DO-EM: Density Operator Expectation Maximization

Anonymous Author(s)
Affiliation
Address
email

Abstract

Density operators, quantum generalizations of probability distributions, are gaining prominence in machine learning due to their foundational role in quantum computing. Generative modeling based on density operator models (**DOMs**) is an emerging field, but existing training algorithms – such as those for the Quantum Boltzmann Machine – do not scale to real-world data, such as the MNIST dataset. The Expectation-Maximization algorithm has played a fundamental role in enabling scalable training of probabilistic latent variable models on real-world datasets. *In this paper, we develop an Expectation-Maximization framework to learn latent variable models defined through **DOMs** on classical hardware, with resources comparable to those used for probabilistic models, while scaling to real-world data.* However, designing such an algorithm is nontrivial due to the absence of a well-defined quantum analogue to conditional probability, which complicates the Expectation step. To overcome this, we reformulate the Expectation step as a quantum information projection (QIP) problem and show that the Petz Recovery Map provides a solution under sufficient conditions. Using this formulation, we introduce the Density Operator Expectation Maximization (DO-EM) algorithm – an iterative Minorant-Maximization procedure that optimizes a quantum evidence lower bound. We show that the **DO-EM** algorithm ensures non-decreasing log-likelihood across iterations for a broad class of models. Finally, we present Quantum Interleaved Deep Boltzmann Machines (**QiDBMs**), a **DOM** that can be trained with the same resources as a DBM. When trained with **DO-EM** under Contrastive Divergence, a **QiDBM** outperforms larger classical DBMs in image generation on the MNIST dataset, achieving a 40–60% reduction in the Fréchet Inception Distance.

1 Introduction

Recent advances in quantum hardware and hybrid quantum-classical algorithms have fueled a surge of interest in developing learning models that can operate effectively in quantum regimes [1]. Classical models rely on probability distributions; quantum systems generalize these to density operators - positive semi-definite, unit-trace operators on Hilbert spaces—that encode both classical uncertainty and quantum coherence [2]. While there is considerable progress made in quantum supervised learning, there is relatively less progress in unsuperviced learning [3].

Latent variable models (LVMs) are a cornerstone of unsupervised learning, offering a principled approach to modeling complex data distributions through the introduction of unobserved or hidden variables [4]. These models facilitate the discovery of underlying structure in data and serve as the foundation for a wide range of tasks, including generative modeling, clustering, and dimensionality reduction. Classical examples such as Gaussian Mixture Models, Factor Analysis, and Hidden Markov Models [5, 6] exemplify the power of latent variable frameworks in capturing dependencies and variability in observed data. In recent years, LVMs have formed the conceptual backbone of

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

deep generative models including Variational Autoencoders [7], Generative Adversarial Networks [8], and Diffusion-based models [9]. The EM algorithm [10, 11] has been instrumental in deriving procedures for learning latent variables models. These algorithms are often preferred over algorithms which directly maximizes likelihood.

The study of Density Operator-based Latent Variable Models (**DO-LVM**) remains in its early stages, with foundational questions around expressivity, inference, and learning still largely unexplored [12–14]. Leveraging the modeling power of **DO-LVMs** on real-world data remains a significant challenge. Existing approaches rarely scale beyond 12 visible units—limited by restricted access to quantum hardware, the exponential cost of simulating quantum systems, and the memory bottlenecks associated with representing and optimizing **DO-LVMs** on classical devices. As a result, it is currently infeasible to empirically assess whether **DO-LVMs** offer any practical advantage on real-world datasets in terms of modeling power. EM based algorithms can provide a simpler alternative to existing learning algorithms for **DO-LVMs** which directly maximizes the likelihood. However deriving such algorithms in Density operator theoretic setup is extremely challenging for a variety of reasons, Most notably there are operator theoretic inequalities, such as Jensen Inequality, which can be directly applied to derive an Evidence lower bound(ELBO) style bound for **DO-LVMs**. Precise characterization of models which are compatible with such bounds and their computational behaviour remains an important area of investigation. In this paper we bridge these research gaps by making the following contributions.

- A Density Operator Expectation-Maximization (**DO-EM**) algorithm is specified using Quantum Information Projection in Algorithm 1. **DO-EM** guarantees log-likelihood ascent in Theorem 4.4 under mild assumptions that retain a rich class of models.
- A Quantum Evidence Lower Bound (`QELBO`) for the log-likelihood is derived in Lemma 4.1 from a minorant-maximization perspective leveraging the Monotonicity of Relative Entropy.
- **DO-LVMs** are specialized to train on classical data in Section 5 using the **DO-EM** algorithm. This specialization we call **CQ-LVMs**, a class of models with quantum latent variables, can train real world data due to a decomposition proved in Theorem 5.1.
- Quantum-interleaved deep Boltzmann machines (`QiDBM`), a quantum analog of the `DBM` is defined in Section 5.1. The well known Contrastive Divergence (CD) algorithm for Boltzmann machines is adapted to the `QiDBM`, which when used with **DO-EM** algorithm in Section 5.1, allows `QiDBM`s to be trained on MNIST-scale data.
- First empirical evidence of a modeling advantage when training **DO-LVMs** on standard computers with real-world data is provided in Section 6. `QiDBM`s trained using CD on the MNIST dataset achieve a 40–60% lower Fréchet Inception Distance compared to state-of-the-art deep Boltzmann machines.

2 Preliminaries

Notation The ℓ^2-norm of a column vector $\mathbf{v}$ in a Hilbert space $\mathcal{H}$ is given by $||\mathbf{v}||_2 = \sqrt{\mathbf{v}^\dagger \mathbf{v}}$ where $\mathbf{v}^\dagger$ denotes the conjugate transpose of $\mathbf{v}$. The set of Hermitian (self-adjoint) operators $\mathcal{O} = \mathcal{O}^\dagger$ on $\mathcal{H}$ is denoted by $\mathfrak{L}(\mathcal{H})$. The positive-definite subset of $\mathfrak{L}(\mathcal{H})$ is denoted by $\mathfrak{L}_+(\mathcal{H})$. The Kronecker product between two operators is denoted $A \otimes B$ and their direct sum is denoted $A \oplus B$ [15]. The identity operator on $\mathcal{H}$ is denoted $\mathrm{I}_{\mathcal{H}}$. The null space of an operator $A \in \mathcal{H}$ is denoted by $\ker(A)$.

