Density Operator Expectation Maximization

Adit Vishnu   ADITVISHNU@IISC.AC.IN
Abhay Shastry   ABHAYSHASTRY@FSID-IISC.IN
Dhruva Kashyap   KDHRUVA@IISC.AC.IN
Chiranjib Bhattacharyya   CHIRU@IISC.AC.IN
Department of Computer Science and Automation
Indian Institute of Science
Bangalore, KA 560012, India

Abstract

Machine learning with density operators, the mathematical foundation of quantum mechanics, is gaining prominence with rapid advances in quantum computing. Generative models based on density operators cannot yet handle tasks that are routinely handled by probabilistic models. The progress of latent variable models, a broad and influential class of probabilistic unsupervised models, was driven by the Expectation-Maximization framework. Deriving such a framework for density operators is challenging due to the non-commutativity of operators. To tackle this challenge, an inequality arising from the monotonicity of relative entropy is demonstrated to serve as an evidence lower bound for density operators. A minorant-maximization perspective on this bound leads to Density Operator Expectation Maximization (DO-EM), a general framework for training latent variable models defined through density operators. Through an information-geometric argument, the Expectation step in DO-EM is shown to be the Petz recovery map. The DO-EM algorithm is applied to Quantum Restricted Boltzmann Machines, adapting Contrastive Divergence to approximate the Maximization step gradient. Quantum interleaved Deep Boltzmann Machines and Quantum Gaussian-Bernoulli Restricted Boltzmann Machines, new models introduced in this work, outperform their probabilistic counterparts on generative tasks when trained with similar computational resources and identical hyperparameters.

Keywords: density operators, latent variable models, expectation-maximization algorithm, quantum information theory, Quantum Boltzmann Machines

1 Introduction

Recent advances in quantum hardware and hybrid classical-quantum algorithms have sparked growing interest in machine learning models capable of operating within quantum regimes (Preskill, 2018). Quantum machine learning promises to harness quantum mechanics to enhance learning, either by achieving computational speedups on quantum devices (Kerenidis et al., 2019, 2020; Miyahara et al., 2020; Kim et al., 2023) or by modeling richer data structures through quantum representations (Schuld et al., 2015; Biamonte et al., 2017; Amin et al., 2018). The latter approach employs density operators—positive semi-definite, unit-trace operators on Hilbert spaces—that generalize the notion of probability distributions and underpin the mathematical foundations of quantum mechanics (von Neumann, 1955; Gleason, 1957). While significant progress has been made in the use of density operators in supervised learning, there is relatively less progress in the unsupervised setting (Gujju et al., 2024).

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 (Bishop, 2006). 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 (Jordan et al., 1999; Ghahramani, 2001). Classical examples such as Factor Analysis (Basilevsky, 2009), Independent Component Analysis (Bell and Sejnowski, 1995), Hidden Markov Models (Baum and Petrie, 1966), and Boltzmann Machines (Ackley et al., 1985) 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 deep generative models including Variational Autoencoders (Kingma and Welling, 2014) and Diffusion-based models (Ho et al., 2020). The evidence lower bound (ELBO) and the Expectation-Maximization (EM) algorithm (Baum and Petrie, 1966; Dempster et al., 1977) has been instrumental in deriving procedures for learning latent variables models. These techniques remain the standard approach for learning latent variable models as they offer a principled and tractable means of handling incomplete data.

Density operators were first studied as latent variable models by Warmuth (2005). There is renewed interest in density operator latent variable models (DO-LVM) due to their relevance in quantum computing, yet development remains in its early stages, with fundamental questions concerning expressivity, inference, and learning still largely unresolved (Amin et al., 2018; Kieferová and Wiebe, 2017; Kappen, 2020; Coopmans and Benedetti, 2024). Most existing learning algorithms for DO-LVMs fail to scale beyond 12-bit binary strings, restricting evaluation to toy datasets and preventing any meaningful assessment of their modeling power on real-world data. EM-based algorithms can provide a simpler alternative to existing learning algorithms for DO-LVMs which directly maximizes the likelihood. However, deriving analogous algorithms in the density-operator setting is challenging for several reasons, primarily because operator-theoretic inequalities—such as the operator Jensen inequality—cannot be directly applied to obtain an ELBO-style bound for DO-LVMs. A precise characterization of models compatible with such bounds, along with their computational behavior, remains an important direction for investigation.

1.1 Latent Variable Models

Latent variable models (LVMs) (Bishop, 2006) model the joint probability distribution of a random variable $X = (X_1, \ldots, X_{d_V})$ with parameters θ as

$$\Pr(X = x \mid \theta) = \sum_z \Pr(X = x, Z = z \mid \theta)$$

where $Z = (Z_1, \ldots, Z_{d_L})$ are unobserved *hidden variables*. Maximum likelihood-based methods estimate the parameters of an LVM by maximizing the log-likelihood

$$\mathcal{L}(\mathcal{D}, \theta) = \frac{1}{N} \sum_{i=1}^{N} \ell_i(\theta) \text{ where } \ell_i(\theta) = \log \Pr(X = x^{(i)} \mid \theta)$$

from a data set $\mathcal{D} = \{\mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(N)}\}$. However, this maximization problem is intractable in most cases and gradient-based algorithms are difficult to implement because of unwieldy computations in $\ell_i(\theta)$ due to the marginalization within the logarithm.

Boltzmann Machines (BM) are a class of LVMs inspired by the Ising model in statistical physics (Ackley et al., 1985). BM variants such as Restricted Boltzmann Machines (RBM, Smolensky, 1986), Gaussian-Bernoulli RBMs (GRBM, Welling et al., 2004), and Deep Boltzmann Machines (DBM, Salakhutdinov and Hinton, 2009) could be trained on image data sets, learning meaningful latent representations and generating coherent samples. RBMs and GRBMs were successfully applied to natural images, digits, and simple textures, while DBMs extended this capability to deeper hierarchical representations—an early precursor to modern generative models.

1.2 EM algorithm

We recall the EM algorithm used to train probabilistic latent variable models (Bishop, 2006).The evidence lower bound (ELBO),

$$\ell_i(\theta) \geq \sum_{\mathrm{z}} q_i(\mathrm{z}) \log \frac{p_\theta(\mathrm{x}^{(i)}, \mathrm{z})}{q_i(\mathrm{z})}, \tag{ELBO}$$

is a well known lower bound on the log-likelihood for any choice of a *variational distribution* $q_i(\mathrm{z})$, with equality when $q_i(\mathrm{z}) = p(\mathrm{z} \mid \mathrm{x}^{(i)})$. The Expectation-Maximization algorithm (Baum and Petrie, 1966; Dempster et al., 1977) arises by choosing $q_i(\mathrm{z}) = p_{\theta^{(\mathrm{old})}}(\mathrm{z} \mid \mathrm{x}^{(i)})$, which maximizes the ELBO with respect to the variational distribution q_i (E step), followed by maximizing it with respect to model parameters θ (M step):

$$Q_i(\theta \mid \theta^{(\mathrm{old})}) = \sum_{\mathrm{z}} p_{\theta^{(\mathrm{old})}}(\mathrm{z} \mid \mathrm{x}^{(i)}) \log \left(\frac{p_\theta(\mathrm{x}^{(i)}, \mathrm{z})}{p_{\theta^{(\mathrm{old})}}(\mathrm{z} \mid \mathrm{x}^{(i)})} \right), \tag{E step}$$

$$\theta^{(\mathrm{new})} = \arg \max_\theta \frac{1}{N} \sum_{i=1}^{N} Q_i(\theta \mid \theta^{(\mathrm{old})}). \tag{M step}$$

Here, Q_i is a *minorant* of the log-likelihood (de Leeuw, 1994; Kenneth Lange and Yang, 2000), satisfying

$$\ell_i(\theta) \geq Q_i(\theta \mid \theta^{(k)}) \text{ for all } \theta \text{ and } \ell_i(\theta^{(\mathrm{old})}) = Q_i(\theta^{(\mathrm{old})} \mid \theta^{(\mathrm{old})}), \tag{1}$$

which ensures $\mathcal{L}(\mathcal{D}, \theta^{(\mathrm{new})}) \geq \mathcal{L}(\mathcal{D}, \theta^{(\mathrm{old})})$.

The EM algorithm applied to different latent variable models has led to several unsupervised learning algorithms such as k-means clustering and the Baum-Welch algorithm for Hidden Markov Models (Bishop, 2006). Its formulation—optimizing a surrogate lower bound when direct likelihood maximization is intractable—motivated variational inference, where complex problems are turned into simpler ones by decoupling the degrees of freedom and introducing additional variational parameters that can be optimized to approximate otherwise intractable distributions (Jordan et al., 1999). The goal of this paper is to derive an EM algorithm for density operators.

1.3 Problem Statement and Contributions

Quantum Boltzmann Machines (`QBM`) are density operator extensions of BMs based on the transverse field Ising model (Amin et al., 2018). Unlike their probabilistic counterparts, `QBMs`, have yet to demonstrate comparable success in generative modeling. Their formulation remains largely theoretical and on toy data sets. Practical training methods for `QBMs` on complex data remain an open challenge. To bridge this gap, we ask the following questions:

1. **Operator theoretic challenges:** The non-commutativity of operators restricts the use of techniques from probabilistic latent variable models when deriving training algorithms for DO-LVMs. Is it possible to derive computationally efficient iterative schemes in the spirit of EM for DO-LVMs with guarantees similar to Equation (1)?

2. **Classical data sets:** Since traditional machine learning datasets are not quantum mechanical in nature, can DO-LVMs be adapted to learn from such data? Specifically, could classical-quantum states used in quantum error correction, where classical information guides the recovery of quantum states, provide a framework for addressing such machine learning tasks?

3. **Scaling to standard ML data sets:** Current methods cannot train DO-LVMs on standard image data sets such as MNIST due to algorithmic limitations. Are there variants of DO-LVMs that can handle such data? Specifically, can varaints of the classical Boltzmann Machine, such as GRBMs and DBMs, be extended to `QBMs` and trained on image data sets? If so, is there a modeling advantage to using density operator models over classical probabilistic models for learning such data?

 In this paper, we answer these questions by making the following contributions.

1. **EM algorithm for Density Operators**: The Density Operator Expectation Maximization (DO-EM) algorithm is derived in Section 6. Using the information geometric interpretation of the EM algorithm, the E step is shown to be the well known Petz Recovery Map. Furthermore, Theorem 21 guarantees log-likelihood ascent across iterations of DO-EM under mild assumptions, while still accommodating a rich class of models. The two steps of the EM algorithm are computationally more efficient than existing techniques for computing the gradient of the log-likelihood in DO-LVMs.

2. **Classical-Quantum Models:** The DO-EM algorithm is specialized to standard ML datasets in Section 7. We introduce a new class of DO-LVMs, termed Classical-Quantum LVMs (CQ-LVM), which feature classical visible states and quantum latent states. CQ-LVMs preserve likelihood ascent under the DO-EM algorithm and eliminate the dependence of DO-EM on the size of the visible sample space. Quantum counterparts of DBMs and GRBMs are provided as concrete examples of CQ-LVMs.

3. **Generative capability:** The first empirical evidence of a modeling advantage for DO-LVMs trained on classical hardware using image datasets is presented in Section 8. Under comparable computational budgets and using identical hyperparameters—without any additional tuning for the quantum variants—quantum DBMs trained on MNIST

achieve a 40-60% reduction in Fréchet Inception Distance compared to Deep Boltzmann Machines that approximate the log-likelihood gradient using contrastive divergence. Similarly, quantum GRBMs trained on Fashion-MNIST and CelebA-32 achieve a 20-30% reduction in Fréchet Inception Distance relative to Gaussian-Bernoulli Restricted Boltzmann Machines trained via contrastive divergence.

1.4 Organization

Section 2 provides the necessary background on training latent variable models. Section 3 introduces key properties of density operators along with useful trace inequalities. Section 4 reviews prior efforts to train `QBMs` and outlines the limitations of existing approaches. Section 5 develops a formal treatment of DO-LVMs and establishes the quantum evidence lower bound. Section 6 presents the DO-EM algorithm and proves its central properties. Section 7 then specializes DO-LVMs and Algorithm 1 to classical data sets and introduces new models consistent with this specialization. Finally, Section 8 provides empirical validation for the claims developed in the preceding sections.

2 Training Latent Variable Models

This section provides the relevant background on training probabilistic LVMs.

2.1 EM through the lens of the Data Processing Inequality

The Expectation-Maximization (EM) algorithm (Baum and Petrie, 1966; Dempster et al., 1977), a two-step iterative method for training latent-variable models, has long been the workhorse of unsupervised learning. In this section, we derive the EM algorithm using the data processing inequality to prepare for the subsequent quantum formulation.

Let $q(X)$ be the density of the data distribution to be learned by the LVM $p_\theta(X, Z)$. Let $q(X, Z)$ be any joint distribution that marginalizes to $q(X) = \sum_z q(X, Z)$. From the data processing inequality (Shannon, 1948) applied to marginalization, we know

$$\mathrm{D_{KL}}\left(q(X, Z), p_\theta(X, Z)\right) \geq \mathrm{D_{KL}}\left(q(X), p_\theta(X)\right). \tag{DPI}$$

From the chain rule of KL Divergence (Cover and Thomas, 2006),

$$\mathrm{D_{KL}}\left(q(X, Z), p_\theta(X, Z)\right) = \mathrm{D_{KL}}\left(q(X), p_\theta(X)\right) + \sum_x q(X{=}x)\mathrm{D_{KL}}\left(q(Z|X{=}x), p_\theta(Z|X{=}x)\right)$$

we know that the data processing inequality is saturated when $p_\theta(Z|x) = q(Z|X{=}x)$ for all x, that is, the model and data posteriors are the same.

