

---

# Quantum Boltzmann Machines for Sample-Efficient Reinforcement Learning

---

**Thore Gerlach**

University of Bonn

tgerlach1@uni-bonn.de

**Michael Schenk**

CERN

michael.schenk@cern.ch

**Verena Kain**

CERN

verena.kain@cern.ch

## Abstract

We introduce theoretically grounded Continuous Semi-Quantum Boltzmann Machines (CSQBM) that supports continuous-action reinforcement learning. By combining exponential-family priors over visible units with quantum Boltzmann distributions over hidden units, CSQBM yield a hybrid quantum-classical model that reduces qubit requirements while retaining strong expressiveness. Crucially, gradients with respect to continuous variables can be computed analytically, enabling direct integration into Actor-Critic algorithms. Building on this, we propose a continuous  $Q$ -learning framework that replaces global maximization by efficient sampling from the CSQBM distribution, thereby overcoming instability issues in continuous control.

## 1 Introduction

In recent years, Reinforcement Learning (RL) has been increasingly investigated for deployment in complex real-world scenarios [1]. One major challenge is the restricted access to training data, as in the case of *beam control* in high-energy physics experiments at CERN [2–5]. This task is further complicated by demands for higher beam intensities, smaller beam sizes, and diverse experimental requirements. While physics models support many beam tasks, several systems still rely on manual tuning. These systems often present complex non-linear continuous-control optimization problems where *sample efficiency* is critical due to limited beam times in accelerator operation. Classical RL algorithms, however, often fail to meet the stringent efficiency and stability requirements.

Energy-based models, and in particular Boltzmann Machines (BM) [6], offer a principled probabilistic framework for approximating value functions in RL, while being sample-efficient [7–9]. Since BMs can approximate any probability distribution arbitrarily well, they are capable of capturing highly non-linear correlations in data. At the same time, general BMs are notoriously difficult to train, as sampling from a Boltzmann distribution is NP-hard [10]. Markov Chain Monte Carlo methods allow sampling from Boltzmann machines [11], yet obtaining reliable high-quality samples continues to be a major challenge [12].

Quantum Computing (QC) [13] offers a remedy: due to the phenomena of *superposition*, *entanglement* and the probabilistic nature of QC, it is used for efficiently sampling from distributions with exponentially-sized spaces and highly-correlated variable structures [14]. Quantum Annealers (QAs) [15–17] and gate-based quantum devices designed to prepare Gibbs states [18–20] have been proposed as quantum samplers for Boltzmann distributions. Thus, QC could accelerate BM training and prediction, thereby enhancing their practical relevance. Furthermore, expressiveness of BMs is increased by introducing quantum terms into their energy function, leading to Quantum BMs (QBM). While gradient computation is intractable for arbitrary connectivity [21], restricting the structure yields trainable models [22–24]. Classical visible units enable efficient gradient computation, yielding Semi-QBMs (SQBM) that surpass their classical counterparts in expressiveness [24].

$Q$ -learning is among the most sample-efficient RL methods [25], but its applicability is limited to discrete action spaces. Proof-of-concept studies have demonstrated that quantum annealers can improve sample efficiency in this setting by leveraging SQBMs [26, 27]. Extending  $Q$ -learning to continuous actions is notoriously unstable, as it requires global maximization over a non-linear function approximator, an intractable problem. To overcome this, a hybrid Actor-Critic (AC) framework with an SQBM-based critic was proposed [28]. While AC methods naturally support continuous actions, they continue to struggle with action constraints, training instability, and sample inefficiency. Moreover, these approaches often lack solid theoretical grounding and rely on approximations [27, 29], which can substantially degrade performance.

This work is motivated by these challenges and the lack of a framework for continuous-valued visible units for SQBMs. We address the limitations of energy-based learning together with the instability of continuous-action RL, thereby advancing sample-efficient control methods for complex physical systems such as those at CERN and beyond. Our main contributions are of theoretical nature and are summarized as follows (see also Fig. 1):

- We propose the first theoretically sound SQBM formulation with continuous visible units, called continuous SQBM (CSQBM), which largely reduces the number of required qubits. Our formulation naturally extends classical continuous-valued BMs,
- Regarding the limitations of AC algorithms, we present the first continuous  $Q$ -learning algorithm based on sampling from the hybrid quantum-classical probability distribution of a CSQBM for obtaining the best action, overcoming recent assumptions on the expressiveness of the chosen  $Q$ -function approximation while maintaining sample efficiency.

## 2 Related Work

Quantum RL (QRL) comprises several distinct approaches [30]. At the most fundamental level, fully quantum formulations and subroutine-based methods (e.g., amplitude amplification [31]) promise asymptotic advantages but typically require fault-tolerant hardware. In contrast, variational approaches [32] are better suited to near-term devices, replacing classical function approximators with parameterized quantum circuits, though they face challenges such as barren plateaus and noise. Finally, energy-based models, in particular quantum Boltzmann machines (QBM), exploit quantum sampling for richer representations and are able to capture multi-modalities.

