

Universal spin models are universal approximators in machine learningTobias Reinhart ^{*}*Institute for Theoretical Physics, University of Innsbruck, Technikerstr. 21a, A-6020 Innsbruck, Austria*Gemma De les Coves *Departament d'Enginyeria, Universitat Pompeu Fabra, Carrer Tànger 122, 08018 Barcelona, Catalonia and ICREA, Institució Catalana d'Estudis i Recerca Avançats, Passeig Lluís Companys 23, 08010 Barcelona, Catalonia*

(Received 11 July 2025; accepted 9 March 2026; published 14 April 2026)

One of the theoretical pillars that sustain certain machine learning models are universal approximation theorems, which prove that they can approximate all functions from a function class to arbitrary precision. Independently, classical spin models are termed universal if they can reproduce the behavior of any other spin model in their low energy sector. Universal spin models have been characterized via sufficient and necessary conditions, showing that simple models such as the $2d$ Ising with fields are universal. In this work, we prove that universal spin models are universal approximators of probability distributions. This enables us to leverage the characterization of the former to reveal conditions which are sufficient for universal approximation. Deriving universal approximation theorems thus amounts to verifying these conditions, yielding a unified recipe for universal approximation theorems applicable to a wide range of models. We explicitly test this recipe for restricted and deep Boltzmann machines, as well as for deep belief networks. This work illustrates that independently discovered universality statements may be intimately related, enabling the transfer of results.

DOI: [10.1103/b3xp-wtwx](https://doi.org/10.1103/b3xp-wtwx)

Introduction. Spin models are essentially generalizations of the Ising model [1]. They are toy models of complex systems that consist of many interacting, classical degrees of freedom, called spins. Their versatile nature has made them amenable for problems not only in physics, but also in computational complexity, biology, or the social sciences (see, e.g., [2] for a list of applications), as well as machine learning. In the latter, they give rise to energy based models, such as restricted Boltzmann machines (RBMs) or deep Boltzmann machines (DBMs) [3], where one is interested in the Boltzmann distribution, which describes the behavior of the spin model at thermal equilibrium. This probability distribution is parametrized by couplings that represent interactions between spins. The learning task consists of finding parameters such that the output distribution in the so-called visible spins approximates a given target distribution.

Several machine learning models, including RBMs and DBMs, are theoretically supported by universal approximation theorems (UATs), which prove that they can approximate any function from a given class to arbitrary precision [4–7]. In particular, the UATs for RBMs and DBMs guarantee that both can approximate arbitrary probability distributions. In

this sense, they are universal approximators of probability distributions.

Independently, certain spin models (such as the $2d$ Ising model with fields or the $3d$ Ising model) have been proven to be universal, meaning that their low energy sector can reproduce the behavior of any other spin model [8,9]. While universal spin models have been fully characterized in terms of sufficient and necessary conditions, to the best of our knowledge no such characterization exists for universal approximators in machine learning (cf. [7]).

In this paper, we prove that universal spin models are universal approximators of probability distributions by leveraging [9] (see Fig. 1). It follows that the characterization of the former extends to the latter, yielding a characterization of universal approximators for energy based models [10] and in turn a unified recipe for the derivation of UATs. We apply this recipe to RBMs and DBMs, resulting in proofs of the respective UATs. As a first step beyond energy based models, we apply the procedure to derive a UAT for deep belief networks (DBNs). Overall, we find that universality statements of spin models and some machine learning models are intimately connected.

Spin systems, spin models, and simulations. We shall consider the following simplified versions of definitions of [9]. Spin models are potentially infinite sets of spin systems. Spin systems consist of finitely many classical, binary degrees of freedom, or “spins” interacting according to a hypergraph (V, E) . Each spin is associated to a vertex $v \in V$, and each interaction between spins is associated to a hyperedge $e \in E$ and given by a local energy function. A configuration $\mathbf{s}$ assigns a number from $\{0, 1\}$ to each spin. Mapping configurations $\mathbf{s}$