Expanding the definition of KL divergence in the data processing inequality results in

$$\sum_x q(X{=}x) \log p_\theta(X{=}x) \geq -\mathrm{D_{KL}}\left(q(X{=}x, Z{=}z), p_\theta(X{=}x, Z{=}z)\right) - \mathrm{H_S}(q(Z))$$

where $\mathrm{H_S}(q(Z)) = -\sum_x q(X{=}x) \log q(X{=}x)$ is the Shannon entropy of the data distribution. When the data distribution is an empirical distribution over a finite set of points,

$q(\mathrm{X}{=}\mathrm{x}) = {}^1\!/\!N \sum_{i=1}^{N} \delta(\mathrm{x}^{(i)})$ where $\delta(\mathrm{x}^{(i)})$ is the Dirac measure at point $\mathrm{x}^{(i)}$, we may simplify the chain rule as,

$$\frac{1}{N} \sum_{i=1}^{N} \log p_\theta(\mathrm{X}{=}\mathrm{x}^{(i)}) \geq \frac{1}{N} \sum_{i=1}^{N} \sum_{\mathrm{z}} q(\mathrm{Z}{=}\mathrm{z}|\mathrm{X}{=}\mathrm{x}^{(i)}) \log \frac{p_\theta(\mathrm{X}{=}\mathrm{x}^{(i)}, \mathrm{Z}{=}\mathrm{z})}{q(\mathrm{Z}{=}\mathrm{z}|\mathrm{X}{=}\mathrm{x}^{(i)})}.$$

We observe that this inequality is the averaged ELBO for all data points. Alternatively maximizing the bound with respect to q and θ results in the EM algorithm described in Section 1.

2.1.1 E Step as Information Projection

The Expectation step of the EM algorithm may be seen as a solution to an information projection problem (Amari, 1995; Hino et al., 2024).

Definition 1 (Information Projection) *The information projection of a probability distribution p onto a set of probability distributions $\mathcal{S}$ is*

$$r^* = \operatorname*{argmin}_{r \in \mathcal{S}} \mathrm{D}_{\mathrm{KL}}(r, p).$$

The E step of the EM algorithm is an information projection of the model $p_{\theta^{(\mathrm{old})}}(\mathrm{X}, \mathrm{Z})$ onto the set of joint distributions that marginalize to the empirical data distribution $q(\mathrm{X})$. Formally, the solution to the information projection problem

$$q^*(\mathrm{X}, \mathrm{Z}) = \operatorname*{argmin}_{\sum_{\mathrm{z}} r(\mathrm{X}, \mathrm{z}) = q(\mathrm{X})} \mathrm{D}_{\mathrm{KL}}(r(\mathrm{X}, \mathrm{Z}), p_\theta(\mathrm{X}, \mathrm{Z})),$$

is given by $q^*(\mathrm{X}, \mathrm{Z}) = q(\mathrm{X}) p_\theta(\mathrm{Z}|\mathrm{X})$ (Cover and Thomas, 2006). It is easy to see that $q*$ saturates the DPI and gives us the Equation (E step).

2.2 Boltzmann Machines

This section discusses variants of Boltzmann Machines that played a key role in the development of modern deep learning (Larochelle and Bengio, 2008; Salakhutdinov et al., 2007; Hinton and Salakhutdinov, 2006), and the contrastive divergence algorithm for learning them.

2.2.1 Boltzmann Machines and Variants

Boltzmann Machines (BM) are stochastic neural networks that define a probability distribution over binary vectors based on the Ising model in statistical physics (Ackley et al., 1985). A Boltzmann Machine (BM) consists of a visible layer $\mathrm{x} \in \{0,1\}^m$ and a hidden layer $\mathrm{z} \in \{0,1\}^n$. It is parameterized by

$$\theta = \{\mathbf{W}, \mathbf{W}^{(\mathrm{V})}, \mathbf{W}^{(\mathrm{L})}, \mathbf{a}, \mathbf{b}\},$$

where $\mathbf{W} \in \mathbb{R}^{m \times n}$ denotes the interaction weights between visible and hidden layers, $\mathbf{W}^{(\mathrm{V})} \in \mathbb{R}^{m \times m}$ and $\mathbf{W}^{(\mathrm{L})} \in \mathbb{R}^{n \times n}$ are symmetric, hollow within-layer weight matrices, $\mathbf{a} \in \mathbb{R}^m$ is the

visible bias vector, and $\mathbf{b} \in \mathbb{R}^n$ is the hidden bias vector. The joint distribution is defined through an energy function E_θ as

$$E_\theta(\mathrm{x}, \mathrm{z}) = -\mathrm{x}^\top \mathbf{W} \mathrm{z} - \frac{1}{2}\mathrm{x}^\top \mathbf{W}^{(\mathrm{V})} \mathrm{x} - \frac{1}{2}\mathrm{z}^\top \mathbf{W}^{(\mathrm{L})} \mathrm{z} - \mathbf{a}^\top \mathrm{x} - \mathbf{b}^\top \mathrm{z},$$

$$\Pr(\mathrm{x}, \mathrm{z} \mid \theta) = \frac{1}{\mathcal{Z}(\theta)} \exp(E_\theta(\mathrm{x}, \mathrm{z})) \quad \text{and} \quad \mathcal{Z}(\theta) = \sum_{\mathrm{x}',\mathrm{z}'} \exp\left(E_\theta(\mathrm{x}', \mathrm{z}')\right).$$

Restricted Boltzmann Machines (RBM) were introduced with no intra-layer connections due to the intractability of learning in fully connected BMs (Smolensky, 1986; Hinton, 2002; Carreira-Perpiñán and Hinton, 2005). The energy function of an RBM simplifies to

$$E_\theta(\mathrm{x}, \mathrm{z}) = -\mathrm{x}^\top \mathbf{W} \mathrm{z} - \mathbf{a}^\top \mathrm{x} - \mathbf{b}^\top \mathrm{z}, \tag{RBM}$$

enabling efficient Gibbs sampling using the Contrastive Divergence algorithm discussed in the next section.

Gaussian-Bernoulli RBMs (GRBM) extend RBMs to model continuous visible units $\mathrm{x} \in \mathbb{R}^m$ and discrete hidden units $\mathrm{z} \in \{0, 1\}^n$ (Welling et al., 2004; Liao et al., 2022). It is additionally parameterized by a variance term $\mathbf{s} \in \mathbb{R}^m_{>0}$. The energy and partition functions are given by

$$E_\theta(\mathrm{x}, \mathrm{z}) = -\frac{1}{2}\sum_{i=1}^{m} \frac{(\mathrm{x}_i - \mathbf{a}_i)^2}{s_i} - \sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{W}_{ij}\frac{\mathrm{x}_i}{s_i}\mathrm{z}_j - \sum_{j=1}^{n} \mathbf{b}_j\mathrm{z}_j,$$

$$\mathcal{Z}(\theta) = \int_{-\infty}^{\infty} \sum_{\mathrm{z}} \exp(E_\theta(\mathrm{x}, \mathrm{z}))\mathrm{dx}. \tag{GRBM}$$

Deep Boltzmann Machines (DBM) stack RBMs using undirected connections and allow for joint training of all layers (Salakhutdinov and Hinton, 2009). The energy function of a DBM with L layers, $\Pr(\mathrm{x}, \mathrm{z}_{[1]}, \ldots, \mathrm{z}_{[L]})$ is defined as

$$E_\theta(\mathrm{x}, \mathrm{z}_{[1]}, \ldots, \mathrm{z}_{[L]}) = -\mathbf{a}^\top \mathrm{x} - \sum_{i=1}^{L} \mathbf{b}_i^\top \mathrm{z}_{[i]} - \mathrm{x}^\top \mathbf{W}^{(1)}\mathrm{z}_{[1]} - \sum_{i=1}^{L-1} \mathrm{z}_{[i]}^\top \mathbf{W}^{(i+1)}\mathrm{z}_{[i+1]}. \tag{DBM}$$

An RBM can be seen as a DBM with a single hidden layer.

2.2.2 CONTRASTIVE DIVERGENCE

The Contrastive Divergence (CD) algorithm provides a computationally tractable approximation to maximum-likelihood learning in Boltzmann Machines by replacing the intractable model expectation with a short-run Markov chain (Hinton, 2002; Carreira-Perpiñán and Hinton, 2005). The resulting estimator is biased but substantially more efficient, making it suitable for large-scale models where exact sampling would be prohibitively expensive.

The gradient of the log-likelihood of an RBM for the interaction terms $\mathbf{W}$ is

$$\frac{\partial}{\partial \mathbf{W}}\mathcal{L}(\mathcal{D}, \theta) = \mathbb{E}_{q(\mathrm{X})p_\theta(\mathrm{Z}|\mathrm{X})}(\mathrm{xz}^\top) - \mathbb{E}_{p_\theta(\mathrm{X},\mathrm{Z})}(\mathrm{xz}^\top)$$

where $q(\mathrm{X})$ is the data distribution. The second term in this expression is computationally intractable. Since an RBM has no intra-layer interactions, all units within a layer are conditionally independent given the state of the opposite layer

$$p_\theta(\mathrm{x}|\mathrm{z}) = \prod_{i=1}^{m} p_\theta(\mathrm{x}_i|\mathrm{z}) \quad \text{and} \quad p_\theta(\mathrm{z}|\mathrm{x}) = \prod_{i=1}^{n} p_\theta(\mathrm{z}_i|\mathrm{x}).$$

For a data point x, the CD-k algorithm approximates the expectation values in the gradients by the sampling scheme

$$\mathrm{z}_j(t) \sim p_\theta(\mathrm{z}_j = 1 \mid \mathrm{x}(t)) = \sigma \left(\sum_i \mathbf{W}_{ij} \mathrm{x}(t)_i + \mathbf{b}_j \right),$$

$$\mathrm{x}(t+1)_i \sim p_\theta(\mathrm{x}_i = 1 \mid \mathrm{z}(t)) = \sigma \left(\sum_j \mathbf{W}_{ij} \mathrm{z}(t)_j + \mathbf{a}_i \right),$$

$$\mathbb{E}_{p_\theta(\mathrm{Z}|\mathrm{X}=\mathrm{x})}(\mathrm{x}\mathrm{z}^\top) \approx \mathrm{x}\mathrm{z}(0)^\top,$$

$$\mathbb{E}_{p_\theta(\mathrm{X},\mathrm{Z})}(\mathrm{x}\mathrm{z}^\top) \approx \mathrm{x}(k)\mathrm{z}(k)^\top,$$

with $\mathrm{x}(0) = \mathrm{x}$. The gradients for the bias terms are calculated similarly.

The CD approach can be extended to variants of RBMs. In DBMs, as all units within a layer are conditionally independent given the state of the adjacent layers which leads to a sampling scheme similar to RBMs (Salakhutdinov and Hinton, 2009, 2012). In a GRBM, the conditional probabilities for the layers are given by

$$\mathrm{z}_j(t) \sim p_\theta(\mathrm{Z}_j = 1 \mid \mathrm{x}(t)) = \sigma \left(-\mathbf{b}_j - \sum_i \mathbf{W}_{ij} \frac{\mathrm{x}(t)_i}{s_i} \right),$$

$$\mathrm{x}(t+1)_i \sim p_\theta(\mathrm{X}_i \mid \mathrm{z}(t)) = \text{Normal} \left(\mu_i + \sum_j \mathbf{W}_{ij} \mathrm{z}(t)_j, s_i \right).$$

These contrastive divergence techniques made it possible to iteratively train variants of the Boltzmann Machines at scale, paving the way for their use as early generative models.

3 Quantum Information

Density operators were introduced by von Neumann (1955) as the mathematical foundation of quantum mechanics. The formalism built from them is therefore described as *quantum*, in contrast to the *classical* theory rooted in probability. While this paper does not engage in quantum mechanics directly, the ideas and results carry over to contexts such as quantum computing. We use the adjective quantum to describe the non-probabilistic nature of density operator. This section introduces the foundations of quantum information that are useful to work with the models in this paper.

3.1 Notation

This paper works in finite-dimensional Hilbert spaces unless stated otherwise, for the sake of clarity. All results admit natural extensions to the infinite-dimensional setting. A short introduction to matrix analysis is provided in Appendix A.

The absolute value of a complex number α is denoted by $|\alpha|$. $\mathcal{H}$ denotes a Hilbert space over $\mathbb{C}$ with dimension $d_{\mathcal{H}}$. The conjugate transpose of a column vector $\mathbf{v}$ in $\mathcal{H}$ is given by $\mathbf{v}^{\dagger}$ and its rank-one projectors are written $\Lambda(\mathbf{v}) = \mathbf{v}\mathbf{v}^{\dagger}$. Operators on $\mathcal{H}$, linear maps that takes vectors to vectors, form the set $\mathcal{T}(\mathcal{H})$. The set of density operators on $\mathcal{H}$ is $\mathcal{P}(\mathcal{H})$. The identity operator is denoted I with the relevant dimension or Hilbert space in the subscript. The support of A is denoted $\mathrm{supp}(A)$ and the null space of A is denoted $\ker(A)$. For operators A and B, the Kronecker product is $A \otimes B$ and the direct sum is $A \oplus B$. The Hilbert-Schmidt inner product is $\langle X, Y \rangle = \mathrm{Tr}(X^{\dagger}Y)$. The commutator of X and Y is denoted $[X, Y] = XY - YX$.

3.2 Density Operators

A quantum state is described by a density operator (Nielsen and Chuang, 2010).