By employing the free energy (FE) of a QBM to approximate the  $Q$ -value, prior works [1, 26, 27, 33] demonstrated improved sample efficiency compared to classical neural networks (NNs) in prototypical environments with discrete action spaces. An extension to continuous-valued environments was investigated in [28], where an AC approach was applied to the AWAKE beam line at CERN. This method represents continuous states and actions by encoding them into Bernoulli-distributed binary visible units, thereby compromising the theoretical soundness of the BM model. This forces reliance on finite-difference approximations for gradient computation over the visible units, a procedure that is both computationally inefficient and prone to numerical instability.

Several extensions of Boltzmann machines to continuous visible domains have been developed to better accommodate real-valued data. Prominent examples include Gaussian-Bernoulli models [34, 35] and more general formulations based on exponential-family distributions [36, 37]. While these approaches enhance expressiveness, they suffer from training instabilities caused by sampling difficulties. To address this, a continuous-valued QBM has been proposed [38], leveraging imaginary-time evolution on photonic quantum hardware for more efficient sampling. Although promising, this method is inherently tied to that specific platform and does not readily extend to other QC paradigms.

In continuous environments, AC algorithms often achieve state-of-the-art performance, yet they suffer from persistent challenges such as handling action constraints, training instability, and, most critically, sample inefficiency. By contrast,  $Q$ -learning [25] is highly sample-efficient but restricted to low-dimensional discrete action spaces, since it requires global maximization over the input of an NN—an intractable task due to high non-linearity [39]. While reformulations of the maximization step have been investigated [40, 41], their scalability remains uncertain. Sampling-based methods can efficiently approximate local optima [42–44], but they risk unstable or divergent training. Alternatively, analytic solutions for the global optimum are possible under strong structural assumptions on the NN [45–47], though at the cost of severely limiting representational power.

Our approach addresses these limitations by introducing CSQBM, which combine a theoretically sound hybrid quantum–classical distribution: an exponential-family prior over the visible units and a quantum Boltzmann distribution over the hidden units. Gradients with respect to the visible units can be computed analytically, making the model directly applicable within AC frameworks. In addition, we propose a continuous  $Q$ -learning algorithm based on CSQBM that enables efficient sampling for global  $Q$ -value maximization.

### 3 Background

For notational convenience, we denote vectors/matrices by bold lower/upper case letters and the identity matrix as  $\mathbf{I}$ . Further,  $\text{tr}[\cdot]$  denotes the trace of a matrix.

#### 3.1 Quantum Boltzmann Machines

A BM is a recurrent binary neural network (NN) and consists of two types of neurons: visible units  $\mathbf{v} \in \{-1, 1\}^n$  and hidden (latent) units  $\mathbf{h} \in \{-1, 1\}^m$ . The visible units are observed and encode the data, while hidden units give the model its representational power. Weighted connections between units define a quadratic energy function  $E(\mathbf{v}, \mathbf{h})$ , characterizing a Boltzmann distribution  $p(\mathbf{v}, \mathbf{h}) \propto e^{-E(\mathbf{v}, \mathbf{h})}$ . It has the capability of approximating every distribution arbitrarily well, making BMs useful for learning tasks. Since drawing exact samples from this distribution is intractable [10], one inevitably faces a trade-off between sample quality and computational efficiency.

QBM are promising in overcoming the sampling limitation. Instead of considering binary units, QBM assume every unit to be represented by a quantum bit (qubit). Instead of taking a definite value in  $\{-1, 1\}$ , the state of a qubit is represented by a 2-dimensional complex vector  $|\psi\rangle \in \mathbb{C}^2$

$$|\psi\rangle = a|0\rangle + b|1\rangle, \quad |0\rangle = (1, 0)^\top, \quad |1\rangle = (0, 1)^\top, \quad a, b \in \mathbb{C}, \quad |a|^2 + |b|^2 = 1.$$

Even though a qubit can exist in a mixture/*superposition* of the basis states  $|0\rangle$  and  $|1\rangle$  (corresponding to  $-1$  and  $1$ ), its exact state cannot be observed. The only information retrieval possible is through *measurement*, which leads the state  $|\psi\rangle$  to collapse to  $|0\rangle$  with probability  $|a|^2$  and to  $|1\rangle$  with probability  $|b|^2$ . Considering  $N$  qubits with states  $|\psi\rangle_1, \dots, |\psi\rangle_N$ , their joint quantum state  $|\psi\rangle$  is implicitly exponentially large, that is  $|\psi\rangle = |\psi\rangle_1 \otimes \dots \otimes |\psi\rangle_N \in \mathbb{C}^{2^N}$ . Through the quantum mechanical phenomenon of *entanglement*, single qubits become strongly correlated, creating an exponentially large state space that classical computers cannot efficiently simulate. Combined with their probabilistic nature, a system of  $N$  qubits allows the encoding of arbitrary discrete probability distributions over a  $2^N$ -dimensional space. Through measurement of the quantum state, QC can thus be used for efficiently obtaining samples. A more sophisticated introduction into QC is given in [13].