^{*}Contact author: tobias.reinhart@uibk.ac.at

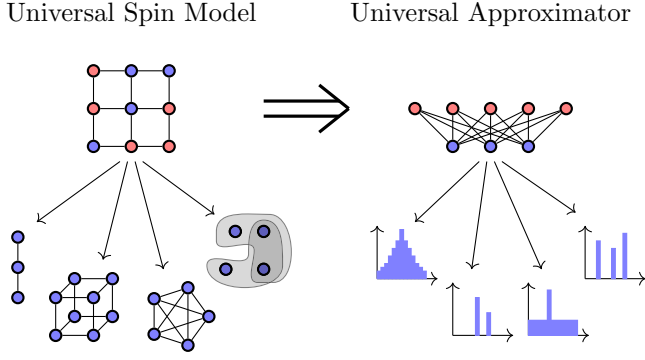


FIG. 1. We prove that universal spin models are universal approximators of probability distributions in machine learning. We leverage the characterization of the former [9] to derive UATs for the latter. We explicitly do so for RBMs, DBMs, and DBNs.

to the sum of their local energy contributions $H(\mathbf{s})$ defines the Hamiltonian H of the spin system. The Boltzmann distribution,

$$p(\mathbf{s}) = \frac{1}{Z} e^{-H(\mathbf{s})}, \quad (1)$$

is the probability distribution of configurations at thermal equilibrium (at inverse temperature $\beta = 1$), where the normalization Z is the partition function.

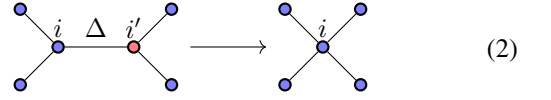
Simulations are transformations of spin systems that preserve the low energy sector. Given spin systems S and T (for source and target) with Hamiltonians H_S and H_T and with $V_T \subseteq V_S$ and a cut-off energy Δ , we say that S simulates T with cut-off Δ (and write $S \rightarrow T$) if up to a constant energy shift Γ (Fig. 2).

(i) For each target configuration $\mathbf{t}$ with $H_T(\mathbf{t}) \leq \Delta$ there exists a unique configuration $\mathbf{h}_{\mathbf{t}}$ of auxiliary (or hidden) spins $V_S \setminus V_T$ such that the combined source configuration $(\mathbf{t}, \mathbf{h}_{\mathbf{t}})$ satisfies $H_S(\mathbf{t}, \mathbf{h}_{\mathbf{t}}) = H_T(\mathbf{t})$;

(ii) All other source configurations have energy at least Δ .

For a simple example, consider any spin system T and any spin i from T , and construct S by adding a single auxiliary spin i' that interacts with i with a large ferromagnetic coupling Δ . Then, for any target configuration $\mathbf{t}$, taking i' to be in the same state as i yields a source configuration with equal energy (up to a shift). Conversely, all other source configurations, i.e., where i and i' are in different states, yield energies above the cut-off. This simulation allows one to reduce the number of

adjacent spins of i in T ,



Applying it repeatedly, one can prove that spin systems on $2d$ square lattices can simulate those on arbitrary planar graphs [9].

We say that S simulates a real-valued function f of spin configurations if it simulates any spin system with Hamiltonian f . We say that S simulates a probability distribution p over spin configurations on V_T if for all such configurations $\mathbf{t}$,

$$p_{V_T}[S](\mathbf{t}) = p(\mathbf{t}) + \mathcal{O}(e^{-\Delta}), \quad (3)$$

that is, $p_{V_T}[S]$, the marginal of the source Boltzmann distribution $p[S]$ over physical spins V_T approximates the target distribution p with an error that decreases exponentially with the cut-off.

Simulations also approximately preserve several properties of spin systems [9], illustrating that they can be interpreted as approximations whose error is controlled by the cut-off. In particular, they preserve Boltzmann distributions (Lemma 2 of [11]), that is, whenever $S \rightarrow T$, then S also simulates the Boltzmann distribution of T [according to Eq. (3)]. Note that the pointwise approximation of Eq. (3) implies approximations in terms of the Kullback-Leibler divergence and the total variation distance. Further, simulations have a modular structure, meaning they can be combined in several ways, allowing to construct complicated simulations from simpler ones [12]

Simulations are easily promoted to spin models. We say that a spin model $\mathcal{M}$ simulates a spin system T (or a function f or a probability distribution p) with cut-off Δ if there exists a spin system $S \in \mathcal{M}$ that simulates T (or f , or p) with cut-off Δ . We say that $\mathcal{M}$ simulates T (or f , or p) if this holds for all cut-offs. Finally, a spin model is called universal if it simulates all spin systems. It is called a universal approximator (of probability distributions) if it simulates all probability distributions over spin configurations.