Definition 2 (Density Operator) *Density operators on a Hilbert space $\mathcal{H}$ is the set $\mathcal{P}(\mathcal{H})$ of Hermitian, positive semi-definite operators with unit trace.*

Normalized elements $\mathbf{u}$ of the Hilbert space $\mathcal{H}$ are *pure states* with density operator $\rho = \Lambda(\mathbf{u})$ while other states are *mixed*. A density operator can be interpreted as a probability distribution over pure states since the spectral theorem stipulates that a density operator $\rho \in \mathcal{P}(\mathcal{H})$ can be expressed as a mixture

$$\rho = \sum_{i=1}^{d_{\mathcal{H}}} \alpha_i \Lambda(\mathbf{u}_i) \quad \text{with} \quad \sum_{i=1}^{d_{\mathcal{H}}} \alpha_i = 1, \ \alpha_i \geq 0 \text{ and } \mathbf{u}_i^{\dagger}\mathbf{u}_i = 1 \text{ for all } i$$

where $(\alpha_i, \mathbf{u}_i)$ are the eigenvalues and eigenvectors of ρ. A density operator over a Hilbert space with dimension $d_{\mathcal{H}}$ generalizes the notion of a probability distribution over a sample space with cardinality $d_{\mathcal{H}}$. The standard sample space of such a probability space correspond to the standard basis $\{\mathbf{e}_i\}_{i=1}^{d_{\mathcal{H}}}$ of the Hilbert space $\mathcal{H}$. Hence, classical probability distributions are diagonal density operators with $\mathbf{u}_i = \mathbf{e}_i$. It is often useful to assume additional properties for a density operator.

Definition 3 (Faithful Density Operator) *A density operator ρ on a finite-dimensional Hilbert space $\mathcal{H}$ is* faithful *if the following equivalent conditions hold:*

- *ρ is positive definite,*
- *$\mathrm{supp}(\rho) = \mathcal{H}$,*
- *ρ is invertible,*
- *$\ker(\rho) = 0$.*

This manuscript uses the operator theoretic term *faithful* to reflect the conditions required by our theorems in the infinite-dimensional setting (Bratteli and Robinson, 1987, 1997).

Density operators on a composite Hilbert space $\mathcal{H}_A \otimes \mathcal{H}_B$ play the role of joint probability distributions. A well-studied subclass is that of classical-quantum states, which have been extensively analyzed in the quantum information literature (Holevo, 1998; Wilde, 2016).

Definition 4 (Classical-Quantum State) *A density operator ρ in $\mathcal{P}(\mathcal{H}_A \otimes \mathcal{H}_B)$ is called a* classical-quantum state *if there is a predetermined orthonormal basis $\{\mathbf{u}_i\}_{i=1}^{d_A}$ such that*

$$\rho = \sum_{i=1}^{d_A} \alpha_i \Lambda(\mathbf{u}_i) \otimes \rho_B(i)$$

where $\rho_B(i)$ are density operators in $\mathcal{P}(\mathcal{H}_B)$ and $\alpha_i \geq 0$ with $\sum_{i=1}^{d_A} \alpha_i = 1$.

These models are called classical-quantum as the eigenvectors of ρ over the Hilbert space $\mathcal{H}_A$ is fixed. The component of ρ corresponding to the subsystem $\mathcal{H}_A$ is fully determined by the classical probability distribution $\{\alpha_i\}_{i=1}^{d_A}$. We use classical-quantum states to formulate classical-quantum latent variable models, a family of DO-LVMs that can be trained on standard ML data sets on classical hardware, as detailed in Section 7.

3.2.1 OBSERVABLES AND MEASUREMENT

Hermitian operators on $\mathcal{H}$ are the quantum mechanical analogs of random variables. The expected value of an observable $\mathcal{O}$ in $\mathcal{T}(\mathcal{H})$ with respect to a density operator ρ is given by the Hilbert-Schmidt inner product $\langle \mathcal{O} \rangle_\rho = \mathrm{Tr}(\rho\mathcal{O})$. This quantity reduces to the classical expectation value if ρ and $\mathcal{O}$ are diagonal as the quantum correlations encoded in the off-diagonal elements does not make an appearance in the expression for the trace. A sample is drawn from $\mathcal{O}$ using a *measurement*.

Definition 5 (Projective Measurement) *A* projective measurement *is described by a Hermitian observable $\mathcal{O}$ in $\mathcal{T}(\mathcal{H})$. If the observable has spectral decomposition $\mathcal{O} = \sum_{i=1}^{d_\mathcal{H}} \lambda_i \Lambda_i$ where Λ_i is the projector onto the eigenspace of $\mathcal{O}$ with eigenvalue λ_i, the measurement results in outcome λ_i with probability $\mathrm{Pr}(\lambda_i) = \mathrm{Tr}(\rho\Lambda_i)$.*

It is possible to identify the projection operator that resulted in the outcome if the observable $\mathcal{O}$ has non-degenerate eigenvalues. Multiple measurements on ρ is physically equivalent to measuring independent copies of same system and result in independent and identically distributed samples. Drawing a sample from a classical random variable X is equivalent to the measurement of the observable $\mathcal{O} = \sum_{i=1}^{d_\mathcal{H}} \mathrm{x}_i \Lambda(\mathbf{e}_i)$ on a diagonal density operator ρ which will result in x_i with probability $\mathrm{Tr}(\rho\Lambda(\mathbf{e}_i))$.

3.3 CPTP maps and Conditional Operators

In this section we detail the properties of completely positive trace preserving maps that act on density operators (Wilde, 2016; Watrous, 2018). A linear map $\mathcal{N} : \mathcal{T}(\mathcal{H}_A) \to \mathcal{T}(\mathcal{H}_B)$ is *positive* if $\mathcal{N}(X)$ is positive semi-definite for all positive semi-definite operators X in $\mathcal{T}(\mathcal{H})$. Such a map is *completely positive* if $\mathrm{I}_R \otimes \mathcal{N}$ is a positive map for an arbitrary reference system R. A map is *trace preserving* if $\mathrm{Tr}(\mathcal{N}(X)) = \mathrm{Tr}(X)$.

Definition 6 (CPTP map) *A CPTP map $\mathcal{N} : \mathcal{T}(\mathcal{H}_A) \to \mathcal{T}(\mathcal{H}_B)$ is a linear, completely positive, trace preserving map.*

The *adjoint* of a CPTP map $\mathcal{N} : \mathcal{T}(\mathcal{H}_A) \to \mathcal{T}(\mathcal{H}_B)$ is the unique linear map $\mathcal{N}^\dagger : \mathcal{T}(\mathcal{H}_B) \to \mathcal{T}(\mathcal{H}_A)$ satisfying $\langle Y, \mathcal{N}(X) \rangle = \langle \mathcal{N}^\dagger(Y), X \rangle$ for all X in $\mathcal{T}(\mathcal{H}_A)$ and Y in $\mathcal{T}(\mathcal{H}_B)$. The adjoint of a CPTP map is completely positive and unital, that is, $\mathcal{N}^\dagger(\mathrm{I}_B) = \mathcal{N}^\dagger(\mathrm{I}_A)$.

Projective measurements described in Definition 5 is a CPTP map that takes a density operator to a probability distribution (Watrous, 2018). Such maps are called *quantum-to-classical* maps. CPTP maps also allow the notion of marginalization of density operators over joint spaces. A density operator defined on a system $\mathcal{H}_A \otimes \mathcal{H}_B$ can be *marginalized* to a subspace by the partial trace operation.

Definition 7 (Partial Trace) *Let $\{\mathbf{u}_i\}_{i=1}^{d_B}$ be an orthonormal basis of the Hilbert space $\mathcal{H}_B$. The partial trace of a joint density operator $\rho \in \mathcal{P}(\mathcal{H}_A \otimes \mathcal{H}_B)$ with respect to $\mathcal{H}_B$ is the map $\mathrm{Tr}_B : \mathcal{P}(\mathcal{H}_A \otimes \mathcal{H}_B) \to \mathcal{P}(\mathcal{H}_A)$ given by*

$$\rho_A = \mathrm{Tr}_B(\rho) = \sum_{i=1}^{d_B} (I_A \otimes \mathbf{u}_i^{\dagger}) \rho (I_A \otimes \mathbf{u}_i).$$

It is easy to see that the partial trace operation is a CPTP map. The adjoint of the map $\mathcal{N}(\rho) = \mathrm{Tr}_B(\rho)$ is given by $\mathcal{N}^{\dagger}(\mathrm{Tr}_B(\rho)) = \mathrm{Tr}_B(\rho) \otimes \mathrm{I}_B$. Several concepts from probability theory, such as expectation values and marginalization, have direct analogs in the framework of density operators. However, satisfactory counterparts for conditional probability and Bayes' theorem are still lacking. We outline two existing proposals for such analogs.

3.3.1 Conditional Probability

There is no formal notion of conditional probability in density operators (Wilde, 2016). The conditional amplitude operator was introduced by Cerf and Adami (1997, 1999) as an operator analog to conditional probability. This operator mirrors the role of conditional probability in the definition of quantum conditional entropy.

Definition 8 (Conditional Amplitude Operator) *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*

$$\rho_{B|A} = \exp(\log \rho - \log \rho_A \otimes \mathrm{I}_B) \tag{CAO}$$

The `CAO` reduces to conditional probability if ρ is diagonal. In general, the `CAO` is not a perfect notion of a conditional density operator as it generally produces trace deficient operators. An EM-like algorithm to learn ρ_A from $\rho_{B|A}$ was conjectured in Warmuth and Kuzmin (2010). However, this algorithm did not attempt to learn from data and was not supported by correctness guarantees.

3.3.2 Bayes' Theorem

The Petz recovery map, a density operator analog of Bayes' theorem was proposed in (Petz, 1986, 1988) in the context of recoverability in information theory.

Definition 9 (Petz Recovery Map) *The Petz recovery map $\mathcal{R}_{\mathcal{N},\rho} : \mathcal{T}(\mathcal{H}_B) \to \mathcal{T}(\mathcal{H}_A)$ with respect to a CPTP map $\mathcal{N} : \mathcal{T}(\mathcal{H}_A) \to \mathcal{T}(\mathcal{H}_B)$ and a positive semi-definite operator ρ in $\mathcal{T}(\mathcal{H}_A)$ is*

$$\mathcal{R}_{\mathcal{N},\rho}(\omega) = \rho^{1/2} \mathcal{N}^{\dagger} \left(\mathcal{N}(\rho)^{-1/2} \omega \, \mathcal{N}(\rho)^{-1/2} \right) \rho^{1/2}. \tag{PRM}$$

The `PRM` is a completely positive, trace non-increasing map that does not always map density operators to density operators (Wilde, 2016). It can be seen from the proof in Wilde (2016) that

$$\mathrm{Tr}(\mathcal{R}_{\mathcal{N},\rho}(\omega)) = \mathrm{Tr}(\Lambda(\mathcal{N}(\rho))\omega) \leq \mathrm{Tr}(\omega)$$

where $\Lambda(\mathcal{N}(\rho))$ is the projector onto $\mathrm{supp}(\mathcal{N}(\rho))$. This allows for a condition on when the Petz recovery map is a CPTP map.

Corollary 10 *The Petz recovery map $\mathcal{R}_{\mathcal{N},\rho} : \mathcal{T}(\mathcal{H}_B) \to \mathcal{T}(\mathcal{H}_A)$ with respect to a CPTP map $\mathcal{N} : \mathcal{T}(\mathcal{H}_A) \to \mathcal{T}(\mathcal{H}_B)$ and a positive semi-definite operator ρ in $\mathcal{T}(\mathcal{H}_A)$ is a CPTP map if $\mathcal{N}(\rho)$ is a faithful density operator.*

The Petz recovery map reduces to the Bayes' theorem when ρ and $\mathcal{N}$ are probabilistic (Wilde, 2016). The Petz recovery map with respect to the partial trace is not a perfect notion of Bayes rule as

$$\omega \neq \mathrm{Tr}_B\left(\mathcal{R}_{\mathrm{Tr}_B,\rho}(\omega)\right).$$

for ω in $\mathcal{P}(\mathcal{H}_A)$ and ρ in $\mathcal{P}(\mathcal{H}_A \otimes \mathcal{H}_B)$.

3.4 Quantum Relative Entropy and Inequalities

In this section, we collect results about quantum relative entropy that will be useful in the specification of the EM algorithm for density operators.

A real valued function $f : \mathbb{R} \to \mathbb{R}$ can be naturally extended to act on density operators via the spectral theorem,

$$f(\rho) = \sum_{i=1}^{d_{\mathcal{H}}} f(\alpha_i)\Lambda(\mathbf{u}_i)$$

as discussed in Bhatia (1997). This construction was used by von Neumann (1955) to generalize the classical Shannon entropy to the quantum setting, defining the entropy of a density operator as

$$\mathrm{S}_{\mathrm{VN}}(\rho) = -\mathrm{Tr}(\rho \log \rho).$$

This quantity, now called the von Neumann entropy, is only a function of the eigenvalues of the density operator. Umegaki (1962) extended the classical Kullback-Leibler divergence (Cover and Thomas, 2006) between probability distributions to the quantum domain.

Definition 11 (Umegaki Relative Entropy) *Let ω and ρ be density operators in $\mathcal{P}(\mathcal{H})$ with $\ker(\rho) \subseteq \ker(\omega)$. Their relative entropy is given by*

$$\mathrm{D}_{\mathrm{U}}(\omega,\rho) = \mathrm{Tr}(\omega \log \omega) - \mathrm{Tr}(\omega \log \rho).$$

The Umegaki relative entropy inherits several properties of KL divergence such as non-negativity and joint convexity with $\mathrm{D}_{\mathrm{U}}(\omega,\rho) = 0$ if and only if $\omega = \rho$. Lindblad (1975) showed that the relative entropy is monotonic under the action of a CPTP map.