The energy of a quantum system can be described by a *Hamiltonian*, which is a Hermitian matrix  $\mathbf{H} \in \mathbb{C}^{2^N \times 2^N}$  ( $\mathbf{H}^\dagger = \mathbf{H}$ ) with  $N = n + m$ . The Hamiltonian of a QBM takes the form

$$\mathbf{H} = \mathbf{H}^v + \mathbf{H}^{vv} + \mathbf{H}^{vh} + \mathbf{H}^h + \mathbf{H}^{hh} = \sum_P \mathbf{H}_P^v + \mathbf{H}_P^h + \sum_Q \mathbf{H}_{PQ}^{vv} + \mathbf{H}_{PQ}^{vh} + \mathbf{H}_{PQ}^{hh}, \quad (1)$$

where  $\mathbf{H}_P^v = \sum_i w_i^v P_i$  encode the linear biases for the visible and hidden units and  $\mathbf{H}_{PQ}^{hh} = \sum_{i,j} w_{ij}^{hh} P_i P_j$  encode pairwise interactions.  $P, Q \in \{\mathbf{X}, \mathbf{Y}, \mathbf{Z}\}$  are Pauli matrices and  $P_i$  denotes applying operator  $P$  to qubit  $i$ . Choosing  $P = Q = \mathbf{Z}$  leads to a classical BM, while using more Pauli terms leads to more “quantumness”. Encoding weights into the three different Pauli basis configuration results in a significantly larger amount of weights, making QBM more expressive than classical BMs. For more details, we refer the reader to [24].

The underlying quantum Boltzmann distribution is characterized by the Gibbs state  $\rho = e^{-\beta \mathbf{H}} / \mathcal{Z} \in \mathbb{C}^{2^N \times 2^N}$ , where  $\beta > 0$  is an inverse temperature and  $\mathcal{Z} = \text{tr}[e^{-\beta \mathbf{H}}]$ . The diagonal of  $\rho$  encodes the probability of all configurations and the marginal distribution of the visible units is obtained by

$$p(\mathbf{v}) = \text{tr}[\rho_{\mathbf{v}}] = \frac{\text{tr}[e^{-\beta \mathbf{H}} \Delta_{\mathbf{v}}]}{\mathcal{Z}} = \frac{e^{-\beta F_{\mathbf{H}}(\mathbf{v})}}{\mathcal{Z}}, \quad F_{\mathbf{H}}(\mathbf{v}) = -\frac{1}{\beta} \log \text{tr}[e^{-\beta \mathbf{H}} \Delta_{\mathbf{v}}], \quad (2)$$

where  $\Delta_{\mathbf{v}}$  is a projection matrix onto the configuration  $\mathbf{v}$  and  $F_{\mathbf{H}}(\mathbf{v})$  is called the free energy of  $\mathbf{v}$ .

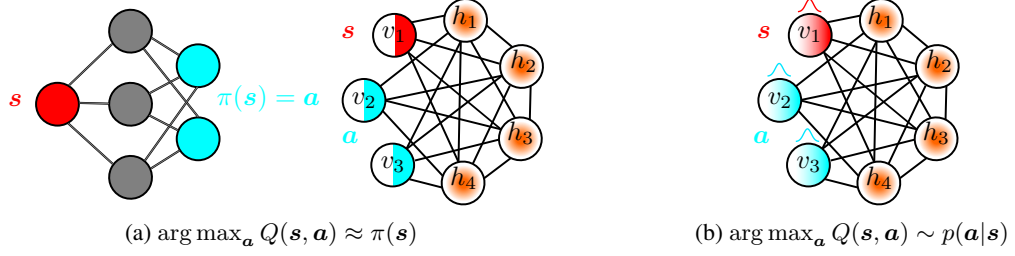


Figure 1: Illustration of our proposed CSQBM and how it enables continuous-action  $Q$ -learning. We overcome the limitations of current AC approaches using SQBMs (a) by introducing theoretically sound continuous SQBMs (CSQBMs), which utilize exponential-family priors (e.g. Gaussian) (b). The best action is obtained by sampling from the hybrid quantum–classical distribution.

Even though QBMs are powerful in theory, training them in practice is intractable, since gradients can not be computed analytically for arbitrary Hamiltonian structures. In [24], it was shown that gradients of  $\text{tr} [e^{-\beta H} \Delta_v]$  are analytically computable when only Pauli- $\mathbf{Z}$  terms are used for the visible units, since in that case  $H \Delta_v = \Delta_v H$ . This leads to a reduced number of trainable weights, however, these models were shown to still be more expressive than their classical BM counterpart.