Universal spin models can be characterized in terms of seemingly weak properties, implying that fairly simple models like the $2d$ Ising model with fields or the $3d$ Ising model are universal [9]. Here we characterize them by flag completeness and closure. S is a flag system for configuration $\mathbf{x}$ if up to an energy shift,

$$H_S(\mathbf{s}, \mathbf{h}) = \begin{cases} 0 & \text{if } \mathbf{s} = \mathbf{x} \text{ and } \mathbf{h} = \mathbf{h}_{\mathbf{x}} \\ 1 & \text{if } \mathbf{s} \neq \mathbf{x} \text{ and } \mathbf{h} = \mathbf{h}_{\mathbf{x}} \\ \geq 1 & \text{if } \mathbf{s} = \mathbf{x} \text{ and } \mathbf{h} \neq \mathbf{h}_{\mathbf{x}} \\ \geq 2 & \text{if } \mathbf{s} \neq \mathbf{x} \text{ and } \mathbf{h} \neq \mathbf{h}_{\mathbf{x}}. \end{cases} \quad (4)$$

In words, given a fixed configuration $\mathbf{s}$ of physical spins, the energy of S is minimal whenever the configuration $\mathbf{h}$ of flag spins correctly signals whether $\mathbf{s} = \mathbf{x}$ or not, i.e., if $\mathbf{h}$ equals the truth value of “ $\mathbf{s} = \mathbf{x}$ ” with configurations $\mathbf{h}_{\mathbf{x}}$ corresponding to “TRUE,” and $\mathbf{h}_{\neq \mathbf{x}}$ corresponding to “FALSE.” Thus, in the low energy sector, information over the physical spins can be retrieved from the flag spins, only. See Eqs. (9) and (12) for examples.

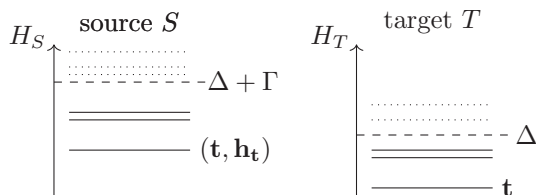


FIG. 2. If S simulates T , $H_S(\mathbf{t}, \mathbf{h}_{\mathbf{t}}) = H_T(\mathbf{t})$ below the cut-off Δ up to a shift Γ .

A spin model is flag complete if it contains flag systems for arbitrary configurations. Additionally, a spin model $\mathcal{M}$ is closed for a subset $\mathcal{N} \subset \mathcal{M}$ if it simulates arbitrary non-negative linear combinations from $\mathcal{N}$.

Theorem 1. A spin model $\mathcal{M}$ is universal if and only if

- (i) it is flag complete, and
- (ii) it is closed for any subset of flag systems with disjoint flag spins.

Proof. The “only if” direction is immediate, since then $\mathcal{M}$ simulates all spin systems.

We prove the “if” direction constructively. Without loss of generality we assume that all flag systems have zero energy shift, i.e., satisfy Eq. (4) exactly.

First, we prove that linear combinations of flag systems can simulate an arbitrary function f . The idea is that $\lambda S_{\mathbf{x}}$ for large λ moves $\mathbf{x}$ to the low energy sector. Thus, the linear combination $S = \sum_i \lambda_i S_{\mathbf{x}_i}$ over all low energy target configurations $f(\mathbf{x}_i) \leq \Delta$, for appropriate λ_i , reproduces the low energy sector of f . Specifically, we define

$$\begin{aligned} f_{\text{diff}} &= \max\{f(\mathbf{x}_i) - f(\mathbf{x}_j) \mid f(\mathbf{x}_i), f(\mathbf{x}_j) \leq \Delta\} \\ \lambda_i &= \Delta + f_{\text{diff}} - f(\mathbf{x}_i) \\ \Gamma &= \sum_i \lambda_i - (\Delta + f_{\text{diff}}). \end{aligned} \quad (5)$$

Then, high energy target configurations $\mathbf{y}$ yield $H_S(\mathbf{y}, \mathbf{h}) \geq \Delta + \lambda$, regardless of the configuration of flag spins, $\mathbf{h}$. Low energy target configurations $\mathbf{x}_i$ yield $H_S(\mathbf{x}_i, \mathbf{h}) = \Gamma + f(\mathbf{x}_i)$ if the flag spins flag correctly, and $H_S(\mathbf{x}_i, \mathbf{h}) \geq \Gamma + \Delta$ otherwise. Thus, S simulates f with cut-off Δ and energy shift $-\Gamma$.