Theorem 12 (Monotonicity of Relative Entropy) *For all density operators ω and ρ in $\mathcal{P}(\mathcal{H})$ such that $\ker(\omega) \subset \ker(\rho)$, $\mathrm{D_U}(\omega, \rho) \geq \mathrm{D_U}(\mathcal{N}(\omega), \mathcal{N}(\rho))$.*

The Monotonicity of Relative Entropy (`MRE`) is the quantum information theoretical analog of the data processing inequality. Petz (1986, 1988) showed that `MRE` is saturated if and only if the action of the CPTP map is *reversible*. We provide a restatement of this result from Hayden et al. (2004).

Theorem 13 (Theorem 3, Hayden et al., 2004) *For states ω and ρ in $\mathcal{P}(\mathcal{H}_A)$ and a CPTP map $\mathcal{N} : \mathcal{P}(\mathcal{H}_A) \to \mathcal{P}(\mathcal{H}_B)$, $\mathrm{D_U}(\omega, \rho) = \mathrm{D_U}(\mathcal{N}(\omega), \mathcal{N}(\rho))$ if and only if there exists a CPTP map $\mathcal{R}$ such that $\mathcal{R}(\mathcal{N}(\omega)) = \omega$ and $\mathcal{R}(\mathcal{N}(\rho)) = \rho$. Furthermore, on the support of $\mathcal{N}(\rho)$, $\mathcal{R}$ is explicity given by the Petz recovery map.*

Ruskai (2002) proved an equivalent condition using the matrix logarithm for the saturation of the monotonicity of relative entropy.

Theorem 14 (Ruskai's condition) *For states ω and ρ in $\mathcal{P}(\mathcal{H}_A)$ and a quantum chanenl $\mathcal{N} : \mathcal{P}(\mathcal{H}_A) \to \mathcal{P}(\mathcal{H}_B)$, $\mathrm{D_U}(\omega, \rho) = \mathrm{D_U}(\mathcal{N}(\omega), \mathcal{N}(\rho))$ if and only if*

$$\log \omega - \log \rho = \mathcal{N}^\dagger (\log \mathcal{N}(\omega) - \log \mathcal{N}(\rho)).$$

Ruskai's condition, which can obtained by differentiating the equality condition in Petz's characterization of monotonicity (Petz, 2003), provides a criterion for perfect recoverability under the Petz recovery map.

4 A Survey on QBMs

In this section, we introduce the Quantum Boltzmann Machine, the predominant density operator latent variable model, and outline the main challenges in training it.

4.1 Mathematical Qubits

A mathematical bit used to define Boltzmann machines and other probabilistic models can take the value $+1$ with probability p and the value -1 with probability $1 - p$. Quantum Boltzmann Machines are defined using mathematical qubits (Nielsen and Chuang, 2010), the density operator analog of a mathematical bit.

A mathematical qubit is a density operator over a Hilbert space $\mathcal{H}$ of dimension $d_{\mathcal{H}} = 2$ over $\mathbb{C}$ with the $+1$ state corresponding to the vector $(1, 0)^\dagger$ and the -1 state corresponding to $(0, 1)^\dagger$. The vectors $(1, 0)^\dagger$ and $(0, 1)^\dagger$ are known as *computational basis states*. Notably, the Hilbert space also contains an infinite number of states $u = \alpha(1, 0)^\dagger + \beta(0, 1)^\dagger$ with α and β being complex numbers such that $|\alpha|^2 + |\beta|^2 = 1$. When both α and β are non-zero, the state u is said to be in *superposition*, a state in the continuum between the computational basis states. It is well known that such qubits can be represented as unit vectors on a sphere, known as the Bloch sphere, by the parameterization

$$u = \cos \frac{\theta}{2} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + e^{i\varphi} \sin \frac{\theta}{2} \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

While a mathematical bit can only be sampled for $+1$ or -1, a mathematical qubit can be sampled for $+1$ or -1 along any direction in the Bloch sphere with each direction giving different statistics. The observables $\mathcal{O} \in \mathcal{T}(\mathcal{H})$ that correspond to sampling along the three coordinate axes are given by the Pauli operators

$$\sigma^{(x)} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \ \sigma^{(y)} = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \ \text{and } \sigma^{(z)} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

which together with the identity matrix, spans the set of all Hermitian observables on a qubit. The Pauli operators are traceless with eigenvalues $+1$ and -1. When sampled along a coordinate axis k, a density operator ρ will return $+1$ or -1 with average $\mathrm{Tr}(\rho\sigma^{(k)})$ where $k \in \{x, y, z\}$. Sampling from a classical bit is equivalent to sampling along the z-axis from a diagonal density operator.

4.2 Quantum Boltzmann Machines

The Hamiltonian is a Hermitian operator $\mathrm{H} \in \mathcal{T}(\mathcal{H})$ representing the total energy and generalizes the notion of an energy function in classical energy-based models (Mahan, 2013). A Hamiltonian-based model can be defined using the Gibbs state density matrix of a Hamiltonian given by

$$\rho(\theta) = \frac{\exp(\mathrm{H}(\theta))}{Z(\theta)}, \quad \text{with } \mathrm{H}(\theta) = \sum_r \theta_r \mathrm{H}_r \text{ and } Z(\theta) = \mathrm{Tr}\exp(\mathrm{H}(\theta))$$

where $\mathrm{H}_r \in \mathcal{T}(\mathcal{H})$ are fixed Hermitian operators and $\theta_r \in \mathbb{R}$ are model parameters. If $\mathcal{H} = \mathcal{H}_\mathrm{V} \otimes \mathcal{H}_\mathrm{L}$ with only $\mathcal{H}_\mathrm{V}$ observed, the model is a latent variable model with density operator over the visible units given by $\rho_\mathrm{V}(\theta) = \mathrm{Tr}_\mathrm{L}\rho(\theta)$. The Quantum Boltzmann Machine (QBM) introduced by Amin et al. (2018) is a Hamiltonian-based model inspired by the transverse field Ising model (Mahan, 2013). The Hamiltonian of QBM with m visible units and n hidden units is given by

$$\mathrm{H}(\theta) = \begin{cases} -\sum_{i=1}^{m} \mathbf{a}_i \sigma_i^{(z)} - \sum_{i=1}^{n} \mathbf{b}_i \sigma_{m+i}^{(z)} - \sum_{i=1}^{m+n} \Gamma_i \sigma_i^{(x)} \\ -\sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{W}_{ij} \sigma_i^{(z)} \sigma_j^{(z)} - \sum_{i=1}^{m}\sum_{j=1}^{m} \mathbf{W}_{ij}^{(\mathrm{V})} \sigma_i^{(z)} \sigma_j^{(z)} - \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{W}_{ij}^{(\mathrm{L})} \sigma_{m+i}^{(z)} \sigma_{m+j}^{(z)} \end{cases} \quad \text{(QBM)}$$

where $\sigma_i^{(z)}$ and $\sigma_i^{(x)}$ are $2^{m+n} \times 2^{m+n}$ spin operators that act on $\mathcal{H}_\mathrm{V} \otimes \mathcal{H}_\mathrm{L}$ defined by

$$\sigma_i^{(k)} = \mathrm{I}_{2^{i-1}} \otimes \sigma^{(k)} \otimes \mathrm{I}_{2^{m+n-i}} \quad \text{for } k \text{ in } \{x, z\}. \tag{2}$$

The Hamiltonian of the Heisenberg model is also used to define a QBM (Kappen, 2020). Setting $\Gamma_i = 0$ for all i in this Hamiltonian gives a diagonal Hamiltonian with each entry corresponding to the energy of a state in a classical Boltzmann Machine. These density operator models are plagued by their $2^{m+n} \times 2^{m+n}$ dimensionality.

4.3 Learning QBMs

In this section, we introduce the two learning objectives for Quantum Boltzmann Machines (QBM) proposed in the literature and survey work done in learning them.

The first objective, termed the *projective log-likelihood* (PL), is based on projective measurements in the visible Hilbert space and was initially introduced for QBM training by Amin et al. (2018). Since then, it has been adopted in numerous subsequent works (Anschuetz and Cao, 2019; Zoufal et al., 2021; Demidik et al., 2025) and is closely related to the likelihood objective employed in quantum state tomography (Hradil, 1997; Rehácek et al., 2007). In this framework, the model maximizes the classical log-likelihood of a data set $\mathcal{D} = \{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(N)}\}$ given by

$$\mathcal{L}_{\mathrm{P}}(\mathcal{D}, \theta) = \frac{1}{N} \sum_{i=1}^{N} \log \mathrm{Tr}\left(\left(\Lambda(\mathbf{v}^{(i)}) \otimes \mathrm{I}_{\mathrm{L}}\right)\rho(\theta)\right) = \frac{1}{N} \sum_{i=1}^{N} \log \mathrm{Tr}\left(\Lambda(\mathbf{v}^{(i)})\mathrm{Tr}_{\mathrm{L}}(\rho(\theta))\right), \quad \text{(PL)}$$

where $\mathrm{Tr}\left(\Lambda(\mathbf{v}^{(i)})\mathrm{Tr}_{\mathrm{L}}(\rho(\theta))\right)$ is the probability of obtaining sample $\mathbf{v}^{(i)}$ under a projective measurement. The presence of projection operators in the learning objective complicates gradient estimation. To circumvent this, Amin et al. (2018) introduced a bound-based optimization scheme derived from the Golden-Thompson inequality. However, this approach cannot update the transverse field parameters that control the model's quantum effects.

The second objective, termed the *quantum log-likelihood* (QL), is based on the Umegaki relative entropy and was introduced in Kieferová and Wiebe (2017). A data set with inbuilt quantum correlations or a data set with artificially induced quantum correlations (Kappen, 2020) is used to build an empirical data density operator η_{V} in $\mathcal{P}(\mathcal{H}_{\mathrm{V}})$. The maximization of the quantum log-likelihood

$$\mathcal{L}_{\mathrm{U}}(\eta_{\mathrm{V}}, \theta) = \mathrm{Tr}\left(\eta_{\mathrm{V}} \log \mathrm{Tr}_{\mathrm{L}}\rho(\theta)\right) = \mathrm{Tr}\left(\eta_{\mathrm{V}} \log \rho_{\mathrm{V}}(\theta)\right) \quad \text{(QL)}$$

is equivalent to the minimization of the Umegaki relative entropy between η_{V} and $\rho_{\mathrm{V}}(\theta)$. The appearance of the logarithm of a partial trace in this expression prevents the derivation of a closed-form gradient (Wiebe and Wossnig, 2019). This challenge closely parallels the difficulty encountered in classical latent variable models, where the objective involves a logarithm of a summation. Although both approaches reduce to maximizing the classical log-likelihood when the density operators are diagonal, no formal relationship between the two objective functions has been established to date.

Recent work on QBMs has primarily focused on formulations without hidden units (Kappen, 2020; Patel and Wilde, 2024; Patel et al., 2024; Huijgen et al., 2024), owing to the operator-theoretic complications introduced by the partial trace in QL during gradient computation. Studies that incorporate hidden variables have relied on heuristic gradient evaluation techniques (Amin et al., 2018; Wiebe and Wossnig, 2019; Anschuetz and Cao, 2019; Zoufal et al., 2021), typically involving approximate preparation of Gibbs state density operators (Bilgin and Boixo, 2010; Chen et al., 2023; Ding et al., 2025) without any theoretical guarantees. A related yet orthogonal direction concerns the realization of QBMs on quantum devices, which is a subsequent stage following the algorithmic development of the underlying algorithms (Benedetti et al., 2017; Nelson et al., 2022; Shibukawa et al., 2024). These methods typically fail to scale beyond toy data sets of 12-bit binary strings, largely because

of limited access to quantum hardware, the exponential cost of simulating quantum systems, and the memory demands of representing and optimizing QBMs on classical machines. Even when memory is not the bottleneck, computing the gradient remains a major challenge for models with hidden units. Therefore, the empirical evaluation of these models has been limited to toy data sets, and their performance on realistic data remains an open question.

5 Density Operator Latent Variable Models

In this section, we undertake a formal study of LVMs arising out of density operators. We prove a fundamental inequality between the differing notions of log-likelihood used to train QBMs and showcase a novel quantum evidence lower bound.

We assume a data set is obtained from N arbitrary projective measurements of an underlying density operator ψ_V defined on the Hilbert space $\mathcal{H}_V$. As discussed in Section 3, each projective measurement yields a pure state corresponding to a projection operator, so each sample in the data set can be represented by a unit vector in $\mathcal{H}_V$. The *empirical data density operator* $\eta_V \in \mathcal{P}(\mathcal{H}_V)$ for a data set $\mathcal{D} = \{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(N)}\}$ is defined as

$$\eta_V = \frac{1}{N} \sum_{i=1}^{N} \Lambda(\mathbf{v}^{(i)}). \tag{3}$$

We now formally define a density operator-based LVM.

Definition 15 *A Density Operator Latent Variable Model (DO-LVM) specifies the density operator $\rho_V \in \mathcal{P}(\mathcal{H}_V)$ on observables in $\mathcal{H}_V$ through a joint density operator $\rho \in \mathcal{P}(\mathcal{H}_V \otimes \mathcal{H}_L)$ as $\rho_V = \mathrm{Tr}_L(\rho(\theta))$ where the space $\mathcal{H}_L$ is not observed.*

The objective of learning a DO-LVM is to obtain parameters θ such that $\rho_V(\theta)$ is as close as possible to a ground truth density operator ψ_V from which samples are drawn. This is can be achieved by training the projective log-likelihood (PL) or the quantum log-likelihood (QL) depending on the data provided. As noted in Section 4.3, the use of either log-likelihood with hidden units comes with its own distinct set of challenges. The use of QL, however, has quantum information theoretic interpretations related to the Umegaki relative entropy and potentially paves the way for employing operator theoretic inequalities to aid the learning process. We now show that for a data set $\mathcal{D}$ arising out of projective measurements, the quantum log-likelihood provides a lower bound to the projective log-likelihood.