Latent variable models and EM algorithm Latent Variable Models (LVMs) [4] specify the probability distribution of random variables $V = [V_1, \ldots, V_{d_\mathrm{V}}]$ through a joint probability model

$$P(V = \mathrm{v} \mid \theta) = \sum_h P(V = \mathrm{v}, H = \mathrm{h} \mid \theta)$$

where $H = [H_1, \ldots, H_{d_\mathrm{L}}]$ are unobserved random variables. Learning an LVM from data, a problem of great interest in Unsupervised Learning [5], refers to estimating the model parameters θ from a dataset $\mathcal{D} = \{\mathrm{v}^{(1)}, \ldots, \mathrm{v}^{(N)}\}$ consisting of i.i.d instances drawn from the LVM. Maximum likelihood-based methods aim to maximize $\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^N \ell_i(\theta)$ where $\ell_i(\theta) = \log P(V = \mathrm{v}^{(i)} \mid \theta)$. The maximization problem is not only intractable in most cases but even gradient-based algorithms, which can only discover local optima, are difficult to implement because of unwieldy computations in $\ell_i(\theta)$. The EM algorithm [10, 11] is an alternative iterative algorithm with the scheme

$$\theta^{(k+1)} = \underset{\theta}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} Q_i(\theta \mid \theta^{(k)}), \text{ where } \ell_i(\theta) \geq Q_i(\theta|\theta^{(k)}) \text{ and } \ell_i(\theta^{(k)}) = Q_i(\theta^{(k)}|\theta^{(k)}).$$

Boltzmann machines Boltzmann Machines (BM) are stochastic neural networks that define a probability distribution over binary vectors based on the Ising model in statistical physics [16]. Due to the intractability of learning in fully connected BMs, the Restricted Boltzmann Machine (RBM) was introduced with no intra-layer connections, enabling efficient Gibbs sampling [17–19]. Deep Boltzmann Machines (DBM) [20] stacks RBMs uisng undirected connections and allow for joint training of all layers. The joint probability of a DBM with L layers, $P(\mathbf{v}, \mathbf{h}^1, \ldots, \mathbf{h}^L)$ is defined as

$$P(\mathbf{v}, \mathbf{h}_1, \ldots, \mathbf{h}_{d_\mathrm{L}}) = \frac{1}{Z} e^{-E(\mathbf{v}, \mathbf{h}_1, \ldots, \mathbf{h}_{d_\mathrm{L}})} \tag{DBM}$$

where $E(\mathbf{v}, \mathbf{h}^1, \ldots, \mathbf{h}^L)$ is called the *Energy Function*, and $Z = \sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v}, \mathbf{h}^1, \ldots, \mathbf{h}^L)}$ is the *Partition Function* which is typically intractable to compute. Learning in DBMs is difficult due to intractable posterior dependencies. DBMs are usually trained using variants of the Contrastive Divergence algorithm [18, 21, 22]. A detailed discussion on Boltzmann machines and the Contrastive Divergence algorithm is provided in the Appendix A.

2.1 Density operators

A density operator on a Hilbert space $\mathcal{H}$ is a Hermitian, positive semi-definite operator with unit trace [2, 23]. The set of Density operators will be denoted by $\mathcal{P}(\mathcal{H})$, and can be regarded as generalizations of probability distributions. A joint density operator $\rho \in \mathcal{P}(\mathcal{H}_A \otimes \mathcal{H}_B)$ can be *marginalized* to $\rho_A \in \mathcal{P}(\mathcal{H}_A)$ by the partial trace operation $\rho_A = \mathrm{Tr}_B(\rho) = \sum_{i=1}^{d_B} (I_A \otimes \mathbf{x}_i^\dagger) \rho (I_A \otimes \mathbf{x}_i)$ where $\{\mathbf{x}_i\}_{i=1}^{d_B}$ is an orthonormal basis of $\mathcal{H}_B$. Such a ρ is *separable* if it is a convex combination of *product states* $\rho_A \otimes \rho_B$ with $\rho_A \in \mathcal{P}(\mathcal{H}_A)$ and $\rho_B \in \mathcal{P}(\mathcal{H}_B)$.

Definition 2.1 (Umegaki [24] Relative Entropy)**.** Let ω and ρ be density operators in $\mathcal{P}(\mathcal{H}_A \otimes \mathcal{H}_B)$ with $\ker(\rho) \subseteq \ker(\omega)$. Their relative entropy is given by $\mathrm{D_U}(\omega, \rho) = \mathrm{Tr}(\omega \log \omega) - \mathrm{Tr}(\omega \log \rho)$.

Lindblad [25] showed that the relative entropy does not increase under the action of the parital trace.

Theorem 2.2 (Monotonicity of Relative Entropy)**.** *For density operators ω and ρ in $\mathcal{P}(\mathcal{H}_A \otimes \mathcal{H}_B)$ such that* $\ker(\omega) \subset \ker(\rho)$, $\mathrm{D_U}(\omega, \rho) \geq \mathrm{D_U}(\mathrm{Tr}_B \omega, \mathrm{Tr}_B \rho)$.

Petz [26, 27] showed that Theorem 2.2 is saturated if and only if the Petz Recovery Map reverses the partial trace operation.

Definition 2.3 (Petz Recovery Map)**.** For a density operator ρ in $\mathcal{P}(\mathcal{H}_A \otimes \mathcal{H}_B)$, the Petz Recovery Map *for the partial trace* $\mathcal{R}_\rho : \mathcal{H}_A \to \mathcal{H}_A \otimes \mathcal{H}_B$ is the map

$$\mathcal{R}_\rho(\omega) = \rho^{1/2} \left(\left(\rho_A^{-1/2} \omega \rho_A^{-1/2} \right) \otimes \mathrm{I}_B \right) \rho^{1/2}. \tag{PRM}$$

Theorem 2.4 (Ruskai's condition)**.** *For density operators ω and ρ in $\mathcal{P}(\mathcal{H}_A \otimes \mathcal{H}_B)$ such that* $\ker(\omega) \subset \ker(\rho)$, $\mathrm{D_U}(\mathrm{Tr}_B \omega, \mathrm{Tr}_B \rho) = \mathrm{D_U}(\omega, \rho)$ *if and only if* $\log \omega - \log \rho = (\mathrm{Tr}_B \omega - \mathrm{Tr}_B \rho) \otimes \mathrm{I}_B$.

Ruskai's condition can be interpreted as ω and ρ having the same Conditional Amplitude Operator.

Definition 2.5 (Conditional Amplitude Operator[28])**.** The conditional amplitude operator of a density operator ρ in $\mathcal{P}(\mathcal{H}_A \otimes \mathcal{H}_B)$ with respect to $\mathcal{H}_A$ is $\rho_{B|A} = \exp(\log \rho - \log \rho_A \otimes \mathrm{I}_B)$.

A detailed discussion on density operators and quantum channels is provided in Appendix B.

3 Density operator latent variable models

In this section, we introduce Density Operator Latent Variable Models (**DO-LVM**) and recover existing models such as the Quantum Boltzmann Machine (QBM) as special cases. We discuss the computational challenges of learning such models from observations.