Now any spin system T can be decomposed into the sum of its local energy functions $T = \sum_e T_e$, each of which can be simulated by a linear combination of flag systems $\sum_{i_e} \lambda_{i_e} S_{\mathbf{x}_{i_e}}$. Additivity of simulations [9] then implies that

$$\sum_e \sum_{i_e} \lambda_{i_e} S_{\mathbf{x}_{i_e}} \rightarrow \sum_e T_e = T. \quad (6)$$

Finally, closure and flag completeness imply that $\mathcal{M}$ simulates the left hand side of (6). Since simulations can be composed, this proves the claim (see Fig. 2 of [9] for an illustration). ■

Constructing the simulation of T as a sum of simulations of its local energy functions T_e is more efficient than directly simulating T , as the former requires $\mathcal{O}(\sum_e 2^{|e|})$ instead of $\mathcal{O}(2^{|V_T|})$ auxiliary spins, which, for typical targets T , is polynomially instead of exponentially many.

Restricted and deep Boltzmann machines. We now apply Theorem 1 to certain machine learning models, more precisely to their underlying spin model. We will prove that for both RBMs and DBMs, the underlying spin model is universal.

For our purposes, we identify RBMs with their underlying spin model. In our eyes, an RBM is a spin system with complete bipartite interaction graph between visible spins V_v and hidden spins V_h , with Hamiltonian

$$H(\mathbf{v}, \mathbf{h}) = \mathbf{v}^t \mathbf{b}_v + \mathbf{h}^t \mathbf{b}_h + \mathbf{v}^t \mathbf{W} \mathbf{h}, \quad (7)$$

parametrized by coupling vectors $\mathbf{b}_v, \mathbf{b}_h$ and coupling matrix $\mathbf{W}$, where ^{t} denotes transpose.

Similarly, a DBM is a spin system whose interaction graph consists of fully connected layers $V = \cup_i V_i$, such that each

subsystem R_i with spins $V_i \cup V_{i+1}$ defines an RBM, with Hamiltonian H_i . Denoting the configuration of layer i by $\mathbf{h}_i$, the total Hamiltonian is given by

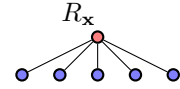
$$H(\mathbf{h}_0, \dots, \mathbf{h}_k) = \left(\sum_{i=0}^{k-2} H_i(\mathbf{h}_i, \mathbf{h}_{i+1}) \right) + H_{k-1}(\mathbf{h}_{k-1}, \mathbf{h}_k). \quad (8)$$

We identify $\mathbf{h}_0 = \mathbf{v}$ and call spins from V_0 visible spins, and those from $V \setminus V_0$ hidden spins.

Theorem 2. RBMs are universal.

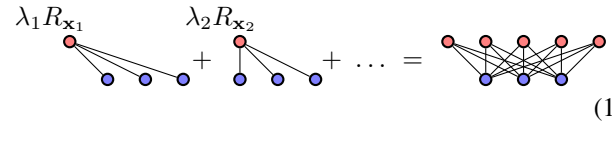
That is, the spin model given by all RBMs can simulate all spin systems. Moreover, the physical spins of each simulation are the visible spins of the RBM, and the auxiliary spins are the hidden spins. In addition, simulating a function with k low energy configurations requires an RBM with k hidden spins.

Proof. We exploit Theorem 1. Flagging configuration $\mathbf{x}$ can be achieved by an RBM

$$R_{\mathbf{x}} \quad (9)$$


with a single hidden spin serving as flag spin, and with couplings such that components of $\mathbf{W}$ are given by $W_i = -2$ if $x_i = 1$ and $W_i = +2$ if $x_i = 0$, $\mathbf{b}_v = \mathbf{0}$, and $\mathbf{b}_h = 2\mathbf{d}(\mathbf{x}, \mathbf{0}) - \mathbf{1}$, where $\mathbf{d}(\cdot, \cdot)$ is the Hamming distance [13]. It is easy to check that $R_{\mathbf{x}}$ satisfies Eq. (4) up to a shift of $+1$, with $h_{\mathbf{x}} = +1$ and $h_{\neq \mathbf{x}} = -1$.

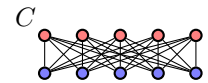
Closure is trivial since scaling flag RBMs amounts to changing their couplings, and sums of flag RBMs with disjoint flag spins are trivially RBMs. Following Theorem 1, the universal RBM is a linear combination of flag RBMs $R_{\mathbf{x}_i}$, that is,

$$\lambda_1 R_{\mathbf{x}_1} + \lambda_2 R_{\mathbf{x}_2} + \dots = \quad (10)$$


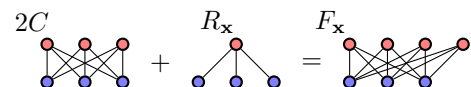
Theorem 3. DBMs of constant width are universal.