Lemma 16 *For a data set $\mathcal{D} = \{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(N)}\}$ arising out of projective measurements and a DO-LVM $\rho(\theta)$, let the empirical data density operator η_V be as defined in Equation (3). Then, $\mathcal{L}_P(\mathcal{D}, \theta) \geq \mathcal{L}_U(\eta_V, \theta)$.*

The proof uses stated Theorem 31 stated in Appendix A.
Proof The logarithm is a continuous concave function in the positive real and each $\Lambda(\mathbf{v}^{(i)})$ is a positive semi-definite, unit trace operator in $\mathcal{T}(\mathcal{H}_V)$. Then by Theorem 31, we have the inequality

$$\log \mathrm{Tr}\left(\Lambda(\mathbf{v}^{(i)}) \rho_V(\theta)\right) \geq \mathrm{Tr}\left(\Lambda(\mathbf{v}^{(i)}) \log \rho_V(\theta)\right)$$

for each data point $\mathbf{v}^{(i)} \in \mathcal{D}$. Averaging the inequality across all data points results in

$$\frac{1}{N} \sum_{i=1}^{N} \log \mathrm{Tr} \left(\Lambda(\mathbf{v}^{(i)}) \rho_\mathrm{V}(\theta) \right) \geq \frac{1}{N} \sum_{i=1}^{N} \mathrm{Tr} \left(\Lambda(\mathbf{v}^{(i)}) \log \rho_\mathrm{V}(\theta) \right),$$

which by the linearity of trace can be written as

$$\frac{1}{N} \sum_{i=1}^{N} \log \mathrm{Tr} \left(\Lambda(\mathbf{v}^{(i)}) \rho_\mathrm{V}(\theta) \right) \geq \mathrm{Tr} \left(\left(\frac{1}{N} \sum_{i=1}^{N} \Lambda(\mathbf{v}^{(i)}) \right) \log \rho_\mathrm{V}(\theta) \right),$$

$$\mathcal{L}_\mathrm{P}(\mathcal{D}, \theta) \geq \mathcal{L}_\mathrm{U}(\eta_\mathrm{V}, \theta)$$

concluding our proof. ∎

The value of the quantum log-likelihood provides a lower bound on the the projected log-likelihood. However, training on `QL` is also intractable due to the partial trace.

The intractability of the gradient of the log-likelihood in probabilistic LVMs is addressed by the EM algorithm. As probabilistic LVMs are a special case of density operator latent variable models (DO-LVM), the training challenges they face persist in DO-LVMs. DO-LVMs also introduce additional operator-theoretic complications due to the absence of conditional density operators and the limited applicability of existing operator versions of Jensen's inequality. Consequently, the standard derivation of the EM algorithm does not extend naturally to the setting of density operator models. In this light, we investigate possible EM-like algorithms to train `QL` using quantum information. The first step in the derivation of an EM algorithm for DO-LVMs is the derivation of an evidence lower bound for data operators η_V that can be used in an EM framework. Since the classical ELBO inequality can be derived using the data processing inequality for marginalization, we derive a quantum ELBO using Theorem 12.

Lemma 17 (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 data $\eta_\mathrm{V} \in \mathcal{P}(\mathcal{H}_\mathrm{V})$. Then for a DO-LVM $\rho(\theta)$ and $\eta \in \mathcal{J}(\eta_\mathrm{V})$,*

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

Proof Since $\mathrm{Tr}_\mathrm{L}\eta = \eta_\mathrm{V}$, by Theorem 12 we obtain

$$\mathrm{D}_\mathrm{U}(\eta, \rho(\theta)) \geq \mathrm{D}_\mathrm{U}(\eta_\mathrm{V}, \rho_\mathrm{V}(\theta)).$$

Expanding the expression for Umegaki relative entropy and rearranging the terms results in

$$\mathrm{Tr}(\eta \log \eta) - \mathrm{Tr}(\eta \log \rho(\theta)) \geq \mathrm{Tr}(\eta_\mathrm{V} \log \eta_\mathrm{V}) - \mathrm{Tr}(\eta_\mathrm{V} \log \rho_\mathrm{V}(\theta)), \text{ and}$$
$$\mathrm{Tr}(\eta_\mathrm{V} \log \rho_\mathrm{V}(\theta)) \geq \mathrm{Tr}(\eta \log \rho(\theta)) - \mathrm{Tr}(\eta \log \eta) + \mathrm{Tr}(\eta_\mathrm{V} \log \eta_\mathrm{V}),$$
$$\mathcal{L}_\mathrm{U}(\eta_\mathrm{V}, \theta) \geq \mathrm{Tr}(\eta \log \rho(\theta)) + \mathrm{S}_\mathrm{VN}(\eta) - \mathrm{S}_\mathrm{VN}(\eta_\mathrm{V}).$$

which is the required inequality. ∎

The `QELBO` expression does not involve projection operators or the partial trace operation that make the optimization of the projected log-likelihood or the quantum log-likelihood

hard. The `QELBO` serves as the density-operator analog of the ELBO, since the monotonicity of relative entropy reduces to the data processing inequality when the density operators are diagonal. The classical EM algorithm is a consequence of the evidence lower bound being a minorant of the log-likelihood. As noted in Section 2, this can be seen as a saturation of the data processing inequality for marginalization. However, it is well known that the monotonicity of relative entropy is often not saturated for the partial trace operation (Lesniewski and Ruskai, 1999; Berta et al., 2015; Wilde, 2015; Carlen and Vershynina, 2020; Cree and Sorce, 2022). Hence, a similar derivation is not directly possible. In the following section, we investigate an expectation-maximization framework for the quantum ELBO.

6 The DO-EM Framework

In this section, we derive an expectation-maximization framework for density operators using `QELBO`. Inspired by the information geometric interpretation of the EM algorithm detailed in Section 2, we study a quantum information projection problem to saturate `QELBO`.

6.1 A Quantum Information Projection Problem

In this section we study the I-projection (Cover and Thomas, 2006) problem for density operators and show conditions when `PRM` can solve this problem.

Definition 18 (Quantum Information Projection) *The Quantum Information Projection of a density operator ρ in $\mathcal{P}(\mathcal{H}_A \otimes \mathcal{H}_B)$ onto a density operator ω in $\mathcal{P}(\mathcal{H}_A)$ with respect to the partial trace $\mathrm{Tr}_B : \mathcal{P}(\mathcal{H}_A \otimes \mathcal{H}_B) \to \mathcal{P}(\mathcal{H}_A)$ is the density operator ξ^* in $\mathcal{P}(\mathcal{H}_A \otimes \mathcal{H}_B)$ such that*

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

To the best of our knowledge, this problem has not been studied in literature. We know from Theorem 12 that the theoretical minimum attained by the objective function in `QIP` is $\mathrm{D_U}(\omega, \mathrm{Tr}_B(\rho))$ though it may not always be attained. According to Theorem 13, the solution to `QIP` should be the Petz recovery map when this lower bound is achieved. Inspired by this connection, we explore sufficiency conditions for when the Petz recovery map is the solution to the `QIP` problem.

As noted in Section 3, the Petz recovery map is not a trace preserving map. The Petz recovery map cannot be a solution to `QIP` if it is not unit trace as it will not partial trace to ω. Hence, the faithfulness of $\mathrm{Tr_L}(\rho)$ is a necessary condition for the Petz recovery map to sovle `QIP`. We now outline two commutative relations sufficient to ensure that the Petz recovery map solves the `QIP` problem.

Definition 19 (Sufficient Conditions) *Two density operators ρ in $\mathcal{P}(\mathcal{H}_A \otimes \mathcal{H}_B)$ and ω in $\mathcal{P}(\mathcal{H}_A)$ satisfy the* sufficient conditions *if:*

1. *$\mathrm{Tr_L}(\rho)$ is faithful*

2. *$[\rho, \mathrm{Tr}_B(\rho) \otimes \mathrm{I}_B] = 0$, and*

3. *$[\omega, \mathrm{Tr}_B(\rho)] = 0$.*

The faithfulness condition guarantees that the Petz recovery map provides a valid density operator. The second condition ensures that ρ is block diagonal when $\text{Tr}_B(\rho)$ is rotated to be diagonal. The final condition ensures that the ω and $\text{Tr}_B(\rho)$ share the same pure states in their spectral decomposition and only differ in probability of each state, making information projection easier by mapping blocks over $\mathcal{H}_B$ to the appropriate pure state in $\mathcal{H}_A$. We defer formal statements on these interpretations to Section 7. When density operators satisfy these sufficiency conditions, we are able to solve the `QIP` problem with the Petz recovery map. We formally state this result in Theorem 20 below.

Theorem 20 *Recall the `QIP` problem in Definition 18. Suppose ρ and ω are two density operators in $\mathcal{P}(\mathcal{H}_A \otimes \mathcal{H}_B)$ and $\mathcal{P}(\mathcal{H}_A)$ respectively such that the Sufficient Conditions are satisfied, the solution to the `QIP` problem is the Petz recovery map*

$$\xi^* = \mathcal{R}_{\text{Tr}_B, \rho}(\omega).$$

The proof of this theorem uses Lemma 32 stated in Appendix A.1.

Proof Since $\text{Tr}_L(\rho)$ is faithful, the Petz recovery map gives a valid density operator by Lemma 10. Note that if two matrices commute, their squareroots also commute because they share the same eigenspace. The Petz recovery map with respect to ρ and $\mathcal{N}$ is equal to

$$\mathcal{R}_{\mathcal{N}, \rho}(\omega) = \log \rho^{1/2} \mathcal{N}^\dagger (\mathcal{N}(\rho)^{-1/2} \omega \mathcal{N}(\rho)^{-1/2}) \rho^{1/2}.$$

Since $[\omega, \text{Tr}_B(\rho)] = 0$, $[\text{Tr}_B(\rho)^{-1}\omega, \text{Tr}_B(\rho)] = 0$ and $[(\text{Tr}_B(\rho)^{-1}\omega) \otimes I_B, \rho] = 0$. Therefore, the Petz recovery map is equal to

$$\begin{aligned}
\mathcal{R}_{\text{Tr}_L, \rho}(\omega) &= \rho(\text{Tr}_B(\rho)^{-1}\omega \otimes I_B), \\
\log \mathcal{R}_{\text{Tr}_B, \rho}(\omega) &= \log \rho + \log(\text{Tr}_B(\rho)^{-1}\omega \otimes I_B), \text{ and} &\text{(commutativity)} \\
\log \mathcal{R}_{\text{Tr}_B, \rho}(\omega) &= \log \rho + \log(\text{Tr}_B(\rho)^{-1}\omega) \otimes I_B. &\text{(Lemma 32)}
\end{aligned}$$

We now show that the Petz recovery map and ρ satisfy the Ruskai condition.

$$\begin{aligned}
\log \rho - \log \mathcal{R}_{\text{Tr}_B, \rho}(\omega) &= \log \rho - \log \rho - \log(\text{Tr}_B(\rho)^{-1}\omega) \otimes I_B \\
&= -\log(\text{Tr}_B(\rho)^{-1}) \otimes I_B - \log(\omega) \otimes I_B &\text{(commutativity)} \\
&= \log(\text{Tr}_B(\rho)) \otimes I_B - \log(\omega) \otimes I_B. &(\log(A^{-1}) = -\log A)
\end{aligned}$$

Hence, `PRM` is a density operator that satisfies Ruskai's condition in Theorem 14. Therefore, $D_U(\mathcal{R}_{\text{Tr}_B, \rho}(\omega), \rho) = D_U(\omega, \text{Tr}_B\rho)$. Theorem 13 implies that $\text{Tr}_B \mathcal{R}_{\text{Tr}_B, \rho}(\omega) = \omega$. We have shown that the Petz recovery map is both a feasible density operator that partial traces to ω and solves the `QIP` problem. ∎

We note that this technique is not generally applicable to arbitrary CPTP maps as Lemma 32 does not apply to arbitrary completely positive unital maps. We now develop our EM framework for density operators.

6.2 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 when the Sufficient Conditions are satisfied.

For fixed parameters $\theta^{(\text{old})}$, 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.

$$\eta(\theta^{(\text{old})}) = \underset{\text{Tr}_{\text{L}}\eta=\eta_{\text{V}}}{\operatorname{argmin}} \, D_{\text{U}}(\eta, \rho(\theta^{(\text{old})}))$$

$$\mathcal{Q}(\theta; \theta^{(\text{old})}) = \text{QELBO}(\eta(\theta^{(\text{old})}), \rho(\theta))$$

We use $\mathcal{Q}$ to define the DO-EM algorithm in Algorithm 1.

Algorithm 1 Density Operator Expectation Maximization

1: **Input:** Data density operator η_{V} and model parameters $\theta^{(0)}$
2: **while** not converged **do**
3:     **E Step:** $\eta^{(t)} = \underset{\eta:\text{Tr}_{\text{L}}\eta=\eta_{\text{V}}}{\operatorname{argmin}} \, D_{\text{U}}(\eta, \rho(\theta^{(t)}))$
4:     **M Step:** $\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}} \, \text{Tr}(\eta^{(t)} \log \rho(\theta))$

As the saturation of monotonicity of relative entropy with respect to the partial trace is not guaranteed, it is not possible to guarantee log-likelihood ascent without additional assumptions. Models and QIPs that obey Ruskai's condition provably achieve log-likelihood ascent under the DO-EM procedure.