Definition 3.1 (**DO-LVM** and the Learning Problem)**.** A Density Operator Latent Variable Model (**DO-LVM**) specifies the density operator $\rho_{\mathrm{V}} \in \mathcal{P}(\mathcal{H}_{\mathrm{V}})$ on observables in $\mathcal{H}_{\mathrm{V}}$ through a joint density operator $\rho_{\mathrm{VL}} \in \mathcal{P}(\mathcal{H}_{\mathrm{V}} \otimes \mathcal{H}_{\mathrm{L}})$ as $\rho_{\mathrm{V}} = \mathrm{Tr}_{\mathrm{L}}\left(\rho_{\mathrm{VL}}(\theta)\right)$ where the space $\mathcal{H}_{\mathrm{L}}$ is not observed. Learning a **DO-LVM** is the estimation of model parameters θ when a target density operator $\eta_{\mathrm{V}} \in \mathcal{P}(\mathcal{H}_{\mathrm{V}})$ is specified. This can be achieved by maximizing the log-likelihood

$$\mathcal{L}(\theta) = \mathrm{Tr}\left(\eta_{\mathrm{V}} \log \rho_{\mathrm{V}}(\theta)\right). \tag{LP}$$

Remark 3.2. Maximizing the log-likelihood of a **DO-LVM** is equivalent to minimzing $\mathrm{D}_{\mathrm{U}}(\eta_{\mathrm{V}}, \rho_{\mathrm{V}}(\theta))$.

We specialize **DO-LVMs** to classical datasets in Section 5.

Hamiltonian-based models The Hamiltonian is a Hermitian operator $\mathrm{H} \in \mathfrak{L}(\mathcal{H})$ representing the total energy and generalizes the notion of an energy function in classical energy-based models. The model is defined using Gibbs state density matrix analogous to the Boltzmann distribution: $\rho(\theta) = \frac{\exp(\mathrm{H}(\theta))}{Z(\theta)}$ with $Z(\theta)$=$\mathrm{Tr}\exp(\mathrm{H}(\theta))$ and $\mathrm{H}(\theta)$=$\sum_r \theta_r \mathrm{H}_r$, where $\mathrm{H}_r \in \mathfrak{L}(\mathcal{H})$ are Hermitian operators and $\theta_r \in \mathbb{R}$ are model parameters. The Quantum Boltzmann Machine is a Hamiltonian-based model inspired by the transverse field Ising model [12]. In this paper, $\mathtt{QBM}_{\mathtt{m,n}}$ denotes a model with m visible and n hidden units with

$$\mathrm{H}(\theta) = -\sum_{i=1}^{m+n} b_i \sigma_i^z - \sum_{i>j} w_{ij} \sigma_i^z \sigma_j^z - \sum_{i=1}^{m+n} \Gamma_i \sigma_i^x \tag{QBM}$$

where σ_i^z and σ_i^x are $2^{m+n} \times 2^{m+n}$ Pauli matrices defined by $\sigma_i^k = \otimes_{j=1}^{i-1} \mathrm{I} \otimes \sigma^k \otimes_{j=i+1}^{m+n} \mathrm{I}$ where $k \in \{x, z\}$, σ^z=$\left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$, and σ^x=$\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$. A `QBM` is hence a **DO-LVM** with $\rho_{\mathrm{V}}(\theta) = \frac{1}{Z(\theta)} \mathrm{Tr}_{\mathrm{L}} \exp(\mathrm{H}(\theta))$.

Setting $\Gamma_i = 0$ recovers the Boltzmann Machine (BM) [12]. However, the density operator representation of these classical models are plagued by their $2^{m+n} \times 2^{m+n}$ dimensionality. The memory requirements for storing and updating models represented by density operators have been prohibitive for `QBMs` to scale beyond about 12 visible units.

Need for an EM algorithm. As probabilistic LVMs are a special case of **DO-LVMs**, the training challenges they face persist in **DO-LVMs**, which also introduce new operator-theoretic difficulties. Maximizing the log-likelihood of a **DO-LVM** involves operators that do not commute [13]. The direct computation of gradient in Equation (LP) is significantly complicated by the partial trace [29]. Due to the difficulty of working with hidden units, recent work on `QBM`s have focused on models without hidden units [30, 14, 31, 32]. Demidik et al. [33] studied a Restricted `QBM` with 12 visible units and 90 hidden units, the largest model studied in literature so far. Refer Appendix B for a detailed survey on `QBM` literature. Hence, training a `QBM`, the most popular **DO-LVM** in literature, on real-world data *remains an open challenge*.

Intractability of the gradient of the log-likelihood in probabilistic LVMs is addressed by the EM algorithm. Classical derivations of the EM algorithm fail with density operators since there is no well-defined way to construct conditional density operators [23]. An EM algorithm for density operators using Conditional Amplitude Operators (`CAO`) was conjectured in Warmuth and Kuzmin [34]. This is insufficient since the `CAO` does not provide a density operator [28]. In the next section, we appeal to well-known results in quantum information theory to derive an ELBO and EM algorithm for density operators.

4 The DO-EM framework

In this section, we develop an algorithmic framework applicable for learning **DO-LVMs** using a density operator expectation maximization framework.

The classical ELBO is derived for each datapoint using conditional probability and Jensen's inequality. This approach fails for density operators due to the absence well-defined quantum conditional probability [23]. In order to derive an ELBO for **DO-LVMs**, we resort to an approach inspired by the chain rule of KL-divergence [35].

Lemma 4.1 (Quantum ELBO)**.** *Let $\mathcal{J}(\eta_{\mathrm{V}}) = \{\eta \mid \eta \in \mathcal{P}(\mathcal{H}_{\mathrm{V}} \otimes \mathcal{H}_{\mathrm{L}})$ & $\mathrm{Tr}_{\mathrm{L}}\eta = \eta_{\mathrm{V}}\}$ be the set of feasible extensions for a target $\eta_{\mathrm{V}} \in \mathcal{P}(\mathcal{H}_{\mathrm{V}})$. Then for a **DO-LVM** $\rho(\theta)$ and $\eta \in \mathcal{J}(\eta_{\mathrm{V}})$,*

$$\mathcal{L}(\theta) \geq \mathrm{QELBO}(\eta, \theta) = \mathrm{Tr}(\eta \log \rho(\theta)) + S(\eta) - S(\eta_{\mathrm{V}}). \tag{QELBO}$$

Proof sketch: We provide a proof due to Theorem 2.2 in Appendix C.

The classical EM algorithm is a consequence of the ELBO being a minorant of the log-likelihood [36, 37]. However, it is well known that Theorem 2.2 is often not saturated [38–42]. Inspired by an information geometric interpretation of the EM algorithm [43], we study an instance of a quantum information projection problem to saturate `QELBO`.

