then, for each $\varepsilon > 0$, there is a DBN p with m binary visible units and $L = \left\lceil \frac{2^{m-1}}{m-b} \right\rceil$ binary hidden layers with m units each such that*

$$\left| \mu(v) - \sum_{h_1, \dots, h_L \in \{0,1\}^m} p(v, h_1, \dots, h_L) \right| \leqslant \varepsilon$$

for all $v \in \{0,1\}^m$.

This proposition allows us to extend the central discrete approximation result of Lemma 4.1 to multi-layer DBNs. To state this result it is convenient to introduce the mapping $\mathfrak{b} : \mathbb{N} \to \mathbb{N}$,

$$\mathfrak{b}(m) = \left\lceil \frac{2^{m-1}}{m - \min\left\{k \in \mathbb{N} : \lceil \log_2(m) \rceil \leqslant 2^{k-1} + k\right\}} \right\rceil.$$

Notice that $\mathfrak{b}$ grows exponentially as $m \to \infty$.

Henceforth, we shall assume that there is a parental density φ and a $\sigma > 0$ such that we have $p(\cdot \,|\, h_1) \in \mathcal{V}_\varphi^\sigma$ for all $h_1 \in H_1$ in the defining identity (21). Furthermore, we restrict to multi-layer DBNs whose hidden layers all have the same dimension m. The set of DBNs with d visible real-valued units, parental density $\varphi \in \mathcal{D}(\mathbb{R}^d)$, L hidden layers of size m each is denoted by $\mathsf{DBN}_\varphi^L(d, m)$. Notice that $\mathsf{DBN}_\varphi^2(d, m) = \mathsf{DBN}_\varphi(d, m, m)$ for the previously used notation introduced in (7).

Lemma 5.2. *Let $q \in [1, \infty]$, $\varphi \in \mathcal{D}_q(\mathbb{R}^d)$, $\sigma > 0$, and $m \in \mathbb{N}$. Then, for every $f \in \mathrm{conv}_m(\mathcal{V}_\varphi^\sigma)$ and every $\varepsilon > 0$, there is a deep belief network $p \in \mathsf{DBN}_\varphi^{\mathfrak{b}(m)}\big(d, \lceil \log_2(m) \rceil\big)$ such that*

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

Proof. Recall that we can write

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

for some $(\alpha_1, \dots, \alpha_m) \in \triangle_m$ and $(\mu_1, \dots, \mu_m) \in \big(\mathbb{R}^d\big)^m$. By fixing an injective mapping $\{1, \dots, m\} \to \{0,1\}^{\lceil \log_2(m) \rceil}$, we can interpret $(\alpha_1, \dots, \alpha_m)$ as a probability distribution on $\{0,1\}^{\lceil \log_2(m) \rceil}$ and the proof is completed similarly to Lemma 4.1. □

Finally, the remaining parts of the proofs of Theorems 3.1 and 3.3 and Corollary 3.5 hold *mutatis mutandis* for narrow multi-layer DBNs leading to the following result:

Corollary 5.3. *The statements of Theorems 3.1 and 3.3 and Corollary 3.5 hold for $p \in \mathsf{DBN}_\varphi(d, m, m+1)$ replaced by*

$$p \in \mathsf{DBN}_\varphi^{\mathfrak{b}(m)}\big(d, \lceil \log_2(m) \rceil\big).$$

6. 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. These bounds were also extended to narrow DBNs with multiple hidden layers whose width only grows logarithmically in the approximation parameter m at the price of an exponentially growing number of them.

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 3.2, 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.

Acknowledgements

JS acknowledges support by the EPSRC Centre for Doctoral Training in Mathematics of Random Systems: Analysis, Modelling and Simulation (EP/S023925/1).

References

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

B. Bailey and M. J. Telgarsky. Size-noise tradeoffs in generative networks. *Advances in Neural Information Processing Systems*, 31, 2018.

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

A. R. Barron. Approximation and estimation bounds for artificial neural networks. *Machine learning*, 14(1):115–133, 1994.