Specifically, the spin model of all DBMs with visible layer of width n and hidden layers of width $n+1$, for any n , can simulate all spin systems. Again, physical spins of the simulations precisely correspond to visible spins of the DBM. Simulating a function with k low energy configurations of n spins requires a DBM with k hidden layers, and thus $(n+1)k$ hidden spins.

Proof. We exploit Theorem 1. First, define the RBM

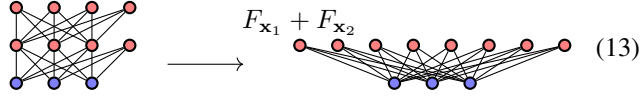
$$C \quad (11)$$


with $\mathbf{b}_v = \mathbf{b}_h = \mathbf{1}$ and $\mathbf{W} = -2\mathbf{I}$, where $\mathbf{I}$ denotes the identity matrix. Then $H_C(\mathbf{v}, \mathbf{h}) = \mathbf{d}(\mathbf{v}, \mathbf{h})$. In words, in the ground state C copies configurations from visible to hidden spins. Defining $F_{\mathbf{x}}$ by

$$2C + R_{\mathbf{x}} = F_{\mathbf{x}} \quad (12)$$


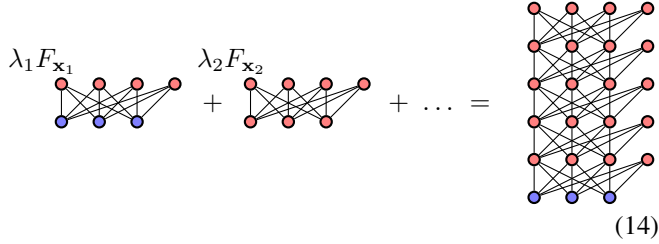
yields a flag system, i.e., satisfies Eq. (4).

Closure follows from simulating the sum of flag systems $F_{\mathbf{x}_1} + F_{\mathbf{x}_2}$ as



This simulation works since in the low energy sector, C copies configurations. Its cut-off can be controlled by scaling the copy systems of the left-hand side. Note that in contrast to Theorem 2, closure is nontrivial since sums of flag systems $F_{\mathbf{x}_i}$ do not yield DBMs of constant width.

Following Theorem 1, the universal DBM is a simulation of a linear combination of flag systems $F_{\mathbf{x}_i}$, that is,



Universal spin models are universal approximators. Let us now show that universal spin models are universal approximators of probability distributions. This means that universal spin models, via their Boltzmann distributions, can approximate arbitrary probability distributions over spin configurations to arbitrary precision.

Observe first that any probability distribution p on $\{0, 1\}^V$ can be simulated by a spin system T . To this end, construct T such that configurations $\mathbf{t}$ with nonzero probability $p(\mathbf{t}) > 0$ are assigned low energy, specifically $-\log(p(\mathbf{t}))$, and those with zero probability are assigned high energy, specifically Δ . Details are provided in Lemma 3 [11].

Theorem 4. Universal spin models are universal approximators.

Proof. Consider any probability distribution p and the corresponding spin system T that simulates p . By assumption, a universal spin model can simulate T for arbitrary cut-offs. Simulations preserve Boltzmann distributions (Lemma 2 [11]), hence the claim. ■

Corollary 5. Both RBMs and DBMs are universal approximators.

Proof. It follows from Theorems 2–4. ■

This is a new proof of the UAT for RBMs and DBMs. While existing proofs of UATs are derived in a case-by-case fashion (cf. [4,6]), ours emanate from an underlying, unified recipe for the derivation for UATs (Theorems 1 and 4) based on conditions that trigger the onset of universality. As such, this work sheds light on the origin and reach of these forms of universality.