4.1 A quantum information projection problem

In this subsection we study the I-projection [35] problem for density operators and show conditions when (`PRM`) can solve this problem. The problem of Quantum Information Projection (`QIP`) is stated as follows. Consider a density operator ω in $\mathcal{P}(\mathcal{H}_A)$ and a density operator ρ in $\mathcal{P}(\mathcal{H}_A \otimes \mathcal{H}_B)$, find ξ^* in $\mathcal{P}(\mathcal{H}_A \otimes \mathcal{H}_B)$ such that

$$\xi^* = \underset{\mathrm{Tr}_B(\xi)=\omega}{\mathrm{argmin}} \ \mathrm{D_U}(\xi, \rho). \tag{QIP}$$

To the best of our knowledge, this problem has not been studied in literature. We know from Theorem 2.2 that the theoretical minimum attained by the objective function in `QIP` is $\mathrm{D_U}(\omega, \mathrm{Tr}_B \rho)$ though it is not always saturated. Inspired by this connection, we explore sufficiency conditions for when `PRM` solves `QIP`.

Definition 4.2 (**Condition S**)**.** Two density operators ω in $\mathcal{P}(\mathcal{H}_A)$ and ρ in $\mathcal{P}(\mathcal{H}_A \otimes \mathcal{H}_B)$ satisfy the sufficiency condition if ρ is full rank, separable, and $[\omega, \mathrm{Tr}_B(\rho)] = 0$.

Theorem 4.3. *Suppose two density operators ω in $\mathcal{P}(\mathcal{H}_A)$ and ρ in $\mathcal{P}(\mathcal{H}_A \otimes \mathcal{H}_B)$ such that **Condition S** is satisfied, the solution to the information projection problem* `QIP` *is* `PRM`*.*

Proof sketch: The statement holds due to the fact that $[\rho, \mathcal{R}_\rho(\omega)] = 0$ under the conditions in the theorem. Thus, ρ and $\mathcal{R}_\rho(\omega)$ obey Ruskai's condition. A detailed proof is provided Appendix C.

4.2 DO-EM through the lens of Minorant Maximization

In this section, we present the **D**ensity **O**perator **E**xpectation **M**aximization (**DO-EM**) algorithm from a Minorant-Maximization perspective and discuss its advantages over direct maximization of the log-likelihood. We prove that the **DO-EM** algorithm can achieve log-likelihood ascent at every iteration under **Condition S**.

For a fixed $\theta^{(\mathrm{old})}$, the `QELBO` is maximized when η is the `QIP` of $\rho(\theta)$ onto the set of feasible extensions. This allows us to define a potential minorant $\mathcal{Q}$ for the log-likelihood.

$$\eta(\theta^{(\mathrm{old})}) = \underset{\mathrm{Tr_L}\eta=\eta_\mathrm{V}}{\mathrm{argmin}} \ \mathrm{D_U}(\eta, \rho(\theta^{(\mathrm{old})}))$$
$$\mathcal{Q}(\theta; \theta^{(\mathrm{old})}) = \mathrm{QELBO}(\eta(\theta^{(\mathrm{old})}), \rho(\theta))$$

Algorithm 1 DO-EM

1: **Input:** Target density operator η_V and $\theta^{(0)}$
2: **while** not converged **do**
3: **E Step:** $\eta^{(t)} = \underset{\eta:\mathrm{Tr_L}\eta=\eta_\mathrm{V}}{\mathrm{argmin}} \ \mathrm{D_U}(\eta, \rho(\theta^{(t)}))$
4: **M Step:** $\theta^{(t+1)} = \underset{\theta}{\mathrm{argmax}} \ \mathrm{Tr}(\eta^{(t)} \log \rho(\theta))$

We use $\mathcal{Q}$ to define the **DO-EM** algorithm in Algorithm 1. Models and QIPs that obey Ruskai's condition provably achieve log-likelihood ascent under the **DO-EM** procedure.

Theorem 4.4 ($\mathcal{Q}$ is a minorant)**.** *Let η_V be a target density matrix and $\rho(\theta)$ be a **DO-LVM** trained by the **DO-EM** algorithm. If $\rho(\theta^{(t)})$ and its* `QIP` *onto the set of feasible extensions, $\eta^{(t)}$, obey Ruskai's condition, then $\mathcal{Q}$ is a minorant of the log-likelihood. Then, $\mathcal{L}(\theta^{(t+1)}) \geq \mathcal{L}(\theta^{(t)})$, where $\theta^{(t+1)} = \mathrm{argmax}_\theta \ \mathcal{Q}(\theta; \theta^{(t)})$.*

Proof sketch: Proof using the saturation of Theorem 2.2 is in Appendix C.

Corollary 4.5. *For a target density operator η_V and model $\rho(\theta)$ satisfying **Condition S**, the E step is the Petz recovery map $\mathcal{R}_\rho(\eta_\mathrm{V})$. Moreover, such a model trained using the **DO-EM** algorithm achieves provable likelihood ascent at every iteration.*

Proof sketch: The proof due to Theorem 4.3 and Theorem 4.4 is given in Appendix C.

The **DO-EM** algorithm can be considered a density operator analog of the classical EM algorithm. We recover the classical EM algorithm from **DO-EM** for discrete models if η_V and $\rho(\theta)$ are diagonal.

The **E Step** in **DO-EM** finds a feasible extension η whose Conditional Amplitude Operator (`CAO`) is equal to that of the model $\rho(\theta)$. The `PRM` under **Condition S** is the `CAO` reweighted by η_{V} to give a valid density operator. This reduces to classical E step when the `CAO` reduces to the conditional probability and `PRM` reduces to Bayes rule. If the model ρ is of the form $\rho_{\mathrm{V}} \otimes \rho_{\mathrm{L}}$, we recover the conjecture in [34].

A log-likelihood involving a partial trace is often intractable. The **M Step** in **DO-EM** algorithm maximizes an expression without the partial trace. The log-likelihood of such expressions may have closed-form expressions for the gradients, for example, using the Lee-Trotter-Suzuki formula [14]. In the classical case, this is equivalent to the EM algorithm maximizing a sum of logarithms instead of a logarithm of sums.

Corollary 4.6. *For a Hamiltonian-based model with E step solution $\eta^{(t)}$, the M step reduces to*

$$\theta^{(t+1)} = \operatorname{argmax}_{\theta} \operatorname{Tr}(\eta^{(t)} \mathrm{H}(\theta)) - \log Z(\theta)$$

Proof sketch: The proof due to properties of the matrix logarithm is given in Appendix C.

However, the memory footprint of **DO-LVM**s remain, preventing the application of these models on real-world data. We specialize **DO-LVM**s and **DO-EM** to train on classical data and achieve practical scale.