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.

L. D. Brown. Fundamentals of statistical exponential families: with applications in statistical decision theory. IMS, 1986.

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.

G. Celeux. EM methods for finite mixtures. In *Handbook of mixture analysis*, Chapman & Hall/CRC Handb. Mod. Stat. Methods, pp. 21–39. CRC Press, Boca Raton, FL, 2019.

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.

W. Dabney, M. Rowland, M. Bellemare, and R. Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

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.

B. S. Everitt and D. J. Hand. *Finite mixture distributions*. Monographs on Applied Probability and Statistics. Chapman & Hall, London-New York, 1981. ISBN 0-412-22420-8.

R. Fakoor, T. Kim, J. Mueller, A. J. Smola, and R. J. Tibshirani. Flexible model aggregation for quantile regression. *arXiv preprint arXiv:2103.00083*, 2021.

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

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

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.

L. K. Jones. A simple lemma on greedy approximation in hilbert space and convergence rates for projection pursuit regression and neural network training. *The Annals of Statistics*, pp. 608–613, 1992.

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.

R. Koenker. *Quantile regression*, volume 38 of *Econometric Society Monographs*. Cambridge University Press, Cambridge, 2005.

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 nonpolynomial 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.

G. McLachlan and K. E. Basford. *Mixture models*, volume 84 of *Statistics: Textbooks and Monographs*. Marcel Dekker, Inc., New York, 1988.

G. McLachlan and D. Peel. *Finite mixture models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York, 2000.

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.

H. D. Nguyen and G. McLachlan. On approximations via convolution-defined mixture models. *Communications in Statistics - Theory and Methods*, 48(16):3945–3955, 2019.

T. T. Nguyen, H. D. Nguyen, F. Chamroukhi, and G. J. McLachlan. Approximation by finite mixtures of continuous density functions that vanish at infinity. *Cogent Mathematics & Statistics*, 7(1):1750861, 2020.

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.

D. Perekrestenko, S. Müller, and H. Bölcskei. Constructive universal high-dimensional distribution generation through deep relu networks. In *International Conference on Machine Learning*, pp. 7610–7619. PMLR, 2020.

C. P. Robert and K. L. Mengersen. Exact Bayesian analysis of mixtures. In *Mixtures: estimation and applications*, Wiley Ser. Probab. Stat., pp. 241–254. Wiley, Chichester, 2011.

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.

N. Tagasovska and D. Lopez-Paz. Single-model uncertainties for deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical analysis of finite mixture distributions*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 1985. ISBN 0-471-90763-4.

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.

A. Details of the Mathematical Results

This appendix provides further details of the mathematical results used in the main text. More specifically, we provide

1. the proof of Proposition 4.2,
2. a detailed proof of Proposition 4.4 for Hilbert spaces as well as an example showing that its approximation rate is optimal in general,
3. the computational details for equations (17) and (18), and
4. the construction of an explicit counterexample to the statement discussed in Remark 4.9.

A.1. Proof of Proposition 4.2

Proof. 1. Observe that

$$\begin{aligned} f(x) - \big(\varphi_\sigma \star f\big)(x) &= \int_{\mathbb{R}^d} \varphi_\sigma(y)\big(f(x) - f(x-y)\big)\,dy \\ &= -\int_{\mathbb{R}^d} \varphi_\sigma(y) \left(\int_0^1 \frac{d}{dt} f(x-ty)\,dt\right) dy \\ &= \int_{\mathbb{R}^d} \varphi_\sigma(y) \left(\int_0^1 \nabla f(x-ty)\cdot y\,dt\right) dy \end{aligned} \tag{22}$$

for all $x \in \mathbb{R}^d$. Applying Lemma 4.7 first with the measure $d\nu_\sigma(y) = \varphi_\sigma(y)\,dy$ and then with the standard Lebesgue in the identity (22), we get