Theorem 21 ($\mathcal{Q}$ is a minorant) *Let η_{V} be a data density operator and $\rho(\theta)$ be a DO-LVM trained by the DO-EM algorithm. If $\rho(\theta^{(t)})$ and its* `QIP`*, $\eta^{(t)}$, obey the Ruskai condition, then $\mathcal{Q}$ is a minorant of the log-likelihood. That is,*

$$\mathcal{L}(\theta^{(t)}) \le \mathcal{L}(\theta^{(t+1)}) \text{ where } \theta^{(t+1)} = \operatorname{argmax}_{\theta} \mathcal{Q}(\theta; \theta^{(t)}).$$

Proof From Ruskai's condition and Theorem 17, we obtain

$$\mathcal{L}(\theta^{(t)}) = \mathcal{Q}(\theta^{(t)}, \theta^{(t)}). \tag{4}$$

By the definition of $\theta^{(t+1)}$,

$$\mathcal{Q}(\theta^{(t)}, \theta^{(t)}) \le \mathcal{Q}(\theta^{(t)}, \theta^{(t+1)}). \tag{5}$$

Combining these inequalities with the QELBO, we obtain

$$\mathcal{Q}(\theta^{(t)}, \theta^{(t+1)}) \le \mathcal{L}(\theta^{(t+1)});$$
$$\mathcal{Q}(\theta^{(t)}, \theta^{(t)}) \le \mathcal{L}(\theta^{(t+1)}) \quad \text{by (5)};$$
$$\mathcal{L}(\theta^{(t)}) \le \mathcal{L}(\theta^{(t+1)}) \quad \text{by (4).}$$

∎

A practical approach to ensure that the model and `QIP` satisfy Ruskai's condition is to ensure that the model and the data operators satisfy the Sufficient Conditions.

Corollary 22 *For a target density operator η_{V} and model $\rho(\theta)$ satisfying Sufficient Conditions, the E step is the Petz recovery map $\mathcal{R}_{\rho,\mathrm{Tr}_{\mathrm{L}}}(\eta_{\mathrm{V}})$. Moreover, such a model trained using the DO-EM algorithm achieves provable likelihood ascent at every iteration.*

Proof We proved in Theorem 20 that the Petz recovery map satisfies the Ruskai condition under the Sufficient Conditions. Hence, the statement is an immediate corollary to Theorem 21. ∎

Thus, in the DO-EM algorithm, the Petz recovery map takes on the role that conditional probability plays in the classical EM algorithm.

6.3 Commentary on the DO-EM Algorithm

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.

When the Sufficient Conditions are satisfied, the **E Step** in DO-EM finds a feasible extension η whose Conditional Amplitude Operator (`CAO`) is equal to that of the model $\rho(\theta)$.

Corollary 23 *If two density operators η and ρ in $\mathcal{P}(\mathcal{H}_{\mathrm{V}} \otimes \mathcal{H}_{\mathrm{L}})$ satisfy the Ruskai condition for the partial trace $\mathrm{Tr}_B : \mathcal{P}(\mathcal{H}_{\mathrm{V}} \otimes \mathcal{H}_{\mathrm{L}}) \to \mathcal{P}(\mathcal{H}_{\mathrm{V}})$, then η and ρ have the same conditional amplitude operator $\eta_{\mathrm{L}|\mathrm{V}} = \rho_{\mathrm{L}|\mathrm{V}}$ with respect to the Hilbert space $\mathcal{H}_{\mathrm{V}}$.*

Proof By the Ruskai condition,

$$\log \eta - \log \rho = \mathcal{N}^{\dagger}(\log \mathcal{N}(\eta) - \log \mathcal{N}(\rho)).$$

Specializing for $\mathcal{N}(\rho) = \mathrm{Tr}_{\mathrm{L}}(\rho)$ and $\mathcal{N}^{\dagger}(\omega) = \omega \otimes \mathrm{I}_{\mathrm{L}}$,

$$\log \eta - \log \rho = (\log \mathrm{Tr}_{\mathrm{L}}(\eta) - \log \mathrm{Tr}_{\mathrm{L}}(\rho)) \otimes \mathrm{I}_{\mathrm{L}},$$
$$\log \eta - \log \mathrm{Tr}_{\mathrm{L}}(\eta) \otimes \mathrm{I}_{\mathrm{L}} = \log \rho - \log \mathrm{Tr}_{\mathrm{L}}(\rho) \otimes \mathrm{I}_{\mathrm{L}}. \qquad \text{(distributivity)}$$

Exponentiating both sides, we obtain

$$\exp\left(\log \eta - \log \mathrm{Tr}_{\mathrm{L}}(\eta) \otimes \mathrm{I}_{\mathrm{L}}\right) = \exp\left(\log \rho - \log \mathrm{Tr}_{\mathrm{L}}(\rho) \otimes \mathrm{I}_{\mathrm{L}}\right),$$
$$\eta_{\mathrm{L}|\mathrm{V}} = \rho_{\mathrm{L}|\mathrm{V}}.$$

the conditional amplitudes operators as defined in Definition 8. ∎

The Petz recovery map under the Sufficient Conditions is the conditional amplitude operator reweighted by η_{V} to give a valid density operator. This reduces to classical E step if the operators involved are diagonal as the `CAO` reduces to the conditional probability and `PRM` reduces to Bayes rule.

A key feature of the classical EM algorithm is the log-of-sums arising out of marginalization in the log-likelihood turns it into a sum-of-logs in the M-step. As discussed in Section 4, log-likelihoods that involve partial traces or projection operators are often computationally intractable. In contrast, the **M-step** of the DO-EM algorithm maximizes an objective that avoids these obstacles entirely. For a Hamiltonian-based model

$$\rho(\theta) = \exp(\mathcal{H}(\theta))/Z(\theta), \qquad \mathcal{H}(\theta) = \sum_r \theta_r \mathcal{H}_r,$$

and an E-step output $\eta^{(t)}$, the gradient of $\mathcal{Q}(\theta; \theta^{\mathrm{old}})$ in the M-step with respect to θ_r is

$$\frac{\partial}{\partial \theta_r} \mathcal{Q}(\theta; \theta^{\mathrm{old}}) = \langle \mathcal{H}_r \rangle_{\eta^{(t)}} - \langle \mathcal{H}_r \rangle_{\rho(\theta)}. \tag{6}$$

Because the M-step maximizes $\mathrm{Tr}(\eta^{(t)} \log \rho(\theta))$, its structure matches the log-likelihood of a fully visible model with no hidden units. This allows direct use of standard training techniques for fully visible density-operator models. A proof of the gradient expression, following the Lie-Trotter-Suzuki argument in Kappen (2020), appears as Corollary Theorem 33 in Appendix A.2.

As expected, the M-step optimization is far simpler than computing gradients for direct maximization of the projective or quantum log-likelihoods. As demonstrated in Section 8, the DO-EM algorithm runs roughly two orders of magnitude faster than directly maximizing the projective log-likelihood for for density operators of size $2^{10} \times 2^{10}$. In the next section, we specialize DO-LVMs and DO-EM for classical data, enabling scaling to standard ML data sets.

7 Classical-Quantum Models

In this section, we focus on adapting DO-LVMs to standard image data sets in machine learning. Training a DO-LVM assumes that the data set was generated by projective measurements on a ground truth density operator ψ. We first define such data sets.

Definition 24 (Classical Data Set) *A data set is classical if it results from repeated projective measurements of a density operator $\psi \in \mathcal{P}(\mathcal{H}_{\mathrm{V}})$ using a fixed measurement operator* $\mathcal{X} = \sum_{i=1}^{d_{\mathrm{V}}} \mathrm{x}_i \Lambda(\mathbf{u}_i)$.

Standard projective measurements of ψ with respect to $\mathcal{X}$ correspond to sampling from a probability distribution (Watrous, 2018). Specifically, outcome x_i occurs with probability

$$\Pr(\mathrm{X} = \mathrm{x}_i) = \mathrm{tr}(\psi \Lambda(\mathbf{u}_i)).$$

Thus, repeated measurements amount to drawing independent and identically distributed samples from this underlying distribution. Standard machine learning data sets can be viewed as arising from such measurements.

For a classical data set, the eigenspace of the empirical data density operator η_{V} given by Equation (3) coincides with the basis in which the projective measurements were performed. The condition $[\eta_{\mathrm{V}}, \mathrm{Tr}_{\mathrm{L}} \rho] = 0$ in the Sufficient Conditions therefore restricts the model class to those whose visible marginals share this same eigenspace. This restriction reflects a fundamental limitation of the data. Since all samples are obtained in a single fixed measurement basis, the observations reveal only the diagonal components of the underlying density operator in that basis. Any coherence or structure present in other orthonormal directions remains completely unobserved and unlearnable for the model. Enforcing that the model align with the measurement basis thus prevents it from trying to infer features that the data does not provide. We now investigate the second commutativity relation in the Sufficient Conditions, $[\rho, \rho_{\mathrm{V}} \otimes \mathrm{I}_{\mathrm{L}}] = 0$, which governs the interaction between the visible and the latent spaces.

Lemma 25 *Consider a density operator $\rho \in \mathcal{P}(\mathcal{H}_V \otimes \mathcal{H}_V)$ such that ρ_V has non-degenerate eigenvalues. Then $[\rho, \rho_V \otimes I_L] = 0$ if and only if ρ is a classical-quantum state*

$$\rho = \sum_{i=1}^{d_V} \alpha_i \Lambda(\mathbf{u}_i) \otimes \rho_L(i).$$

Proof First we prove the forward direction: if $[\rho, \rho_V \otimes I_L] = 0$ and ρ_V has non-degenerate eigenvalues, then ρ is of the form $\rho = \sum_{i=1}^{d_V} \alpha_i \Lambda(\mathbf{u}_i) \otimes \rho_L(i)$. Let the spectral decomposition of ρ_V be given by $\rho_V = \sum_{i=1}^{d_V} \alpha_i \Lambda(\mathbf{u}_i)$. Since $[\rho, \rho_V \otimes I_L] = 0$ and ρ_V is non-degenerate, $[\rho, \Lambda(\mathbf{u}_i) \otimes I_L] = 0$ for all i. This gives

$$\rho(\rho_V \otimes I_L) - (\rho_V \otimes I_L)\rho = 0,$$

$$\sum_{i=1}^{d_V} \alpha_i \left(\rho(\Lambda(\mathbf{u}_i) \otimes I_L) - (\Lambda(\mathbf{u}_i) \otimes I_L)\rho\right) = 0.$$

Multiplying by $\Lambda(\mathbf{u}_j) \otimes I_L$ on the left and $\Lambda(\mathbf{u}_k) \otimes I_L$ on the right, we obtain

$$(\alpha_i - \alpha_j)\left((\Lambda(\mathbf{u}_j) \otimes I_L)\rho(\Lambda(\mathbf{u}_k) \otimes I_L)\right) = 0,$$
$$((\Lambda(\mathbf{u}_j) \otimes I_L)\rho(\Lambda(\mathbf{u}_k) \otimes I_L)) = 0, \qquad\qquad (\alpha_j \neq \alpha_k)$$

for all $j \neq k$. Hence,

$$\rho = \sum_{i=1}^{d_V} \sum_{j=1}^{d_V} (\Lambda(\mathbf{u}_i) \otimes I_L)\rho(\Lambda(\mathbf{u}_j) \otimes I_L),$$

$$= \sum_{i=1}^{d_V} (\Lambda(\mathbf{u}_i) \otimes I_L)\rho(\Lambda(\mathbf{u}_i) \otimes I_L).$$

This is equivalent to

$$\rho = \sum_{i=1}^{d_V} \alpha_i \Lambda(\mathbf{u}_i) \otimes \rho_L(i).$$

Since ρ and ρ_V are unit trace, $\rho_L(x_i)$ are unit trace. Similarly, $\rho_L(x_i)$ is positive semi-definite.

We now prove the reverse direction: if ρ is of the form $\rho = \sum_{i=1}^{d_V} \alpha_i \Lambda(\mathbf{u}_i) \otimes \rho_L(i)$, then $[\rho, \rho_V \otimes I_L] = 0$. Applying the partial trace to $\rho = \sum_{i=1}^{d_V} \alpha_i \Lambda(\mathbf{u}_i) \otimes \rho_L(x_i)$ results in

$$\rho_V = \text{Tr}_L \sum_{i=1}^{d_V} \alpha_i \Lambda(\mathbf{u}_i) \otimes \rho_L(i) = \sum_{i=1}^{d_V} \alpha_i \Lambda(\mathbf{u}_i).$$

Since $[\rho_V, \Lambda(\mathbf{u}_i)] = 0$ and $[I_L, \rho_L(i)] = 0$ for all i, we conclude that $[\rho, \rho_V \otimes I_L] = 0$. ∎

DO-LVMs that satisfy the Sufficient Conditions for a classical data set must therefore be classical-quantum states. Furthermore, since the partial trace of the model is required to share the same eigenbasis as the measurement operator, we define the following class of DO-LVMs.

Definition 26 (CQ-LVM) *A DO-LVM is a CQ-LVM with respect to a classical data set generated by the measurement $\mathcal{X} = \sum_{i=1}^{d_V} \mathrm{x}_i \Lambda(\mathbf{u}_i)$ if it can be expressed as*

$$\rho(\theta) = \sum_{i=1}^{d_V} \mathrm{Pr}(\mathrm{X}{=}\mathrm{x}_i|\theta)\Lambda(\mathbf{u}_i) \otimes \rho_L(i|\theta), \qquad \text{(CQ-LVM)}$$

where $\rho_L(i|\theta)$ are in $\mathcal{P}(\mathcal{H}_L)$.

It is easy to see that all CQ-LVMs satisfy the Sufficient Conditions even if ρ_V has degenerate eigenvalues. The special structure of these models allow us to write specialized expressions for quantum log-likelihood and the Petz recovery map.