5 DO-EM for classical data

In this section, we specialize **DO-LVMs** and the **DO-EM** algorithm to classical datasets. We assume, for ease of presentation, that the data $\mathcal{D} = \{\mathrm{v}^{(1)}, \ldots, \mathrm{v}^{(N)}\}$ is sampled from the set $\mathcal{B} = \{+1, -1\}^{d_{\mathrm{V}}}$.We consider a $2^{d_{\mathrm{V}}}$-dimensional Hilbert space $\mathcal{H}_{\mathrm{V}}$ with standard basis $\mathfrak{B} = \{\mathbf{v}_i\}_{i=1}^{2^{d_{\mathrm{V}}}}$. There is a one-to-one mapping between elements of $\mathcal{B}$ and $\mathfrak{B}$. For any dataset $\mathcal{D}$, there is an equivalent dataset on $\mathcal{H}_{\mathrm{V}}$ given by $\mathfrak{D} = \{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(N)}\}$. The target density operator is then $\eta_{\mathrm{V}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{v}_i \mathbf{v}_i^{\dagger}$. A **DO-LVM** on d_{V}-dimensional binary data is therefore a $2^{d_{\mathrm{V}}+d_{\mathrm{L}}} \times 2^{d_{\mathrm{V}}+d_{\mathrm{L}}}$ matrix while the target η_{V} is a $2^{d_{\mathrm{V}}} \times 2^{d_{\mathrm{V}}}$ matrix.

Specializing **Condition S** to diagonal target density operators, allows the decomposition of a **DO-LVM** into direct sums of smaller subspaces, making the **DO-EM** algorithm computationally easier.

Theorem 5.1. *If ρ_{V} is diagonal, ρ is separable if and only if $\rho = \oplus_i \rho_{\mathrm{L}}(i)$ and $P(\mathbf{v}_i) = \operatorname{Tr}(\rho_{\mathrm{L}}(i))$ with $\mathbf{v}_i \in \mathfrak{B}$. The density operator for $\mathcal{H}_{\mathrm{L}}$ for a particular $\mathbf{v}_i$ is then given by $\frac{1}{P(\mathbf{v}_i)}\rho_{\mathrm{L}}(i)$.*

Proof sketch: See Appendix C.

We call models that obey Theorem 5.1 as **CQ-LVMs** since it implies a classical visible probability distribution with a quantum hidden space. `QELBO` can be specialized to each data point for **CQ-LVMs**.

Lemma 5.2. *For diagonal η_{V} in $\mathcal{P}(\mathcal{H}_{\mathrm{V}})$, a **DO-LVM** $\rho(\theta)$ satisfies **Condition S** if and only if it is of the form in Theorem 5.1. The log-likelihood of these models can then expressed as $\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \ell_i(\theta)$ where $\ell_i(\theta) = \log P(\mathbf{v}^{(i)} \mid \theta)$.*

Proof sketch: The proof is an application of Theorem 5.1 and is given in Appendix C.

The decomposition of the log-likelihood into terms for each datapoint, allows the training of models on real-world data since the target densit operator η_{V} does not have to be initialized. We now show that **CQ-LVMs** are a broad class of models that include several Hamiltonian-based models.

Corollary 5.3. *A Hamiltonian-based model $\rho(\theta) = e^{\mathrm{H}(\theta)}/Z(\theta)$ with $\mathrm{H}(\theta) = \sum_r \theta_r \mathrm{H}_r$ is a **CQ-LVMs** if and only if $\mathrm{H} = \oplus_i \mathrm{H}_i$ where H_i are Hermitian operators in $\mathfrak{L}(\mathcal{H}_{\mathrm{L}})$ and $i \in [2^{d_{\mathrm{V}}}]$.*

Proof sketch: The proof, due to the properties block diagonal matrices, is given in Appendix C. We now specialize `QELBO` and Algorithm 1 to **CQ-LVMs**.

Lemma 5.4. *For diagonal η_{V} in $\mathcal{P}(\mathcal{H}_{\mathrm{V}})$ and a **CQ-LVM** $\rho(\theta)$, the log-likelihood of a data point $\mathbf{v}^{(i)} \in \mathfrak{D}$, $\ell_i(\theta)$ is lower bounded by*

$$\ell_i(\theta) \geq \operatorname{Tr}\left(\eta_{\mathrm{L}} \log(P(\mathbf{v}^{(i)}|\theta)\rho_{\mathrm{L}}^{(i)}(\theta))\right) - \operatorname{Tr}(\sigma_{\mathrm{L}} \log \sigma_{\mathrm{L}})$$

for any density operator $\eta_{\rm L}$ in $\mathcal{P}(\mathcal{H}_{\rm L})$ with equality if and only if $\eta_{\rm L} = \rho_{\rm L}^{(i)}(\theta)$. Hence, the `PRM` *is given by $\mathcal{R}_\rho(\eta_{\rm V}) = \oplus_i P_{\mathcal{D}}(V = {\rm v}_i)\rho_{\rm L}(i \mid \theta)$.*

Proof sketch: Application of Lemma 5.4 to Lemma 4.1. Proof is given in Appendix C.

This allows us to specialize Algorithm 1 to Algorithm 2, enabling the implementation of **DO-EM** without being restricted by the dimension of $\eta_{\rm V}$. However, models such as the `QBM` remain intractable for real-world data due to the normalization term, a problem that exists in classical Boltzmann machines as well.

Algorithm 2 DO-EM for **CQ-LVM**

```
Input: Target density operator η_V and θ^(0)
while not converged do
    Q_i(θ; θ^(k)) = Tr( ρ_L^(i)(θ^(k)) e^{H^(i)(θ)} ) − log Z(θ)
    θ^(t+1) = argmax_θ (1/N) Σ_{i=1}^N Q_i(θ; θ^(k))
```

1. **Input:** Target density operator $\eta_{\rm V}$ and $\theta^{(0)}$
2. **while** not converged **do**
3. $\quad \mathcal{Q}_i(\theta;\theta^{(k)}) = \mathrm{Tr}\left(\rho_{\rm L}^{(i)}(\theta^{(k)})e^{\mathrm{H}^{(i)}(\theta)}\right) - \log Z(\theta)$
4. $\quad \theta^{(t+1)} = \mathrm{argmax}_\theta \frac{1}{N}\sum_{i=1}^{N} \mathcal{Q}_i(\theta;\theta^{(k)})$

5.1 Quantum Boltzmann Machine

In this section, we discuss the `QBM` and define variants which are amenable to implementation on high-dimensional classical data. We first describe `QBM`s that are **CQ-LVMs**.

Corollary 5.5. *A* $\mathtt{QBM_{m,n}}$ *is a* ***CQ-LVM*** *if and only if quantum terms on the visible units are zero.*

Proof sketch: The statement is true because of the structure of Pauli matrices which have entries outside the direct sum structure if and only if $i \le m$. A detailed proof can be found in Appendix C.