$$\begin{aligned} \big\|f - \varphi_\sigma \star f\big\|_{L^q} &= \left\| \int_{\mathbb{R}^d} \varphi_\sigma(y) \left(\int_0^1 \nabla f(\cdot - ty)\cdot y\,dt\right) dy \right\|_{L^q} \\ &\leqslant \int_{\mathbb{R}^d} \varphi_\sigma(y) \left\| \int_0^1 \nabla f(\cdot - ty)\cdot y\,dt \right\|_{L^q} dy \\ &\leqslant \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \varphi_\sigma(y) \left(\int_0^1 \big\|\nabla f(\cdot - ty)\cdot y\big\|_{L^q}\,dt\right) dy\,dx \\ &\leqslant \int_{\mathbb{R}^d} |y|\varphi_\sigma(y) \left(\int_0^1 \big\|\nabla f(\cdot - ty)\big\|_{L^q}\,dt\right) dy \\ &= \mathcal{M}_\varphi \|\nabla f\|_{L^q}\sigma, \end{aligned}$$

as required.

2. Fix $\varepsilon > 0$. By uniform continuity of f, we can find a $\delta > 0$ such that

$$\sup_{|\mu|\leqslant\delta} \big|f(x) - f(x-\mu)\big| \leqslant \varepsilon \qquad \forall\, x \in \mathbb{R}^d. \tag{23}$$

In particular, we obtain

$$\begin{aligned} \Big|f(x) - \big(f \star \varphi_\sigma\big)(x)\Big| &\leqslant \int_{\mathbb{R}^d} \varphi_\sigma(\mu)\big|f(x) - f(x-\mu)\big|\,d\mu \\ &\leqslant \int_{\{|\mu|>\delta\}} \varphi_\sigma(\mu)\big|f(x) - f(x-\mu)\big|\,d\mu + \int_{\{|\mu|\leqslant\delta\}} \varphi_\sigma(\mu)\big|f(x) - f(x-\mu)\big|\,d\mu \\ &\leqslant 2\|f\|_{L^\infty} \int_{\{|\mu|>\delta\}} \varphi_\sigma(\mu)\,d\mu + \varepsilon, \end{aligned} \tag{24}$$

where we applied the uniform continuity estimate (23) to the second integral. By substituting $\bar{\mu} = \mu/\sigma$ in the first integral, we get

$$\int_{\{|\mu|>\delta\}} \varphi_\sigma(\mu)\,d\mu = \int_{\{|\bar{\mu}|>\frac{\delta}{\sigma}\}} \varphi(\bar{\mu})\,d\bar{\mu} \xrightarrow{\sigma\downarrow 0} 0,$$

so that (24) implies

$$\lim_{\sigma\downarrow 0}\big\|f - f\star\varphi_\sigma\big\|_{L^\infty} = \lim_{\sigma\downarrow 0}\sup_{x\in\mathbb{R}^d}\Big|f(x) - \big(f\star\varphi_\sigma\big)(x)\Big| \leqslant 2\|f\|_{L^\infty}\lim_{\sigma\downarrow 0}\int_{\{|\mu|>\delta\}}\varphi_\sigma(\mu)\,d\mu + \varepsilon = \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, the proof is complete. □

A.2. Details on Proposition 4.4

While the proof of Proposition 4.4 for a general Banach space is rather technical, we find it instructive to present the simplified argument for a *Hilbert* space. Our proof is inspired by (Jones, 1992), see also (Barron, 1994).

Proposition A.1. *Let $\big(\mathcal{X}, \|\cdot\|_{\mathcal{X}}\big)$ be a Hilbert space. Let $A \subset \mathcal{X}$ and $f \in \overline{\operatorname{conv}}(A)$. Suppose that $\xi = \sup_{g\in A}\|f-g\|_{\mathcal{X}} < \infty$. Then, for each $m \in \mathbb{N}$, we can find an element $g \in \operatorname{conv}_m(A)$ satisfying*

$$\|f-g\|_{\mathcal{X}} \leqslant \frac{\xi}{\sqrt{m}}. \tag{25}$$

Proof. We proceed by induction on $m \in \mathbb{N}$. The base $m = 1$ is trivial, so we can assume that the statement holds for $m \geqslant 1$. Let us declare