Lemma 27 *Consider a classical data set $\mathcal{D} = \{\mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(N)}\}$ from projective measurements with $\mathcal{X} = \sum_{i=1}^{d_V} \mathrm{x}_i \Lambda(\mathbf{u}_i)$ and a CQ-LVM $\rho(\theta) = \sum_{i=1}^{d_V} P(\mathrm{X} = \mathrm{x}_i|\theta)\Lambda(\mathbf{u}_i) \otimes \rho_L(\mathrm{x}_i|\theta)$. The quantum log-likelihood of this model is given by $\mathcal{L}_U(\eta_V, \theta) = \frac{1}{N}\sum_{i=1}^{N} \log P(\mathrm{X} = \mathrm{x}^{(i)}|\theta)$. The Petz Recovery Map for η_V with respect to $\rho(\theta^{(old)})$ is*

$$\mathcal{R}_{\rho,\mathrm{Tr}_L}(\eta_V) = \frac{1}{N}\sum_{i=1}^{N} \Lambda(\mathbf{v}^{(i)}) \otimes \rho_L(\mathrm{x}^{(i)}|\theta^{(old)}).$$

Proof As $\rho(\theta)$ is a CQ-LVM with respect to the basis $\mathcal{B}$, it satisfies the Sufficient Conditions and $[\eta_V, \rho_V(\theta)] = 0$. Since the matrices commute and can be simultaneously diagonalized, $\mathcal{L}_U(\eta_V, \theta) = \mathrm{Tr}(\eta_V \log \rho_V(\theta))$ reduces to a function of the eigenvalues of the matrices,

$$\mathcal{L}_U(\eta_V, \theta) = \sum_{i=1}^{N} \log P(\mathrm{X}{=}\mathrm{x}^{(i)}|\theta).$$

Since the Sufficient Conditions are satisfied, the Petz Recovery Map can be computed using commutativity as follows

$$\begin{aligned}
\mathcal{R}_{\rho,\mathrm{Tr}_L}(\eta_V) &= \rho^{-1/2}\left(\left(\mathrm{Tr}_B(\rho)^{-1/2}\eta_V\,\mathrm{Tr}_B(\rho)^{-1/2}\right) \otimes \mathrm{I}_B\right)\rho^{-1/2}, \\
&= \rho(\rho_V^{-1} \otimes \mathrm{I}_L)(\eta_V \otimes \mathrm{I}_L), && \text{(commutativity)} \\
&= \left(\sum_{i=1}^{d_V} \Lambda(\mathbf{u}_i) \otimes \rho_L(\mathrm{x}_i)\right)(\eta_V \otimes \mathrm{I}_L), && \text{(spectral theorem)} \\
&= \left(\sum_{i=1}^{d_V} \Lambda(\mathbf{u}_i) \otimes \rho_L(\mathrm{x}_i)\right)\left(\sum_{i=1}^{N} \Lambda(\mathbf{v}^{(i)}) \otimes \mathrm{I}_L\right), && \text{(spectral theorem)} \\
&= \frac{1}{N}\sum_{i=1}^{N} \Lambda(\mathbf{v}^{(i)}) \otimes \rho_L(\mathrm{x}^{(i)}). && \text{(distributivity)}
\end{aligned}$$

∎

Since all data points are results of the same projective measurement, the log-likelihood and the Petz Recovery Map can be calculated for each data point. Using an analogous result to the chain rule of KL divergence for CQ-LVMs proved as Theorem 34 in Appendix A.3, we decompose the QELBO into smaller terms corresponding to each data point.

Theorem 28 (Decomposition of CQ-ELBO) *For a classical data set and a correspond-ing* `CQ-LVM` $\rho(\theta)$, *the minorant function can be decomposed as* $\mathcal{Q}(\theta; \theta^{(old)}) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{Q}_i(\theta; \theta^{(old)})$ *where*

$$\mathcal{Q}_i(\theta; \theta^{(old)}) = \text{Tr}\left(\rho_{\text{L}}(\text{x}^{(i)} \mid \theta^{(old)}) \log\left(\Pr(\text{X}{=}\text{x}^{(i)} \mid \theta)\rho_{\text{L}}(\text{x}^{(i)} \mid \theta))\right)\right) + S_{\text{VN}}\left(\rho_{\text{L}}(\text{x}^{(i)} \mid \theta^{(old)})\right).$$

Proof Plugging the Petz Recovery Map into Section 6.2 allows the following simplification

$$
\begin{aligned}
\mathcal{Q}(\theta, \theta^{(\text{old})}) &= \text{QELBO}(\eta(\theta^{(\text{old})}), \rho(\theta)), \\
&= -D_{\text{U}}(\eta(\theta^{(\text{old})}), \rho(\theta)) - S_{\text{VN}}(\eta_{\text{V}}), && \text{(definition)} \\
&= -D_{\text{U}}(\eta_{\text{V}}, \rho_{\text{V}}(\theta)) - \frac{1}{N}\sum_{i=1}^{N} D_{\text{U}}(\rho_{\text{L}}(\theta^{(\text{old})}), \rho_{\text{L}}(\theta)) - S_{\text{VN}}(\eta_{\text{V}}), && \text{(Theorem 34)} \\
&= \text{Tr}(\eta_{\text{V}} \log \rho_{\text{V}}(\theta)) - \frac{1}{N}\sum_{i=1}^{N} D_{\text{U}}(\rho_{\text{L}}(\theta^{(\text{old})}), \rho_{\text{L}}(\theta)), && \text{(definition)} \\
&= \frac{1}{N}\sum_{i=1}^{N} \log \Pr(\text{x}^{(i)} \mid \theta) - \frac{1}{N}\sum_{i=1}^{N} D_{\text{U}}(\rho_{\text{L}}(\theta^{(\text{old})}), \rho_{\text{L}}(\theta)) && \text{(commutativity)} \\
&= \frac{1}{N}\sum_{i=1}^{N} \mathcal{Q}_i(\theta, \theta^{(\text{old})})
\end{aligned}
$$

as required. ∎

It is straightforward to verify that $\mathcal{Q}_i(\theta, \theta^{(\text{old})})$ is a minorant of $\log \Pr(\text{X} = \text{x}^{(i)}|\theta)$. Moreover, if the $\rho_{\text{L}}(i|\theta)$ were diagonal, this expression reduces to the classical ELBO. Theorem 27 allows us to specialize Algorithm 1 to CQ-LVMs as a mirror of the classical EM algorithm.

Algorithm 2 DO-EM for CQ-LVM

1: **Input:** $\mathcal{D} = \{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(N)}\}$ and $\theta^{(0)}$
2: **while** not converged **do**
3:    **for** $i = 1$ to N **do**
4:       $\mathcal{Q}_i(\theta; \theta^{(t)}) = \text{Tr}\left(\rho_{\text{L}}(\text{x}^{(i)} \mid \theta^{(t)}) \log P(\text{x}^{(i)} \mid \theta)\rho_{\text{L}}(\text{x}^{(i)} \mid \theta)\right) + S_{\text{VN}}(\rho_{\text{L}}(\text{x}^{(i)} \mid \theta^{(t)}))$
5:    $\theta^{(t+1)} = \text{argmax}_\theta \frac{1}{N}\sum_{i=1}^{N} \mathcal{Q}_i(\theta; \theta^{(t)})$

Algorithm 2 allows us to iterate through the dataset without any dependency on the size of η_{V} as the measurement basis is fixed. If the model $\rho(\theta)$ can be tractably computed, DO-EM allows for the tractable learning of CQ-LVMs. The inequality in Lemma 16 is saturated for CQ-LVMs. Hence, the DO-EM algorithm provably trains both the quantum and projective log-likelihoods. We now show that CQ-LVMs are a broad class of models that include several Hamiltonian-based models.

7.1 CQ Hamiltonian-based models and Contrastive Divergence

In this section, we develop Hamiltonian-based CQ-LVMs and techniques to train them at scale. We assume that the data $\mathcal{D} = \{\text{x}^{(1)}, \ldots, \text{x}^{(N)}\}$ is sampled from the set $S = \{+1, -1\}^m$.

The corresponding Hilbert space $\mathcal{H}_{\mathrm{V}}$ then has dimension 2^m with projective measurements along the standard basis. We begin by deriving a condition on the Hamiltonian that must be satisfied for a Hamiltonian-based model to be a CQ-LVM.

Lemma 29 *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-LVM *with respect to the basis $\{\mathbf{e}_i\}_{i=1}^{2^m}$ if and only if $\mathrm{H} = \oplus_i \mathrm{H}_i$ where H_i are Hermitian operators in $\mathcal{T}(\mathcal{H}_{\mathrm{L}})$.*

Proof Note that a CQ-LVM with respect to the basis $\{\mathbf{e}_i\}_{i=1}^{2^m}$ can be expressed as

$$\rho = \sum_{i=1}^{2^m} \Pr(\mathrm{X}=\mathrm{x}_i)\Lambda(\mathbf{e}_i) \otimes \rho_{\mathrm{L}}(\mathrm{x}_i) = \oplus_{i=1}^{2^m} \Pr(\mathrm{x}_i)\rho_{\mathrm{L}}(\mathrm{x}_i). \tag{7}$$

We first prove that a Hamiltonian of the form $\mathrm{H} = \oplus_i \mathrm{H}_i$ leads to a model of the form Equation (7). The Taylor series of the matrix exponential is given by $e^X = \sum_{k=0}^{\infty} \frac{1}{k!}X^k$. Since H is a block diagonal matrix, the powers of H are also block diagonal. Therefore ρ is also a block diagonal matrix as required.

We now prove that a model of the form Equation (7) arises from a Hamiltonian of the form $\mathrm{H} = \oplus_i \mathrm{H}_i$. The Taylor series of the matrix logarithm is given by $\log X = -\sum_{k=1}^{\infty} \frac{1}{k}(I - X)^k$. If $(\log Z(\theta))\rho = e^{\mathrm{H}(\theta)}$ is a block diagonal matrix, then its logarithm $\mathrm{H}(\theta)$ is also a block diagonal matrix as required. ∎

We now recall QBMs from Section 4. This block diagonal structure of the model allow the specialization enables us to specialize QBMs to CQ-LVMs.

Corollary 30 *A* QBM *with m visible and n hidden units given by the Hamiltonian*

$$\mathrm{H}(\theta) = \begin{cases} -\sum_{i=1}^{m}\mathbf{a}_i\sigma_i^{(z)} - \sum_{i=1}^{n}\mathbf{b}_i\sigma_{m+i}^{(z)} - \sum_{i=1}^{m+n}\Gamma_i\sigma_i^{(x)} \\ -\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{W}_{ij}\sigma_i^{(z)}\sigma_j^{(z)} - \sum_{i=1}^{m}\sum_{j=1}^{m}\mathbf{W}_{ij}^{(\mathrm{V})}\sigma_i^{(z)}\sigma_j^{(z)} - \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{W}_{ij}^{(\mathrm{L})}\sigma_{m+i}^{(z)}\sigma_{m+j}^{(z)} \end{cases}$$

is a CQ-LVM if and only if quantum terms on the visible units are zero.

Proof Theorem 29 implies that the Hamiltonian is CQ if and only if it is a direct sum of 2^m matrices of size $2^n \times 2^n$. The only off-diagonal entries of the Hamiltonian are from the quantum bias terms Γ_i. Recall that $\sigma_i^{(x)}$ are defined as

$$\sigma_i^k = \otimes_{j=1}^{i-1}I \otimes \sigma^k \otimes_{j=i+1}^{m+n} I$$

where $\sigma^{(x)} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ and $I = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$. It is easy to see that $\sigma_i^{(x)} = \oplus_{j=1}^{i}\left(\sigma^k \otimes_{j=i+1}^{m+n} I\right)$ for $i > m$ which satisfies the condition in Theorem 29.

To prove the converse, we observe that the Kronecker product $\sigma^{(x)} \otimes M = \left(\begin{smallmatrix} 0 & M \\ M & 0 \end{smallmatrix}\right)$ for any square matrix M. When $i \leq m$, this results in matrices larger than $2^n \times 2^n$ with non-zero entries which cannot then be represented as a direct sum of $2^n \times 2^n$ matrices. Hence, $\Gamma_i = 0$ if $i \leq m$. ∎

While, we showcase one set of Hamiltonians that are CQ-LVMs, several other Hamiltonians leading to CQ-LVMs can be constructed. Notably, the class of semi-quantum RBMs studied in Demidik et al. (2025) are CQ-LVMs. While the Hamiltonians of CQ-LVMs have additional structure, training such models using real-world data is intractable since the free energy term, $-\log Z(\theta)$, cannot be easily computed even for classical Boltzmann machines.

In Boltzmann machines, this issue is typically side-stepped by the Contrastive Divergence algorithm discussed in Section 2. To achieve tractable training of `QBMs`, we adapt the CD algorithm to the Hamiltonian of a Quantum RBM (`QRBM`) with transverse field only in the hidden units given by

$$H(\theta) = -\sum_{i=1}^{m} \mathbf{a}_i \sigma_i^{(z)} - \sum_{i=1}^{n} \mathbf{b}_i \sigma_{m+i}^{(z)} - \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{W}_{ij} \sigma_i^{(z)} \sigma_{m+j}^{(z)} - \sum_{i=1}^{n} \Gamma_i \sigma_{m+i}^{(z)} \qquad \text{(QRBM)}$$

where $\mathbf{a}$ in $\mathbb{R}^m$ and $\mathbf{b}$ in $\mathbb{R}^n$ are the usual bias terms, $\mathbf{W}$ in $\mathbb{R}^{m \times n}$ is the weight matrix between the two layers, and Γ in $\mathbb{R}^n$ is the transverse bias term. If the non-quantum visible layer is fixed to x, with corresponding Hilbert space vector $\mathbf{v}$, the hidden units of the hidden layer are conditionally independent. The Hamiltonian of the jth unit of the hidden layer L is given by

$$H_{\text{L}}(j|\text{x}, \theta) = -\mathbf{b}_j^{\text{eff}} \sigma^{(z)} - \Gamma_j \sigma^{(x)}$$

with $\mathbf{b}_j^{\text{eff}} = \mathbf{b}_j + \sum_{i=1}^{m} \mathbf{W}_{ij} \text{x}_i$. The Hamiltonian decomposes into n independent 2×2 matrices and allows for the tractable sampling from the quantum layer using the expected values

$$\langle \sigma_j^{(z)} \rangle_{\mathbf{v}} = \frac{\mathbf{b}_j^{\text{eff}}}{D_j} \tanh D_j \text{ and } \langle \sigma_j^{(x)} \rangle_{\mathbf{v}} = \frac{\Gamma_j}{D_j} \tanh D_j \qquad (8)$$

where $D_j = \sqrt{(\mathbf{b}_j^{\text{eff}})^2 + \Gamma_j^2}$. The Gibbs step for the non-quantum layers can be performed as per the classical CD algorithm using the quantum sample along the Z-basis. This closed-form expression for Gibbs sampling without matrices allows CD to run on a `QRBM` with the same memory footprint as an RBM. This scheme can easily be generalized to the class of models considered in (Demidik et al., 2025). We provide the justification for the sampling step stated here in Appendix A.4.