### 3.2 Free Energy-Based Reinforcement Learning

For an in-depth description of RL and the underlying Markov Decision Processes (MDPs), we refer the reader to [48]. Generally, RL is a framework for sequential decision-making in which an agent learns to maximize the cumulative reward by interacting with an environment. Assume we are given an MDP  $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ , where  $\mathcal{S}$  is the set of states,  $\mathcal{A}$  describes the set of actions the agent can take,  $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$  describes the probability of a transition,  $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  rewards the transition and  $\gamma \in (0, 1)$  is a discount factor. Setting  $\mathbf{a}_t \sim \pi(\mathbf{s}_t)$ ,  $\mathbf{s}_{t+1} \sim P(\mathbf{s}_t, \mathbf{a}_t)$ , the goal is

$$\max_{\pi} \max_a Q^{\pi}(\mathbf{s}_0, \mathbf{a}), \quad Q^{\pi}(\mathbf{s}_0, \mathbf{a}_0) = \mathbb{E} \left[ r(\mathbf{s}_0, \mathbf{a}_0) + \sum_{t=1}^{\infty} \gamma^t r(\mathbf{s}_t, \mathbf{a}_t) \right]. \quad (3)$$

In value-based RL, the goal is to learn an accurate approximation of the action-value function  $Q$ , which in turn yields the optimal policy through the recursive Bellman optimality criterion [49]. While deep NNs are often used for approximating  $Q$  directly, free energy-based RL (FERL) approximates  $Q$  by the free energy of a SQBM. Partial derivatives are analytically computable  $\partial_x F(\mathbf{v}) = \text{tr} [\rho_{\mathbf{v}} \partial_x \mathbf{H}]$  (see Sec. A.1), which corresponds to computing the expectation value of  $\partial_x \mathbf{H}$  w.r.t. the Gibbs state  $\rho_{\mathbf{v}}$ . Since  $\partial_x \mathbf{H}$  can be easily computed due to the quadratic form in (1), we obtain the  $Q$ -learning-based weight update formula with  $Q(\mathbf{s}, \mathbf{a}) = -F(\mathbf{v})$

$$w \leftarrow w + \alpha \left[ \left( F(\mathbf{s}, \mathbf{a}) + r(\mathbf{s}, \mathbf{a}) - \gamma \min_{\mathbf{a}'} F(\mathbf{s}', \mathbf{a}') \right) \partial_w F(\mathbf{s}, \mathbf{a}) \right],$$

by encoding  $(\mathbf{s}, \mathbf{a})$  into the visible units (see Fig. 1). Simply assuming continuous-valued  $\mathbf{v}$  violates the theoretical foundations of the model, as encoding continuous variables into a quantum state and computing the corresponding marginal distribution is highly non-trivial. In the following sections, we introduce a principled solution to overcome these challenges.

## 4 Continuous Semi Quantum Boltzmann Machines

Instead of considering a joint quantum state over visible and hidden units, we consider a hybrid quantum-classical model. We assume a prior on the visible units from the exponential family, that is  $e^{c(\mathbf{v}) - A(\boldsymbol{\theta})}$ , with  $c(\mathbf{v}) = \boldsymbol{\theta}^\top \mathbf{s}(\mathbf{v}) + \log g(\mathbf{v})$  and  $A(\boldsymbol{\theta}) = \log \int_{\mathbf{v}} e^{c(\mathbf{v})} d\mathbf{v}$ , similar to [37]. This leads to the following Hamiltonian of our proposed continuous SQBM (CSQBM)

$$\mathbf{H}(\mathbf{v}) = -c(\mathbf{v})\mathbf{I} + \mathbf{H}^{vh}(\mathbf{v}) + \mathbf{H}^h + \mathbf{H}^{hh}, \quad \mathbf{H}^{vh}(\mathbf{v}) = - \sum_{ij} \sum_P w_{ij}^{vh,P} s_i(\mathbf{v}) P_j, \quad (4)$$

where the first summand replaces  $\mathbf{H}^v + \mathbf{H}^{vv}$  and the second one replaces  $\mathbf{H}^{vh}$  in (1). Note that since we do not encode the visible units into qubits anymore, our quantum state space size gets reduced from  $\mathbb{C}^{n+m}$  to  $\mathbb{C}^m$ . We obtain a conveniently computable form of the free energy.

**Theorem 1.** With  $\mathbf{H}'(\mathbf{v}) = \mathbf{H}^{vh}(\mathbf{v}) + \mathbf{H}^h + \mathbf{H}^{hh}$  and  $\rho'_{\mathbf{v}} = e^{-\beta \mathbf{H}'(\mathbf{v})} / \text{tr} [e^{-\beta \mathbf{H}'(\mathbf{v})}]$ , it holds

$$F_{\mathbf{H}}(\mathbf{v}) = -c(\mathbf{v}) + F_{\mathbf{H}'}(\mathbf{v}), \quad \partial_x F_{\mathbf{H}}(\mathbf{v}) = -\partial_x c(\mathbf{v}) + \text{tr} [\rho'_{\mathbf{v}} \partial_x \mathbf{H}'] . \quad (5)$$

A detailed proof is provided in Sec. A.2. Interestingly, we cannot only differentiate over the models' parameters in (5) but also over the visible units. This leads to the applicability in AC-algorithms, due to the need for differentiating w.r.t. the input action, while overcoming the limitations of the method presented in [28], which relies on finite differences for computing the gradient and presents a general continuous-variable alternative to the previously imposed binary Bernoulli prior.