The class of semi-quantum models studied in Demidik et al. [33] are **CQ-LVMs**. Training such a QBM is intractable for real-world data since the free energy term, $-\log Z(\theta)$ is intractable even for classical Boltzmann machines. To achieve tractable training of `QBM`s, we introduce the **Q**uantum **I**nterleaved **D**eep **B**oltzmann **M**achine (`QiDBM`) that can be trained using Contrastive Divergence with a quantum Gibbs sampling step derived here.

A **Q**uantum **I**nterleaved **D**eep **B**oltzmann **M**achine (`QiDBM`) is a `DBM` with quantum bias terms on **non-contiguous hidden layers**. We describe the Hamiltonian of a three-layered $\mathtt{QiDBM}_{\ell,\mathtt{m},\mathtt{n}}$ with ℓ visible units and m and n hidden units respectively in the two hidden layers. For ease of presentation, the quantum bias terms are present in the middle layer.

$$\mathrm{H} = -\sum_{i=1}^{\ell+m+n} b_i\sigma_i^z - \sum_{i=1}^{\ell}\sum_{j=1}^{m} w_{ij}^{(1)}\sigma_i^z\sigma_{\ell+j}^z - \sum_{i=1}^{m}\sum_{j=1}^{n} w_{ij}^{(2)}\sigma_{\ell+i}^z\sigma_{\ell+m+j}^z - \sum_{i=1}^{m}\Gamma_i\sigma_{\ell+i}^x \qquad (\mathtt{QiDBM})$$

The quantum interleaving in a `QiDBM` is necessary to make the Gibbs sampling step tractable. We illustrate the case of the middle layer of $\mathtt{QiDBM}_{\ell,\mathtt{m},\mathtt{n}}$. If the non-quantum visible and hidden layers are fixed to $\mathbf{v}$ and $\mathbf{h}^{(2)}$, the hidden units of the quantum layer are conditionally independent. The Hamiltonian of the $i^{\rm th}$ unit of the quantum layer $\mathrm{L}^{(1)}$ is given by $\mathrm{H}^{\mathrm{L}^{(1)}}(i|\mathbf{v},\mathbf{h}^{(2)},\theta) = -b_i^{\rm eff}\sigma^z - \Gamma_i\sigma^x$. This allows for the tractable sampling from the quantum layer using the expected values

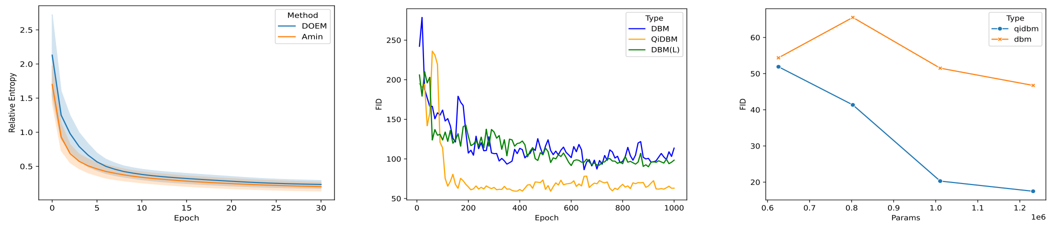
$$\langle\sigma_i^z\rangle_{\mathbf{v},\mathbf{h}^{(2)}} = \frac{b_i^{\rm eff}}{D_i}\tanh D_i \text{ and } \langle\sigma_i^x\rangle_{\mathbf{v},\mathbf{h}^{(2)}} = \frac{\Gamma_i}{D_i}\tanh D_i$$

where $D_i = \sqrt{(b_i^{\rm eff})^2 + \Gamma_i^2}$ and $b_i^{\rm eff} = b_i + \sum_{j=1}^{\ell} w_{ij}^{(1)}\mathbf{v}_j + \sum_j w_{ij}^{(2)}\mathbf{h}_j^{(2)}$. The Gibbs step for the non-quantum layers is done as per the classical CD algorithm using the quantum sample from the Z Pauli operator. This closed-form expression for Gibbs sampling without matrices allows CD to run on a `QiDBM` with the same memory footprint as a `DBM`. See Appendix C for more details.

6 Empirical evaluation

In this work, we propose a quantum model **CQ-LVM**, and a general EM framework, **DO-EM**, to learn them. In this section, we empirically evaluate our methods through experiments to answer the following questions. Details of the compute used to run all our experiments and baselines are provided in Appendix D and E.

(Q1) **Effectiveness of DO-EM.** Is Algorithm 2, a feasible algorithm for **CQ-LVM**s compared to state of the art algorithms for `QBM`s ?



(a) (b) (c)

Figure 1: (a) Relative entropy during training with exact computation of a QBM on a mixture of Bernoulli distribution. Showing that DO-EM does lead to decrease in relative entropy. (b) DBM with 6272 hidden units. QiDBM with 6273 hidden units. DBM(L) with 6273 hidden units. (c) FID scores on Binarized MNIST as a function of model parameters of QiDBM and DBM.

(Q2) **DO-EM on Real World Data.** Does Algorithm 2 scale with the to real world data?

(Q3) **Performance of DO-EM.** Does Algorithm 2 provide reasonable improvement in performance over classical LVMs?

To answer (Q1), we conduct experiments running exact computation to show that the proposed algorithm is feasible and is practical to implement.

Baselines We compare our method with our implementation of Amin et al. [12] which explores an alternate algorithm for training QBMs.

Dataset and Metrics We use a mixture of Bernoulli dataset introduced in Amin et al. [12] described in Appendix D. We measure the efficacy of our proposed method by measuring the average relative entropy during training.

Results of experiment In Figure 1a, we first observe that the relative entropy of our proposed algorithm does decrease during training, validating our theoretical results and showing, to the best our knowledge, the first instance of an expectation maximization algorithm with quantum bias. We also observe that the performance is competitive with Amin et al. [12]. We also note that **CQ-LVM** training with DO-EM is faster than Amin et al. [12] and consumes lesser memory. We provide more experiments using exact computation in Appendix D.

To answer (Q2) and (Q3), we conduct experiments on DBMs of varying sizes with and without the quantum bias term described in Section 5. We present qualitative results of our experiments in Appendix D.

Baselines. We compare our proposed method with Taniguchi et al. [22], the state of the art for training DBMs. We are unable to reproduce the results in their work and we report the results obtained from their official implementation[1] using the hyper parameters described in their work.

Datasets and Metrics Following prior work [22], we perform our experiments on MNIST and Binarized MNIST dataset [44] which contains 60,000 training images and 10,000 testing images of size 28x28. We measure the FID [45] between 10,000 generated images and the MNIST test set to assess the quality of generation. The Fréchet Inception Distance (FID) is a quantitative metric used to evaluate the quality of images generated by generative models by comparing the statistical distribution of their feature representations to those of real images.