$$\Xi_{m+1} = \inf_{g\in\operatorname{conv}_{m+1}(A)}\|f-g\|_{\mathcal{X}}.$$

By the induction hypothesis, we may assume that $\Xi_m \leqslant \frac{\xi}{\sqrt{m}}$ and we can find $h \in \operatorname{conv}_m(A)$ attaining this bound. Consequently, we get

$$\begin{aligned}
&\Xi_{m+1}^2\\
&\leqslant \inf_{\substack{\lambda\in[0,1]\\ g\in A}}\Big\|\lambda(f-g)+(1-\lambda)(f-h)\Big\|_{\mathcal{X}}^2\\
&= \inf_{\substack{\lambda\in[0,1]\\ g\in A}}\Big[\lambda^2\|f-g\|_{\mathcal{X}}^2 + 2\lambda(1-\lambda)\,\langle f-g, f-h\rangle_{\mathcal{X}}\\
&\quad + (1-\lambda)^2\|f-h\|_{\mathcal{X}}^2\Big]\\
&\leqslant \inf_{\lambda\in[0,1]}\Big[\lambda^2\xi^2 + 2\lambda(1-\lambda)\inf_{g\in A}\langle f-g, f-h\rangle_{\mathcal{X}}\\
&\quad + (1-\lambda)^2\Xi_m^2\Big].
\end{aligned} \tag{26}$$

We claim that

$$\inf_{g\in A}\langle f-g, f-h\rangle_{\mathcal{X}} = 0. \tag{27}$$

To see this, let us fix an $\varepsilon > 0$ and observe that, since $f \in \overline{\operatorname{conv}}(A)$, the Cauchy-Schwarz inequality implies that there must be a finite convex combination of elements in A satisfying

$$\sum_{i=1}^k \alpha_i\langle f-a_i, f-h\rangle = \left\langle f - \sum_{i=1}^k\alpha_i a_i, f-h\right\rangle \leqslant \varepsilon.$$

In particular, the inequality $\langle f-a_i, f-h\rangle \leqslant \varepsilon$ holds for at least one vector $a_i \in A$. Since $\varepsilon > 0$ was arbitrary, we have established (27).

Inserting (27) in (26), we arrive at

$$\Xi_{m+1}^2 \leqslant \inf_{\lambda\in[0,1]}\Big[\lambda^2\xi^2 + (1-\lambda)^2\Xi_m^2\Big] \leqslant \frac{\xi^2\Xi_m^2}{\xi^2+\Xi_m^2},$$

where the last step follows by chosing $\lambda = \frac{\Xi_m^2}{\Xi_m^2+\xi^2} \in [0,1]$. Finally, recalling the induction hypothesis $\Xi_m \leqslant \frac{\xi}{\sqrt{m}}$, we conclude

$$\Xi_{m+1}^2 \leqslant \xi^2\frac{\frac{\xi^2}{m}}{\xi^2+\frac{\xi^2}{m}} = \frac{\xi^2}{m+1}.$$

This establishes (25) for $m+1$ and the induction is complete. □

Returning to the original statement of Proposition 4.4 for a general Banach space, the next example shows that its convergence rate is optimal in general:

Example A.2. For $p \in (1,2]$ let us consider the Banach space $\ell^p(\mathbb{R})$ of p-summable real-valued sequences, that is, $(a_n)_{n\in\mathbb{N}} \subset \mathbb{R}$ belongs to $\ell^p(\mathbb{R})$ iff

$$\|a\|_{\ell^p} = \left(\sum_{n=1}^{\infty} |a_n|^p\right)^{\frac{1}{p}} < \infty.$$