While efficient preparation of Gibbs state is a largely open question, the authors of [50] show that this can be done in a runtime being proportional to smallest spectral gap of a quantum Markov chain. When the gap is not exponentially small, the free energy is efficiently estimated in a runtime linear in the complexity of preparing the Gibbs state.

As an example, consider an independent Gaussian prior for every visible unit, that is  $v_i \sim \mathcal{N}(\mu_i, \sigma_i)$ . The form  $e^{c(\mathbf{v}) - A(\boldsymbol{\theta})}$  is obtained by using  $\mathbf{s}(\mathbf{v})^\top = (s_1(v_1), \dots, s_n(v_n))$  and  $\boldsymbol{\theta}^\top = (\theta_1, \dots, \theta_n)$

$$s_i(v_i) = (v_i, v_i^2), \quad \boldsymbol{\theta}_i = (\mu_i/\sigma_i^2, -1/2\sigma_i^2), \quad g(\mathbf{v}) = 1 \quad \Rightarrow \quad c(\mathbf{v}) = (-v_i^2 + 2v_i\mu_i)/2\sigma_i^2 .$$

## 5 Continuous Q-Learning

Policy-based methods, such as AC algorithms, suffer from critical drawbacks, most notably sample inefficiency. In contrast, relying solely on the  $Q$ -learning update for value approximation results in a non-convex global optimization problem. Our approach is to replace the maximization step in  $Q$ -learning with sampling, motivated by the observation that  $\arg \max_{\mathbf{a}} Q(\mathbf{s}, \mathbf{a}) = \arg \max_{\mathbf{a}} e^{-\beta F(\mathbf{s}, \mathbf{a})} = \arg \max_{\mathbf{a}} p(\mathbf{a}|\mathbf{s})$ . While direct sampling from  $p(\mathbf{a}|\mathbf{s})$  remains infeasible, we instead consider sampling from the marginal distribution of the visible units.

**Theorem 2.** Given a configuration  $\mathbf{h}$  of  $\mathbf{H}(\mathbf{v})$  w.r.t. to the Pauli measurement basis  $\mathbf{P}$ , the marginal distribution of the visible units is also from the exponential family

$$p(\mathbf{v}|\mathbf{h}) = e^{c'(\mathbf{v}) - A'(\boldsymbol{\theta}')}, \quad c'(\mathbf{v}) = \boldsymbol{\theta}'^\top \mathbf{s}(\mathbf{v}) + \log g'(\mathbf{v}), \quad (6)$$

with  $\boldsymbol{\theta}' = \beta(\boldsymbol{\theta} + \mathbf{W}\mathbf{h})$ ,  $W_{ij} = w_{ij}^{vh, \mathbf{P}}$ ,  $g'(\mathbf{v}) = g(\mathbf{v})^\beta$  and  $A'(\boldsymbol{\theta}') = \int_{\mathbf{v}} e^{\boldsymbol{\theta}'^\top \mathbf{s}(\mathbf{v}) + \log g'(\mathbf{v})} d\mathbf{v}$ .

The proof is given in Sec. A.3. If the prior is efficiently samplable, so is  $p(\mathbf{a}|\mathbf{s}, \mathbf{h})$ . With the further assumption of an efficiently preparable Gibbs state of  $\mathbf{H}'$ , sampling from  $p(\mathbf{h}|\mathbf{s}, \mathbf{a})$  is also efficient. This leads to the applicability of Gibbs sampling by alternately generating samples from  $p(\mathbf{a}|\mathbf{s}, \mathbf{h})$  and  $p(\mathbf{h}|\mathbf{s}, \mathbf{a})$  to obtain a sample  $\mathbf{a} \sim p(\mathbf{a}|\mathbf{s})$ . The probability of obtaining the best actions increases with the inverse temperature  $\beta$ . Further, samples from  $p(\mathbf{a}|\mathbf{s})$  can be used for exploration, instead of relying on  $\epsilon$ -greedy strategies.

## 6 Conclusion

In this work, we introduced Continuous Semi Quantum Boltzmann Machines (CSQBM) as a theoretically grounded framework that allows for continuous-action  $Q$ -learning. By leveraging exponential family priors for continuous variables and hybrid quantum-classical sampling for action selection, our approach overcomes key limitations of existing Actor-Critic methods, achieving greater expressiveness and sample efficiency while reducing qubit requirements.

For future work, the performance of our methods will be investigated on real-world continuous-control problems—such as particle beam lines at CERN. Further it is interesting to examine the structure and expressiveness of the underlying Hamiltonians of efficiently preparable Gibbs states.

## References

- [1] Daniel Kent et al. “Using Quantum Solved Deep Boltzmann Machines to Increase the Data Efficiency of RL Agents”. In: *2024 IEEE International Conference on Quantum Computing and Engineering (QCE)*. Vol. 2. IEEE. 2024, pp. 311–316.