Experiment: Performance of DO-EM To show the superior performance of the proposed method, we compare the FID of our proposed algorithm on Binarized MNIST. We train a QiDBM and DBM with 498, 588, 686, and 784 hidden units with a learning rate of 0.001 for 1000 epochs with 2 hidden layers with SGD optimizer with a batch size of 600.

Results of Experiments In Figure 1c, we observe that the proposed algorithm outperforms the DBM in all cases, achieving a minimum FID of 14.77 to the DBM's 42.61. This experiment shows that simply adding quantum bias terms to a DBM can *improve the quality* of generations by around 65%.

[1] `https://github.com/iShohei220/unbiased_dbm`

Experiment: DO-EM on High Dimensional Data We run CD on 2 DBMs without quantum bias terms according to Taniguchi et al. [22] and CD with quantum bias for a QiDBM on MNIST. Each image corresponds to 6272 visible binary units. The QiDBM has 78.70M parameters with 2 hidden layers with quantum bias added to the second layer with a hidden size of 6272. Both DBMs have 2 hidden layers and have 78.69M and 78.71M parameters and hidden sizes of 6272 and 6273 respectively. We use a learning rate of 0.001 for all experiments and train with a batch size of 600 with SGD optimizer for 1000 epochs. The purpose of this experiment is to show that it is feasible to train large models with quantum bias terms.

Results of Experiments In Figure 1b, we observe that the proposed method outperforms both classical models of similar size with a 45% reduction in FID. We observe that the FID of the model converges to this value in around 400 epochs whereas both DBM models still exhibit instability after 500 epochs. The QiDBM achieves an FID of 62.77 whereas the classical DBMs achieve an FID of 111.73 and 99.17 for the smaller and larger model respectively. This experiment indicates that scaling QiDBMs is feasible and provides a significant improvement in performance. In Appendix D, we show the qualitative differences between generated samples of the DBM and QiDBM. We observe that the generated samples from the QiDBM appear to be better than that of the DBM after only 250 epochs.

Discussion We design **CQ-LVM**s and implement Algorithm 1 to learn different target distributions. We first show that Algorithm 1 is effective in learning **CQ-LVM**s and is competitive with the state of the art in terms of reduction of relative entropy at lower running times for 10 qubits and can be extended to even 20 qubits where others cannot. Next, we see that the addition of quantum bias terms to a DBM when trained using Algorithm 2 shows superior generation quality compared to classical DBMs with a 60% reduction of FID on Binarized MNIST. Next, we show that **QiDBMs** can learn high dimensional datasets like MNIST using Algorithm 2 by scaling models upto 6272 hidden units. We observe that QiDBMs also achieve better performance, with 40% lower FID compared to DBMs of similar sizes. We also observe that QiDBMs converge in about half the amount of time compared to DBMs.

7 Discussion

The paper makes important progress by proposing **DO-EM**, an EM Algorithm for Latent Variable models defined by Density Operators, which provably achieves likelihood ascent. We propose **CQ-LVM**, a large collection of density operator based models, where **DO-EM** applies. We show that `QiDBM`, an instance of **CQ-LVM**, can easily scale to MNIST dataset which requires working with 6200+ units and outperform DBMs, thus showing that Density Operator models may yield better performance. The specification of **DO-EM** is amenable to implementation on quantum devices.

DO-EM on quantum devices The E Step of the DO-EM algorithm can be implemented on a quantum computer using the method developed by Gilyén et al. [46], where the quantum channel is performing the partial trace operation. The goal is to prepare the Petz recovery map for the partial trace channel $\eta^{(t)} = \mathcal{R}_\rho(\eta_{\mathrm{V}})$ using `PRM`. The requirements for this are (1) Quantum access to the input state η_{V} (2) efficient state preparation of the model's density matrix $\rho(\theta)$ [47, 48] and (3) Block-encodings for the model's density matrix and its marginal $\rho_{\mathrm{V}}(\theta) = \mathrm{Tr}_{\mathrm{L}}\rho(\theta)$ [49]. Given these input assumptions, the quantum algorithm implementing `PRM` consists of three steps [46]: (1) applying $\rho_{\mathrm{V}}^{-1/2}$ on the state η_{V}, (2) applying the adjoint channel which is straight-forward for the partial trace channel and can be operationally achieved by preparing subsystem L in the maximally mixed state, and (3) applying $\rho^{1/2}$ on the combined system. Both $\rho_{\mathrm{V}}^{-1/2}$ and $\rho^{1/2}$ are implemented using *Quantum Singular Value Transformation (QSVT)* techniques, leveraging block-encodings of the relevant states [49].

The M Step proceeds via gradient descent by the computation of the gradient given by $\left(\mathrm{Tr}[H_r\eta(\theta^{(t)})] - \mathrm{Tr}[H_r\rho(\theta)]\right)$ for the different terms in the Hamiltonian $H = \sum_r \theta_r H_r$ [14, 32]. The M Step stops when the gradients are small and an updated parameter $\theta^{(t+1)}$ is obtained. This two-step iterative DO-EM procedure continues until convergence. While the gradients can be estimated on existing near-term quantum devices, the E step requires careful design.

Limitations We discuss the limitations of this work in Appendix F.

References

[1] J. Preskill. Quantum computing in the nisq era and beyond. *Quantum*, 2:79, 2018.

[2] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, 2010. doi: 10.1017/CBO9780511976667.

[3] Yaswitha Gujju, Atsushi Matsuo, and Rudy Raymond. Quantum machine learning on near-term quantum devices: Current state of supervised and unsupervised techniques for real-world applications. *Phys. Rev. Appl.*, 21:067001, Jun 2024.

[4] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006. ISBN 978-0-387-31073-2.

[5] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999. doi: 10.1023/A:1007665907178.

[6] Zoubin Ghahramani. An introduction to hidden markov models and bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 15(01):9–42, 2001.

[7] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2014.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, volume 27, 2014.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

[10] Leonard E. Baum and Ted Petrie. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics*, 37(6):1554 – 1563, 1966.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1): 1–22, 1977.

[12] Mohammad H. Amin, Evgeny Andriyash, Jason Rolfe, Bohdan Kulchytskyy, and Roger Melko. Quantum boltzmann machine. *Phys. Rev. X*, 8:021050, May 2018.

[13] Mária Kieferová and Nathan Wiebe. Tomography and generative training with quantum boltzmann machines. *Phys. Rev. A*, 96:062327, 12 2017.

[14] H J Kappen. Learning quantum models from quantum or classical data. *Journal of Physics A: Mathematical and Theoretical*, 53(21):214001, 5 2020.

[15] Rajendra Bhatia. *Matrix Analysis*, volume 169. Springer, 1997. ISBN 0387948465.

[16] David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. A learning algorithm for boltzmann machines. *Cognitive Science*, 9(1):147–169, 1985. ISSN 0364-0213.