Deep belief networks. Let us extend the characterization of universal approximators to deep belief networks (DBNs) and thus beyond energy based models. Similar to DBMs, DBNs are defined on layerwise fully connected graphs. However, their output is the marginal of

$$q(\mathbf{h}_0, \dots, \mathbf{h}_k) = \left(\prod_{i=0}^{k-2} p_i(\mathbf{h}_i | \mathbf{h}_{i+1}) \right) p_{k-1}(\mathbf{h}_{k-1}, \mathbf{h}_k) \quad (15)$$

over visible spins, where $p_i(\mathbf{h}_i, \mathbf{h}_{i+1})$ is the Boltzmann distribution of the i th RBM [cf. Eq. (8)]. For $k = 2$ ($S = R_0 + R_1$) this marginal can be expressed as

$$q_{V_0}[S](\mathbf{h}_0) = \sum_{\mathbf{h}_1} p[R_0](\mathbf{h}_0 | \mathbf{h}_1) p_{V_1}[R_1](\mathbf{h}_1). \quad (16)$$

Following [6], the corresponding DBM output equals

$$p_{V_0}[S](\mathbf{h}_0) = \sum_{\mathbf{h}_1} p[R_0](\mathbf{h}_0 | \mathbf{h}_1) (p_{V_1}[R_0] * p_{V_1}[R_1])(\mathbf{h}_1), \quad (17)$$

where $*$ denotes the pointwise, normalized product of probability distributions (see [6]).

Thus, the difference between the DBN and DBM output is the factor $p_{V_1}[R_0]$, namely the top-level marginal of R_0 , appearing in the latter. In [6] this factor is termed the backwards signal of R_0 , as it distinguishes the directional information flow of DBNs from the undirected one of DBMs.

Under certain assumptions, the backwards signal of R_0 can be absorbed in a change of couplings of R_1 , yielding a system $S' = R_0 + R'_1$ whose DBN output approximates the DBM output of S . More precisely, when the top-level marginal of R_0 is approximately spiked, i.e., up to $\mathcal{O}(e^{-\Delta})$ satisfies

$$p_{V_1}[R_0](\mathbf{h}_1) = \begin{cases} a & \text{if } \mathbf{h}_1 = \mathbf{x} \\ b & \text{else,} \end{cases} \quad (18)$$

then there exist systems R'_1 and $S' = R_0 + R'_1$ such that

$$q_{V_0}[S'](\mathbf{h}_0) = p_{V_0}[S](\mathbf{h}_0) + \mathcal{O}(e^{-\Delta}). \quad (19)$$

This is proven in Lemma 4 [11].

For $k > 2$ the difference between DBM and DBN output consist of multiple backward signals. In Lemma 5 [11], we prove that under certain assumptions, they can be absorbed by applying Lemma 4 [11] iteratively. Hence, some DBMs can be simulated by DBNs. We call such DBMs effectively directional.

We now prove that effectively directional DBMs are universal approximators (Theorem 4). Since they can be simulated by DBNs, this entails a UAT for DBNs (Theorem 6). To this end, we consider DBMs whose top-level RBM is a linear combination of two flag RBMs for configurations $\mathbf{z}$ and $\mathbf{x}$,

When scaled appropriately, $R_{\mathbf{x},\mathbf{z}}$ simulates

$$f_{\mathbf{x},\mathbf{z}}(\mathbf{v}) = \begin{cases} c & \text{if } \mathbf{v} = \mathbf{x} \\ d & \text{if } \mathbf{v} = \mathbf{z} \\ \Delta & \text{else,} \end{cases} \quad (21)$$

with cut-off Δ .

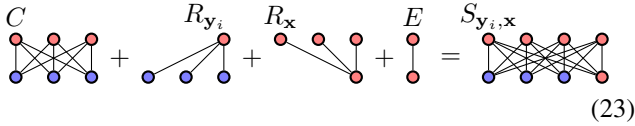
Each lower level RBM simulates

$$f_{y_i,\mathbf{x}}(\mathbf{v}, \mathbf{h}) = \begin{cases} a_i & \text{if } \mathbf{v} = \mathbf{h} \\ b_i & \text{if } \mathbf{v} = \mathbf{y}_i \text{ and } \mathbf{h} = \mathbf{x} \\ \Delta & \text{else,} \end{cases} \quad (22)$$

with cut-off Δ . Following [14] we call systems that simulate Eq. (22) sharing systems and denote them $S_{y_i,\mathbf{x}}$. Their Boltzmann distribution is such that whenever $\mathbf{h} = \mathbf{x}$ there is

a nonzero probability (determined by a_i and b_i) to obtain $\mathbf{v} = \mathbf{y}_i$, while all configurations with $\mathbf{h} \neq \mathbf{x}$ are deterministically copied to $\mathbf{v} = \mathbf{h}$. In this sense, $S_{\mathbf{y}_i, \mathbf{x}}$ shares probability mass between $\mathbf{x}$ and $\mathbf{y}_i$.