- [2] L Gatignon. “Physics at the SPS”. In: *Review of Scientific Instruments* 89.5 (2018).
- [3] Erik Adli et al. “Acceleration of electrons in the plasma wakefield of a proton bunch”. In: *Nature* 561.7723 (2018), pp. 363–367.
- [4] Hannes Bartosik and Giovanni Rumolo. “Performance of the LHC injector chain after the upgrade and potential development”. In: *arXiv preprint arXiv:2203.09202* (2022).
- [5] E Montbarbon et al. “The CERN East Area Renovation”. In: *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms* 461 (2019), pp. 98–101.
- [6] David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. “A learning algorithm for Boltzmann machines”. In: *Cognitive science* 9.1 (1985), pp. 147–169.
- [7] Brian Sallans and Geoffrey E Hinton. “Reinforcement learning with factored states and actions”. In: *Journal of Machine Learning Research* 5.Aug (2004), pp. 1063–1088.
- [8] Ruslan Salakhutdinov and Geoffrey Hinton. “Deep boltzmann machines”. In: *Artificial intelligence and statistics*. PMLR. 2009, pp. 448–455.
- [9] Will Grathwohl et al. “Your classifier is secretly an energy based model and you should treat it like one”. In: *arXiv preprint arXiv:1912.03263* (2019).
- [10] Francisco Barahona. “On the computational complexity of Ising spin glass models”. In: *Journal of Physics A: Mathematical and General* 15.10 (1982), p. 3241.
- [11] Geoffrey E Hinton. “Training products of experts by minimizing contrastive divergence”. In: *Neural computation* 14.8 (2002), pp. 1771–1800.
- [12] Philip M Long and Rocco Servedio. “Restricted Boltzmann machines are hard to approximately evaluate or simulate”. In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. 2010, pp. 703–710.
- [13] Michael A Nielsen and Isaac L Chuang. *Quantum computation and quantum information*. Cambridge University Press, 2010.
- [14] Nico Piatkowski and Christa Zoufal. “Quantum circuits for discrete graphical models”. In: *Quantum Machine Intelligence* 6.2 (2024), p. 37.
- [15] Tadashi Kadowaki and Hidetoshi Nishimori. “Quantum annealing in the transverse Ising model”. In: *Physical Review E* 58.5 (1998), p. 5355.
- [16] Jeremy Liu et al. “Adiabatic quantum computation applied to deep learning networks”. In: *Entropy* 20.5 (2018), p. 380.
- [17] Vivek Dixit et al. “Training restricted boltzmann machines with a d-wave quantum annealer”. In: *Frontiers in Physics* 9 (2021), p. 589626.
- [18] Chi-Fang Chen et al. “Quantum thermal state preparation”. In: *arXiv preprint arXiv:2303.18224* (2023).
- [19] Mirko Consiglio et al. “Variational Gibbs state preparation on noisy intermediate-scale quantum devices”. In: *Phys. Rev. A* 110 (1 July 2024), p. 012445. DOI: 10.1103/PhysRevA.110.012445. URL: <https://link.aps.org/doi/10.1103/PhysRevA.110.012445>.
- [20] Cambyse Rouzé, Daniel Stilck França, and Álvaro M Alhambra. “Optimal quantum algorithm for Gibbs state preparation”. In: *arXiv preprint arXiv:2411.04885* (2024).
- [21] Mohammad H Amin et al. “Quantum boltzmann machine”. In: *Physical Review X* 8.2 (2018), p. 021050.
- [22] Eric R Anschuetz and Yudong Cao. “Realizing quantum Boltzmann machines through eigenstate thermalization”. In: *arXiv preprint arXiv:1903.01359* (2019).
- [23] Luuk Coopmans and Marcello Benedetti. “On the sample complexity of quantum Boltzmann machine learning”. In: *Communications Physics* 7.1 (2024), p. 274.
- [24] Maria Demidik et al. “Expressive equivalence of classical and quantum restricted Boltzmann machines”. In: *arXiv preprint arXiv:2502.17562* (2025).
- [25] Christopher JCH Watkins and Peter Dayan. “Q-learning”. In: *Machine learning* 8.3 (1992), pp. 279–292.
- [26] Daniel Crawford et al. “Reinforcement learning using quantum Boltzmann machines”. In: *Quantum Information and Computation* 18.1&2 (2018), pp. 51–74.
- [27] Anna Levit et al. “Free energy-based reinforcement learning using a quantum processor”. In: *arXiv preprint arXiv:1706.00074* (2017).