It can be shown that this Banach space has Rademacher type $\mathfrak{t} = p$. Let A be the set formed of the standard basis vectors:

$$A = \big\{(1,0,0,0,\dots),(0,1,0,0,\dots),(0,0,1,0,\dots),\dots\big\}.$$

Choosing $f \equiv 0$, we find that

$$\inf_{h\in\mathrm{conv}_m(A)} \|f-h\|_{\ell^p} = \inf_{(\alpha_1,\dots,\alpha_m)\in\triangle_m} \left(\sum_{i=1}^{m} \alpha_i^p\right)^{\frac{1}{p}}.$$

The optimum on the right-hand side is attained by choosing $\alpha_1 = \dots = \alpha_m = \frac{1}{m}$ so that

$$\inf_{h\in\mathrm{conv}_m(A)} \|f-h\|_{\ell^p} = \frac{1}{m^{1-\frac{1}{p}}} = \frac{1}{m^{1-\frac{1}{\mathfrak{t}}}}.$$

A.3. Details of the Computation for Equations (17) and (18)

We begin with a calculus lemma:

Lemma A.3. *Let $\alpha, A, B > 0$. The function $f : (0,\infty) \to \mathbb{R}_+$,*

$$f(\sigma) = A\sigma + \frac{B}{\sigma^\alpha},$$

is minimized for

$$\sigma_\star = \left(\frac{\alpha B}{A}\right)^{\frac{1}{\alpha+1}} \tag{28}$$

with function value

$$f(\sigma_\star) = (\alpha+1)\left(\frac{A}{\alpha}\right)^{\frac{\alpha}{\alpha+1}} B^{\frac{1}{\alpha+1}}. \tag{29}$$

Proof. We have $f'(\sigma) = A - \alpha B\sigma^{-(\alpha+1)}$, whence the minimizer is given by

$$\sigma_\star = \left(\frac{\alpha B}{A}\right)^{\frac{1}{\alpha+1}}.$$

Inserting this into the function expression we find

$$f(\sigma_\star) = A^{\frac{\alpha}{\alpha+1}}(\alpha B)^{\frac{1}{\alpha+1}} + B^{\frac{1}{\alpha+1}}\left(\frac{A}{\alpha}\right)^{\frac{\alpha}{\alpha+1}} = A^{\frac{\alpha}{\alpha+1}} B^{\frac{1}{\alpha+1}}\left(\alpha^{\frac{1}{\alpha+1}} + \alpha^{-\frac{\alpha}{\alpha+1}}\right) = (\alpha+1)\left(\frac{A}{\alpha}\right)^{\frac{\alpha}{\alpha+1}} B^{\frac{1}{\alpha+1}},$$

as required. □

We now apply Lemma A.3 with

$$A = \mathcal{M}_\varphi \|\nabla\|_{L^q}, \qquad B = \frac{2\Upsilon_q \|\varphi_q\|_{L^q}}{m^{1-\frac{1}{\min(q,2)}}}, \qquad \alpha = d - \frac{d}{q}.$$

Since $\frac{1}{\alpha+1} = \frac{q}{d(q-1)+q}$, (28) immediately gives (17). Similarly, we have $\frac{\alpha}{\alpha+1} = \frac{d(q-1)}{d(q-1)+q}$ so that (29) gives

$$\|f - g_m\|_{L^q} \leqslant \left(d - \frac{d}{q} + 1\right) \left(\frac{\mathcal{M}_\varphi \|\nabla\|_{L^q}}{d - \frac{d}{q}}\right)^{\frac{d(q-1)}{d(q-1)+q}} \left(\frac{2\Upsilon_q \|\varphi_q\|_{L^q}}{m^{1-\frac{1}{\min(q,2)}}}\right)^{\frac{1}{d-\frac{d}{q}+1}},$$

which is the estimate (18).

A.4. Construction of the Counterexample in Remark 4.9

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 $(0,1]$. On the other hand, it holds that

$$\sup_{x\in[0,1]} \left|f_m(x) - f(x)\right| = \left|f_m(0) - f(0)\right| = \frac{1}{2} \qquad \forall m \in \mathbb{N}.$$

Consequently, f_m does not converge uniformly to f.

Nevertheless, 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 4.8.