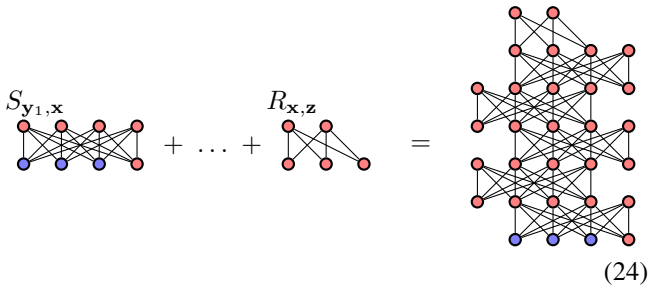
We construct $S_{\mathbf{y}_i, \mathbf{x}}$ from a copy system, C , two flag systems, $R_{\mathbf{y}_i}$ and $R_{\mathbf{x}}$, and an interaction between their flag spins, E ,



$$C + R_{\mathbf{y}_i} + R_{\mathbf{x}} + E = S_{\mathbf{y}_i, \mathbf{x}} \quad (23)$$

In words, C introduces an energy penalty whenever $\mathbf{v} \neq \mathbf{h}$. The flag systems ensure that if $\mathbf{h} = \mathbf{x}$ and $\mathbf{v} = \mathbf{y}_i$, in their low energy sector, both flag spins are in state 1. Choosing E appropriately then cancels the energy penalty of C and moves the configuration with $\mathbf{h} = \mathbf{x}$ and $\mathbf{v} = \mathbf{y}_i$ to the overall low energy sector. Details can be found in Lemma 1 [11].

We now consider DBMs composed of sharing systems and a top level flag RBM. We call such DBMs multisharing systems:



$$S_{\mathbf{y}_1, \mathbf{x}} + \dots + S_{\mathbf{y}_n, \mathbf{x}} + R_{\mathbf{x}, \mathbf{z}} = \text{Multisharing System} \quad (24)$$

In Lemma 5 [11] and Lemma 6 [11], we prove that multisharing systems are effectively directional and universal. This directly implies a UAT for DBNs, and shows that the characterization of universal approximators extends beyond energy based models.

Theorem 6. DBNs of constant width are universal approximators.

Specifically, DBNs with layers of width $n + 1$ can simulate arbitrary probability distributions over configurations of n spins.

Proof. It follows from Lemma 5 [11] and Lemma 6 [11], and Theorem 4. ■

Outlook. The characterization of universal approximators derived in this work not only unifies existing UATs, but also extends beyond known examples. It may thus guide the construction of universal machine learning models, which could be attractive in two situations.

First, for “nonstandard” training algorithms. While “standard” training by contrastive divergence is particularly efficient for bipartite or multipartite architectures [15], the performance of other training algorithms depends on other properties [16,17]. Can we find universal models with the desirable properties?

Second, for nongeneric target distributions. The characterization of universality implies that certain targets can be efficiently simulated, that is, with a polynomial instead of exponential number of parameters. Given a set of target distributions, can we devise a universal model that can efficiently represent all distributions from the set? Conversely, given a model, what can it efficiently represent? Even for RBMs and DBMs this is the subject of ongoing research [18–21].

Further, while our results for DBNs extend the characterization of universal approximators beyond energy based models, other examples of universal approximators are often based on continuous variables instead of discrete spins. These include feed forward neural networks (FFNs) [22,23], graph neural networks [24], and transformers [25]. Can our characterization be extended to continuous variables (cf. [8]) and thus to some of these examples (cf. [5])?

Finally, the mathematical structure underpinning universal spin models is shared with other instances of universality, including universal Turing machines and universal gate sets from the theory of (quantum) computation [26]. Since this work puts neural networks and their feature of universality on an equal footing with these other instances, it raises the following questions. Do other instances allow for a characterization of universality similar to spin models and neural networks? How do different instances of universality compare? For example, comparing neural networks to Turing machines (or weaker automata) might reveal differences of their computational power [27,28] (cf. [2]), possibly even shedding light on the distinction between artificial and natural computation [29].

Acknowledgments. This research was funded in parts by the Austrian Science Fund 10.55776/Y1261.

Data availability. No data were created or analyzed in this study.