- [28] Michael Schenk et al. “Hybrid actor-critic algorithm for quantum reinforcement learning at cern beam lines”. In: *Quantum Science and Technology* 9.2 (2024), p. 025012.
- [29] Masuo Suzuki. “Relationship between d-dimensional quantal spin systems and (d+ 1)-dimensional ising systems: Equivalence, critical exponents and systematic approximants of the partition function and spin correlations”. In: *Progress of theoretical physics* 56.5 (1976), pp. 1454–1469.
- [30] Nico Meyer et al. “A survey on quantum reinforcement learning”. In: *arXiv preprint arXiv:2211.03464* (2022).
- [31] Gilles Brassard et al. “Quantum amplitude amplification and estimation”. In: *Contemporary Mathematics* 305 (2002), pp. 53–74.
- [32] Sofiene Jerbi et al. “Parametrized quantum policies for reinforcement learning”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 28362–28375.
- [33] Niels MP Neumann et al. “Multi-agent reinforcement learning using simulated quantum annealing”. In: *International Conference on Computational Science (ICCS)*. Springer. 2020, pp. 562–575.
- [34] KyungHyun Cho, Alexander Ilin, and Tapani Raiko. “Improved learning of Gaussian-Bernoulli restricted Boltzmann machines”. In: *Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN)*. Springer. 2011, pp. 10–17.
- [35] Jan Melchior, Nan Wang, and Laurenz Wiskott. “Gaussian-binary restricted Boltzmann machines for modeling natural image statistics”. In: *PLoS One* 12.2 (2017).
- [36] Max Welling, Michal Rosen-Zvi, and Geoffrey E Hinton. “Exponential family harmoniums with an application to information retrieval”. In: *Advances in neural information processing systems* 17 (2004).
- [37] Yifeng Li and Xiaodan Zhu. “Exponential family restricted Boltzmann machines and annealed importance sampling”. In: *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2018, pp. 1–10.
- [38] Shikha Bangar et al. “Continuous-variable quantum Boltzmann machine”. In: *Quantum Machine Intelligence* 7.1 (2025), pp. 1–15.
- [39] Guy Katz et al. “Reluplex: An efficient SMT solver for verifying deep neural networks”. In: *International conference on computer aided verification*. Springer. 2017, pp. 97–117.
- [40] Moonkyung Ryu et al. “CAQL: Continuous Action Q-Learning”. In: *Proceedings of the 8th International Conference on Learning Representations (ICLR)*. 2020.
- [41] Radu Burtea and Calvin Tsay. “Constrained continuous-action reinforcement learning for supply chain inventory management”. In: *Computers & Chemical Engineering* 181 (2024), p. 108518.
- [42] Dmitry Kalashnikov et al. “Scalable deep reinforcement learning for vision-based robotic manipulation”. In: *Conference on robot learning*. PMLR. 2018, pp. 651–673.
- [43] Riley Simmons-Edler et al. “Q-learning for continuous actions with cross-entropy guided policies”. In: *arXiv preprint arXiv:1903.10605* (2019).
- [44] Georgia Perakis and Asterios Tsiourvas. “Optimizing objective functions from trained ReLU neural networks via sampling”. In: *arXiv preprint arXiv:2205.14189* (2022).
- [45] Shixiang Gu et al. “Continuous deep q-learning with model-based acceleration”. In: *International conference on machine learning*. PMLR. 2016, pp. 2829–2838.
- [46] Anton Plaksin and Stepan Martyanov. “Continuous deep Q-learning in optimal control problems: Normalized advantage functions analysis”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 22806–22815.
- [47] Brandon Amos, Lei Xu, and J Zico Kolter. “Input convex neural networks”. In: *International conference on machine learning*. PMLR. 2017, pp. 146–155.
- [48] Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*. Vol. 1. 1. MIT press Cambridge, 1998.
- [49] Richard Bellman. “The theory of dynamic programming”. In: *Bulletin of the American Mathematical Society* 60.6 (1954), pp. 503–515.
- [50] Sofiene Jerbi et al. “Quantum enhancements for deep reinforcement learning in large spaces”. In: *PRX Quantum* 2.1 (2021), p. 010328.

## A Technical Appendices and Supplementary Material

### A.1 Gradient of Free Energy

The free energy of an SQBM Hamiltonian  $\mathbf{H}$  of the form given in (1) is defined as

$$F_{\mathbf{H}}(\mathbf{v}) = -\frac{1}{\beta} \log \text{tr} [e^{-\beta \mathbf{H}} \Delta_{\mathbf{v}}]$$

Due to  $\mathbf{H}e^{-\beta \mathbf{H}} = e^{-\beta \mathbf{H}} \mathbf{H}$  and  $\mathbf{H} \Delta_{\mathbf{v}} = \Delta_{\mathbf{v}} \mathbf{H}$ , we obtain  $\partial_x e^{-\beta \mathbf{H}} = -\beta e^{-\beta \mathbf{H}} \partial_x \mathbf{H}$  and thus

$$\partial_x F_{\mathbf{H}}(\mathbf{v}) = -\frac{1}{\beta} \frac{\partial_x \text{tr} [e^{-\beta \mathbf{H}} \Delta_{\mathbf{v}}]}{\text{tr} [e^{-\beta \mathbf{H}} \Delta_{\mathbf{v}}]} = \frac{\text{tr} [e^{-\beta \mathbf{H}} \Delta_{\mathbf{v}} \partial_x \mathbf{H}]}{\text{tr} [e^{-\beta \mathbf{H}} \Delta_{\mathbf{v}}]} = \text{tr} [\rho_{\mathbf{v}} \partial_x \mathbf{H}] .$$