[17] P. Smolensky. Information processing in dynamical systems: Foundations of harmony theory. In *Parallel Distributed Processing, Volume 1: Explorations in the Microstructure of Cognition: Foundations*, chapter 6, pages 194–281. The MIT Press, 07 1986.

[18] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural Comput*, 14(8):1771–1800, Aug 2002.

[19] Miguel Á. Carreira-Perpiñán and Geoffrey Hinton. On contrastive divergence learning. In Robert G. Cowell and Zoubin Ghahramani, editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, volume R5 of *Proceedings of Machine Learning Research*, pages 33–40. PMLR, 06–08 Jan 2005.

[20] Ruslan Salakhutdinov and Geoffrey Hinton. Deep boltzmann machines. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 5 of *JMLR: W&CP*, pages 448–455, Clearwater Beach, Florida, USA, 2009. JMLR.

[21] Ruslan Salakhutdinov and Geoffrey E. Hinton. An efficient learning procedure for deep boltzmann machines. *Neural Computation*, 24(8):1967–2006, 2012. doi: 10.1162/NECO_a_00302.

[22] Shohei Taniguchi, Masahiro Suzuki, Yusuke Iwasawa, and Yutaka Matsuo. End-to-end training of deep boltzmann machines by unbiased contrastive divergence with local mode initialization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 33804–33815. PMLR, 23–29 Jul 2023.

[23] Mark M. Wilde. *Quantum Information Theory*. Cambridge University Press, nov 2016.

[24] H. Umegaki. Conditional expectation in an operator algebra. IV (entropy and information). *Kōdai Mathematical Seminar Reports*, 14:59–85, 1962.

[25] Göran Lindblad. Completely positive maps and entropy inequalities. *Communications in Mathematical Physics*, 40(2):147–151, Jun 1975.

[26] Dénes Petz. Sufficient subalgebras and the relative entropy of states of a von neumann algebra. *Communications in Mathematical Physics*, 105(1):123–131, Mar 1986.

[27] Dénes Petz. Sufficiency Of Channels Over von Neumann Algebras. *The Quarterly Journal of Mathematics*, 39(1):97–108, 03 1988.

[28] N. J. Cerf and C. Adami. Negative entropy and information in quantum mechanics. *Phys. Rev. Lett.*, 79:5194–5197, Dec 1997.

[29] Nathan Wiebe and Leonard Wossnig. Generative training of quantum boltzmann machines with hidden units. *arXiv preprint arXiv:1905.09902*, 2019.

[30] Onno Huijgen, Luuk Coopmans, Peyman Najafi, Marcello Benedetti, and Hilbert J. Kappen. Training quantum boltzmann machines with the β-variational quantum eigensolver. *arXiv preprint arXiv:2304.08631*, 2024.

[31] Dhrumil Patel and Mark M. Wilde. Natural gradient and parameter estimation for quantum boltzmann machines. *arXiv preprint arXiv:2410.24058*, 2024.

[32] Luuk Coopmans and Marcello Benedetti. On the sample complexity of quantum boltzmann machine learning. *Communications Physics*, 7(1):274, 2024.

[33] Maria Demidik, Cenk Tüysüz, Nico Piatkowski, Michele Grossi, and Karl Jansen. Expressive equivalence of classical and quantum restricted boltzmann machines. *arXiv preprint arXiv:2502.17562*, 2025.

[34] Manfred K. Warmuth and Dima Kuzmin. Bayesian generalized probability calculus for density matrices. *Mach. Learn.*, 78(1–2):63–101, January 2010.

[35] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, USA, 2006.

[36] David R. Hunter Kenneth Lange and Ilsoon Yang. Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, 9(1):1–20, 2000. doi: 10.1080/10618600.2000.10474858.

[37] Jan de Leeuw. Block-relaxation algorithms in statistics. In Hans-Hermann Bock, Wolfgang Lenski, and Michael M. Richter, editors, *Information Systems and Data Analysis*, pages 308–324, Berlin, Heidelberg, 1994. Springer Berlin Heidelberg. ISBN 978-3-642-46808-7.

[38] Andrew Lesniewski and Mary Beth Ruskai. Monotone riemannian metrics and relative entropy on noncommutative probability spaces. *Journal of Mathematical Physics*, 40(11):5702–5724, 11 1999.

[39] Mario Berta, Marius Lemm, and Mark M. Wilde. Monotonicity of quantum relative entropy and recoverability. *Quantum Info. Comput.*, 15(15–16):1333–1354, November 2015.

[40] Mark M. Wilde. Recoverability in quantum information theory. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2182):20150338, oct 2015.

[41] Eric A Carlen and Anna Vershynina. Recovery map stability for the data processing inequality. *Journal of Physics A: Mathematical and Theoretical*, 53(3):035204, jan 2020.

[42] Samuel S. Cree and Jonathan Sorce. Geometric conditions for saturating the data processing inequality. *Journal of Physics A: Mathematical and Theoretical*, 55(13):135301, 2022. doi: 10.1088/1751-8121/ac5648.

[43] Hideitsu Hino, Shotaro Akaho, and Noboru Murata. Geometry of em and related iterative algorithms. *Information Geometry*, 7(1):39–77, 2024.

[44] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[45] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, August 2020. Version 0.3.0.

[46] András Gilyén, Seth Lloyd, Iman Marvian, Yihui Quek, and Mark M. Wilde. Quantum algorithm for petz recovery channels and pretty good measurements. *Phys. Rev. Lett.*, 128:220502, Jun 2022.

[47] Ersen Bilgin and Sergio Boixo. Preparing thermal states of quantum systems by dimension reduction. *Phys. Rev. Lett.*, 105:170405, Oct 2010. doi: 10.1103/PhysRevLett.105.170405.

[48] Chi-Fang Chen, Michael J. Kastoryano, Fernando G. S. L. Brandão, and András Gilyén. Quantum thermal state preparation, 2023.

[49] András Gilyén, Yuan Su, Guang Hao Low, and Nathan Wiebe. Quantum singular value transformation and beyond: exponential improvements for quantum matrix arithmetics. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2019, page 193–204, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367059. doi: 10.1145/3313276.3316366.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All contributions tally with the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations discussed in Appendix F.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Assumptions are stated clearly. Proofs provided in Appendix C.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Experiments are clearly laid out in Section 6 and Appendix D. Experimental details for reproducibility are provided in Appendix E. Anonymous code is linked.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Link to anonymous code provided. Details are provided in Appendix E.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Details provided in Appendix E.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: See Appendix D.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Details provided in Appendix E.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Code of ethics followed, no interventions with living beings requiring special processing. Only standard datasets were used. No conflicts of interest.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The paper concerns an algorithm to learn density operator latent variable models and does not directly have societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Sources provided in Appendix E.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Anonymous code contains a ReadMe file.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: Core method development in this research does not involve LLMs.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.