- [1] E. Ising, Beitrag zur Theorie des Ferromagnetismus, *Z. Angew. Phys.* **31**, 253 (1925).
- [2] T. Reinhart and G. De les Coves, The grammar of the Ising model: A new complexity hierarchy, *Proc. R. Soc. A* **481**, 20240579 (2025).
- [3] T. Osogami, Boltzmann machines and energy-based models, *arXiv:1708.06008*.
- [4] N. Le Roux and Y. Bengio, Representational power of restricted Boltzmann machines and deep belief networks, *Neural Comput.* **20**, 1631 (2008).
- [5] N. Le Roux and Y. Bengio, Deep belief networks are compact universal approximators, *Neural Comput.* **22**, 2192 (2010).

- [6] G. Montúfar, Deep narrow Boltzmann machines are universal approximators, in *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, arXiv:1411.3784* (2015).
- [7] G. Montúfar and J. Rauh, Hierarchical models as marginals of hierarchical models, *Int. J. Approximate Reasoning* **88**, 531 (2017).
- [8] G. De las Cuevas and T. S. Cubitt, Simple universal models capture all classical spin physics, *Science* **351**, 1180 (2016).
- [9] T. Reinhart, B. Engel, and G. D. les Coves, The structure of emulations in classical spin models: Modularity and universality, *arXiv:2407.13428*.

- [10] That is machine learning models, defined by a Hamiltonian the corresponding Boltzmann distribution, examples are DBMs and RBMs.
- [11] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/b3xp-wtwx> for additional details on simulations.
- [12] Simulations can be composed, scaled and added [9]. This implies, e.g. that they can be constructed from *gadgets*, spin systems that locally replace a part of the target, similar to Eq. (2).
- [13] We denote by $\mathbf{0}$ the all zero and by $\mathbf{1}$ the all one vector.
- [14] I. Sutskever and G. E. Hinton, Deep, narrow sigmoid belief networks are universal approximators, *Neural Comput.* **20**, 2629 (2008).
- [15] G. Hinton, Training products of experts by minimizing contrastive divergence, *Neural Comput.* **14**, 1771 (2002).
- [16] Y. Song and D. P. Kingma, How to train your energy-based models, [arXiv:2101.03288](https://arxiv.org/abs/2101.03288).
- [17] D. Carbone, Hitchhiker's guide on the relation of energy-based models with other generative models, sampling and statistical physics: A comprehensive review, *Transactions on Machine Learning Research* (2025).
- [18] X. Gao and L.-M. Duan, Efficient representation of quantum many-body states with deep neural networks, *Nat. Commun.* **8**, 662 (2017).
- [19] G. Carleo, Y. Nomura, and M. Imada, Constructing exact representations of quantum many-body systems with deep neural networks, *Nat. Commun.* **9**, 5322 (2018).
- [20] L. Gu, F. Zhou, and L. Yang, Towards the representational power of restricted Boltzmann machines, *Neurocomputing* **415**, 358 (2020).
- [21] J. Martens, A. Chattopadhyay, T. Pitassi, and R. Zemel, On the representational efficiency of restricted Boltzmann machines, in *Advances in Neural Information Processing Systems*, Vol. 26 (2013).
- [22] K. Hornik, M. Stinchcombe, and H. White, Multilayer feedforward networks are universal approximators, *Neural Networks* **2**, 359 (1989).
- [23] G. Cybenko, Approximation by superpositions of a sigmoidal function, *Math. Control Signal Systems* **2**, 303 (1989).
- [24] R. Sato, A survey on the expressive power of graph neural networks, [arXiv:2003.04078](https://arxiv.org/abs/2003.04078).
- [25] C. Yun, S. Bhojanapalli, A. S. Rawat, S. Reddi, and S. Kumar, Are transformers universal approximators of sequence-to-sequence functions? in *Proceedings of the 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia* (2020).
- [26] T. Gonda, T. Reinhart, S. Stengele, and G. D. les Coves, A framework for universality in physics, computer science, and beyond, *Compositionality* **6**, 3 (2024).
- [27] H. Siegelmann and E. Sontag, On the computational power of neural nets, *J. Comput. Syst. Sci.* **50**, 132 (1995).
- [28] J. Pérez, J. Marinković, and P. Barceló, On the Turing completeness of modern neural network architectures, in *Proceedings of the 7th International Conference on Learning Representations, ICLR 2019, New Orleans, Louisiana, USA* (2019).
- [29] P. Fraser, R. Solé, and G. De las Cuevas, Why can the brain (and not a computer) make sense of the liar paradox? *Front. Ecol. Evol.* **9**, 802300 (2021).