### A.2 Proof of Theorem 1

Now assume that  $\mathbf{H}(\mathbf{v})$  describes the energy of a CSQBM given in (4). Theorem 1 is obtained by

$$\begin{aligned} \partial_x F_{\mathbf{H}(\mathbf{v})}(\mathbf{v}) &= -\frac{1}{\beta} \frac{\partial_x \text{tr} [e^{-\beta \mathbf{H}(\mathbf{v})}]}{\text{tr} [e^{-\beta \mathbf{H}(\mathbf{v})}]} = -\frac{1}{\beta} \frac{\partial_x \text{tr} [e^{-\beta(-c(\mathbf{v})\mathbf{I} + \mathbf{H}'(\mathbf{v}))}]}{\text{tr} [e^{-\beta(-c(\mathbf{v})\mathbf{I} + \mathbf{H}'(\mathbf{v}))}]} \\ &= -\frac{1}{\beta} \frac{\partial_x (e^{\beta c(\mathbf{v})} \text{tr} [e^{-\beta \mathbf{H}'(\mathbf{v})}])}{e^{\beta c(\mathbf{v})} \text{tr} [e^{-\beta \mathbf{H}'(\mathbf{v})}]} \\ &= -\frac{1}{\beta} \frac{\partial_x e^{\beta c(\mathbf{v})} \text{tr} [e^{-\beta \mathbf{H}'(\mathbf{v})}] + e^{\beta c(\mathbf{v})} \partial_x \text{tr} [e^{-\beta \mathbf{H}'(\mathbf{v})}]}{e^{\beta c(\mathbf{v})} \text{tr} [e^{-\beta \mathbf{H}'(\mathbf{v})}]} \\ &= \frac{-\partial_x c(\mathbf{v}) \text{tr} [e^{-\beta \mathbf{H}'(\mathbf{v})}] + \text{tr} [e^{-\beta \mathbf{H}'(\mathbf{v})} \partial_x \mathbf{H}'(\mathbf{v})]}{\text{tr} [e^{-\beta \mathbf{H}'(\mathbf{v})}]} = -\partial_x c(\mathbf{v}) + \text{tr} [\rho'_{\mathbf{v}} \partial_x \mathbf{H}'(\mathbf{v})] , \end{aligned}$$

where we used the fact that  $e^{\beta \mathbf{I} + \mathbf{H}} = e^{\beta \mathbf{I}} e^{\mathbf{H}} = e^{\beta} e^{\mathbf{H}}$  for the identity matrix  $\mathbf{I}$ .

### A.3 Proof of Theorem 2

Assume we are given a hidden configuration  $\mathbf{h} \in \{-1, 1\}^m$  w.r.t. to some Pauli basis  $\mathbf{P}$ . Computing the trace of a Hamiltonian only acting on the hidden units  $\mathbf{H}^{\mathbf{h}} + \mathbf{H}^{\mathbf{h}\mathbf{h}}$ , leads to a scalar  $\lambda(\mathbf{h})$  independent of  $\mathbf{v}$ . Thus

$$\begin{aligned} p(\mathbf{v}|\mathbf{h}) &= \frac{\text{tr} [e^{-\beta \mathbf{H}(\mathbf{v})} \Delta_{\mathbf{h}}]}{\int_{\mathbf{v}} \text{tr} [e^{-\beta \mathbf{H}(\mathbf{v})} \Delta_{\mathbf{h}}] d\mathbf{v}} = \frac{e^{-\beta(-c(\mathbf{v}) - (\mathbf{W}\mathbf{h})^\top \mathbf{s}(\mathbf{v}) + \lambda(\mathbf{h}))}}{\int_{\mathbf{v}} e^{-\beta(-c(\mathbf{v}) - (\mathbf{W}\mathbf{h})^\top \mathbf{s}(\mathbf{v}) + \lambda(\mathbf{h}))} d\mathbf{v}} \\ &= \frac{e^{\beta(\boldsymbol{\theta}^\top \mathbf{s}(\mathbf{v}) + \log g(\mathbf{v}) + (\mathbf{W}\mathbf{h})^\top \mathbf{s}(\mathbf{v}))}}{\int_{\mathbf{v}} e^{\beta(\boldsymbol{\theta}^\top \mathbf{s}(\mathbf{v}) + \log g(\mathbf{v}) + (\mathbf{W}\mathbf{h})^\top \mathbf{s}(\mathbf{v}))} d\mathbf{v}} \\ &= \frac{e^{c'(\mathbf{v})}}{\int_{\mathbf{v}} e^{c'(\mathbf{v})} d\mathbf{v}} = e^{c'(\mathbf{v}) - A'(\boldsymbol{\theta}')} , \end{aligned}$$

with  $c'(\mathbf{v}) = \boldsymbol{\theta}'^\top \mathbf{s}(\mathbf{v}) + \log g(\mathbf{v})^\beta$ ,  $\boldsymbol{\theta}' = \beta(\boldsymbol{\theta} + \mathbf{W}\mathbf{h})$  and  $W_{ij} = w_{ij}^{v\mathbf{h}, \mathbf{P}}$ .