7.2 Novel QBM Variants

In this section, we extend the notion of quantum RBMs to Deep Boltzmann Machines and Gaussian-Bernoulli RBMs introducded in Section 2 and adapt the Gibbs sampling step described above to these new models.

7.2.1 QUANTUM INTERLEAVED DBM

A **Q**uantum **I**nterleaved **D**eep **B**oltzmann **M**achine (`QiDBM`) is a Deep Boltzmann Machine with quantum bias terms on **non-contiguous hidden layers**. It can be viewed as a set of stacked classical-quantum RBMs. We describe the Hamiltonian of a three-layered `QiDBM` 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.

$$
H(\theta) = \begin{cases}
-\sum_{i=1}^{m} \mathbf{a}_i \sigma_i^{(z)} - \sum_{i=1}^{m} \mathbf{b}_i^{(1)} \sigma_{\ell+i}^{(z)} - \sum_{i=1}^{n} \mathbf{b}_i^{(2)} \sigma_{\ell+m+i}^{(z)} - \sum_{i=1}^{m} \Gamma_i \sigma_{\ell+i}^{(x)} \\
-\sum_{i=1}^{\ell} \sum_{j=1}^{m} \mathbf{W}_{ij}^{(1)} \sigma_i^{(z)} \sigma_{\ell+j}^{(z)} - \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{W}_{ij}^{(2)} \sigma_{\ell+i}^{(z)} \sigma_{\ell+m+j}^{(z)}
\end{cases} \tag{QiDBM}
$$

We illustrate the case of the middle layer of $\texttt{QiDBM}_{\ell,\mathtt{m},\mathtt{n}}$. If the non-quantum visible and hidden layers are fixed to x and $z_{[2]}$, with corresponding Hilbert space vectors $\mathbf{v}$ and $\mathbf{h}^{(2)}$, the hidden units of the quantum layer are conditionally independent. The Hamiltonian for the jth unit of the quantum layer is given by $H^{(1)}(j|\mathrm{x}, z_{[2]}, \theta) = -\mathbf{b}_j^{\text{eff}} \sigma^{(z)} - \Gamma_i \sigma^{(x)}$. We sample from the quantum layer using the expected values

$$
\langle \sigma_j^{(z)} \rangle_{\mathrm{x},z_{[2]}} = \frac{\mathbf{b}_j^{\text{eff}}}{D_j} \tanh D_j \text{ and } \langle \sigma_j^{(x)} \rangle_{\mathrm{x},z_{[2]}} = \frac{\Gamma_j}{D_j} \tanh D_j
$$

where $D_j = \sqrt{(\mathbf{b}_j^{\text{eff}})^2 + \Gamma_j^2}$ and $b_j^{\text{eff}} = b_j + \sum_{i=1}^{\ell} \mathbf{W}_{ij}^{(1)} \mathrm{x}_i + \sum_i^n \mathbf{W}_{ji}^{(2)} z_{[2],i}$. The Gibbs step for the non-quantum layers is done as per the classical CD algorithm using the quantum sample from the Z-basis.

7.2.2 Quantum Gaussian-Bernoulli RBM

A **Q**uantum **G**aussian-Bernoulli **R**estricted **B**oltzmann **M**achine ($\texttt{QGRBM}$) is a $\texttt{GRBM}$ modified to include the transverse field over the Hidden units. Since the visible data is sampled from $\mathbb{R}$, the visible Hilbert space $\mathcal{H}_{\mathrm{V}}$ should be infinite-dimensional. All separable infinite-dimensional Hilbert spaces over $\mathbb{C}$ has dimension equal to the cardinality of the real numbers (Conway, 1990). We fix $\mathcal{H}_{\mathrm{V}}$ as the space of square-summable complex sequences and associate each real number x with a corresponding basis vector $\mathbf{x}$ in an orthonormal basis of $\mathcal{H}_{\mathrm{V}}$.

The Hamiltonian of the finite-dimensional space $\mathcal{H}_{\mathrm{L}}$ with respect to x in $\mathbb{R}^m$ is

$$
H(\mathrm{x}, \theta) = -\sum_{i=1}^{m} \frac{(\mathrm{x}_i - \mathbf{a}_i)^2}{\mathbf{s}_i^2} I_{\mathrm{L}} - \sum_{j=1}^{n} \mathbf{b}_j^{\text{eff}} \sigma_j^{(z)} - \sum_{j=1}^{n} \Gamma_j \sigma_j^{(x)}, \tag{QGRBM}
$$

where $\mathbf{b}_j^{\text{eff}} = \mathbf{b}_j + \sum_{i=1}^{m} \mathbf{W}_{ij} \frac{\mathrm{x}_i}{\mathrm{s}_i}$. The model density operator ρ is then obtained by integrating over the visible space

$$
\rho(\theta) = \frac{1}{\mathcal{Z}(\theta)} \int_{\mathbf{x}} \Lambda(\mathbf{x}) \otimes \exp(H(\mathrm{x}, \theta)) d\mathbf{x},
$$

$$
\mathcal{Z}(\theta) = \int_{-\infty}^{+\infty} \mathrm{Tr} \exp(H(\mathrm{x}, \theta)) d\mathrm{x}.
$$

The probability density function of the visible units is given by $\Pr(\mathrm{x}|\theta) = \frac{\mathrm{Tr} \exp(H(\mathrm{x},\theta))}{\mathcal{Z}(\theta)}$. If the visible layer is fixed to x, the Hamiltonian for the the jth unit of the hidden layer L is given by $H_{\mathrm{L}}(j|\mathrm{x}, z_{[2]}, \theta) = -\mathbf{b}_j^{\text{eff}} \sigma^{(z)} - \Gamma_i \sigma^{(x)}$. Much like in the $\texttt{QRBM}$ and $\texttt{QiDBM}$ models, sampling over these qubits remains efficient, allowing CD to approximate the gradients. The visible units can are drawn from the Gaussian distribution defined in Section 2, using the hidden-qubit samples obtained in the Z-basis.

8 Experiments

In this work, we introduced Classical-Quantum LVMs—a class of DO-LVMs—and a general EM framework, DO-EM, for learning them. In this section, we empirically evaluate our methods with experiments designed to answer the following questions.

(Q1) **Effectiveness of DO-EM.** Do CQ-LVMs trained with Algorithm 1 provide any computational advantage over gradient-based approaches in training DO-LVMs?

(Q2) **Scalability of DO-EM.** Can CQ-LVMs trained with Algorithm 2 scale to standard ML data sets such as MNIST?

(Q3) **Performance of DO-EM.** Do CQ-LVMs trained with Algorithm 2 provide a measurable improvement over probabilistic LVMs on standard ML data sets?

To address these questions, we train variants of the `QBM` proposed in this work and compare them against benchmarks reproduced using reference implementations from literature. Code for all experiments is available at `https://github.com/aditvishnu/DO-EM`. Details of the computing infrastructure used for all experiments are provided in Appendix B.

8.1 DO-EM vs Gradient-based Maximization of Projective Log-Likelihood

We conduct experiments using *exact log-likelihood computation* to compare DO-EM with gradient-based maximization of the projective log-likelihood. A `QBM` satisfying the Sufficient Conditions and trained using Algorithm 1 is compared against a semi-restricted `QBM` trained via gradient-based maximization of projective log-likelihood (`PL`), following the approach of Amin et al. (2018).

The `QBMs` are trained on a mixture of Bernoulli data set introduced in Amin et al. (2018). A Bernoulli distribution is built on a randomly selected mode and the average probability distribution over M such modes determines the mixture of Bernoulli distribution. The distribution can be written as

$$P_v^{\text{data}} = \frac{1}{M} \sum_{k=1}^{M} p^{m-d_v^k} (1-p)^{d_v^k}$$

where m is the dimension of data, p is the probability of success, and d_v^k is the Hamming distance between v and the randomly selected mode. We use 100 randomly sampled data sets of 1000 samples each with $N = 1000$, $M = 8$, $m = 8$, and $p = 0.9$. A model with $m = 8$ visible units and $n = 2$ hidden units with random initialization is trained on all 100 datasets with a learning rate of 0.1, and we plot the resulting average log-likelihood in Figure 1a. For the semi-restricted QBM, where the Projective and Umegaki log-likelihoods differ, we show both curves in the plot.

Algorithm 1 executes a pair of EM steps in only *0.2 seconds*, while direct gradient ascent on the projective log-likelihood requires approximately *40 seconds* per step. Thus, DO-EM offers a significant speed-up for training `QBMs` in the exact computation setting. This efficiency can be further enhanced by employing Algorithm 2 to train the `QBM`. As shown in Figure 1a, the model trained with DO-EM *converges in roughly half the number of steps*

(a) Log-likelihood during training with exact computation showing DO-EM achieving comparable results faster.

(b) FID on MNIST on a QiDBM (6272 units per hidden layer) and DBMs (6272 and 6273 units per hidden layer).

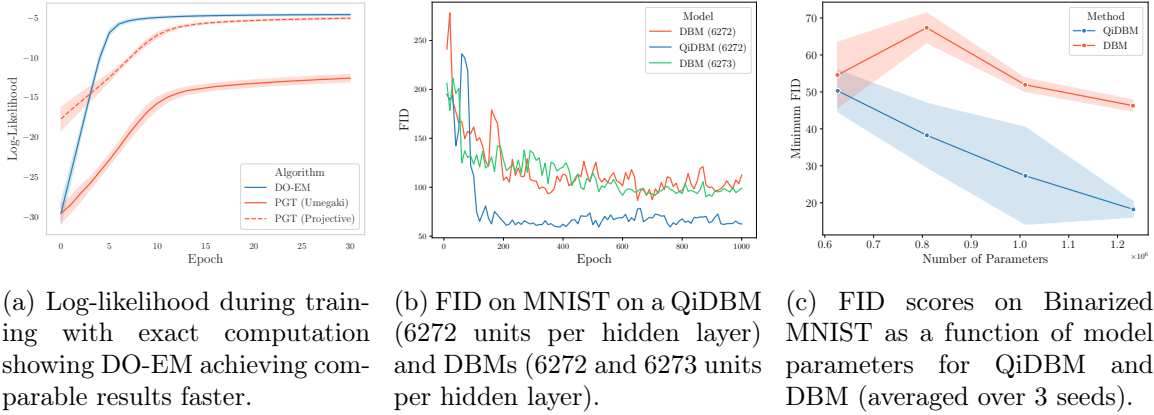(c) FID scores on Binarized MNIST as a function of model parameters for QiDBM and DBM (averaged over 3 seeds).

Figure 1: (a) DO-EM vs Proejctive log-likelihood gradient-based training (PGT) with exact computation on mixture of Bernoulli data set. (b,c) QiDBM vs DBM on MNIST with CD.

compared to direct gradient ascent while reaching *comparable log-likelihood values*. These findings position DO-EM as a strong candidate for `QBM` training, offering clear empirical gains in computational resources while maintaining similar performance.

8.2 CQ-LVMs on Discrete Data

To showcase the scalability and performance of CQ-LVMs using DO-EM, we train Quantum interleaved Deep Boltzmann Machines (`QiDBM`) using contrastive divergence to approximate the M-step gradient. We compare our method with Taniguchi et al. (2023), the state of the art for training DBMs. Since we were unable to reproduce their reported results, we use results obtained from their official implementation[1] with the hyperparameters recommended in their paper. *No additional hyperparameter tuning* is performed on the quantum variant.

Following Taniguchi et al. (2023), we use a $\{0, 1\}$ encoding instead of $\{-1, +1\}$ to perform experiments on MNIST and Binarized MNIST (Deng, 2012). The data sets contain 60,000 training and 10,000 testing images of size 28×28. We evaluate the quality of generated images using the Fréchet Inception Distance (FID, Seitzer, 2020), computed between 10,000 generated samples and the MNIST test set.

8.2.1 MNIST

To demonstrate the scalability of CQ-LVMs on standard ML data sets under the DO-EM scheme, we train `Qi-DBMs` on the MNIST dataset. Following Taniguchi et al. (2023), each 8-bit image is mapped to 6272 binary visible units. We compare a three-layer `QiDBM`—with quantum-bias terms in the middle layer and 6272 units per hidden layer—against two three-layer DBMs: one with 6272 units per hidden layer and another with 6273 units per hidden layer. The `QiDBM` has 78.70M parameters while the DBMs have 78.69M and 78.71M parameters respectively. Training is carried out for 1000 epochs using stochastic gradient descent with Contrastive Divergence, a batch size of 600, and a learning rate of 0.001.

---

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

(a) DBM at 175 epochs

(b) QiDBM at 175 epochs

(c) DBM at 206 epochs

(d) QiDBM at 206 epochs

(e) DBM at 337 epochs

(f) QiDBM at 337 epochs

(g) DBM at 679 Epochs
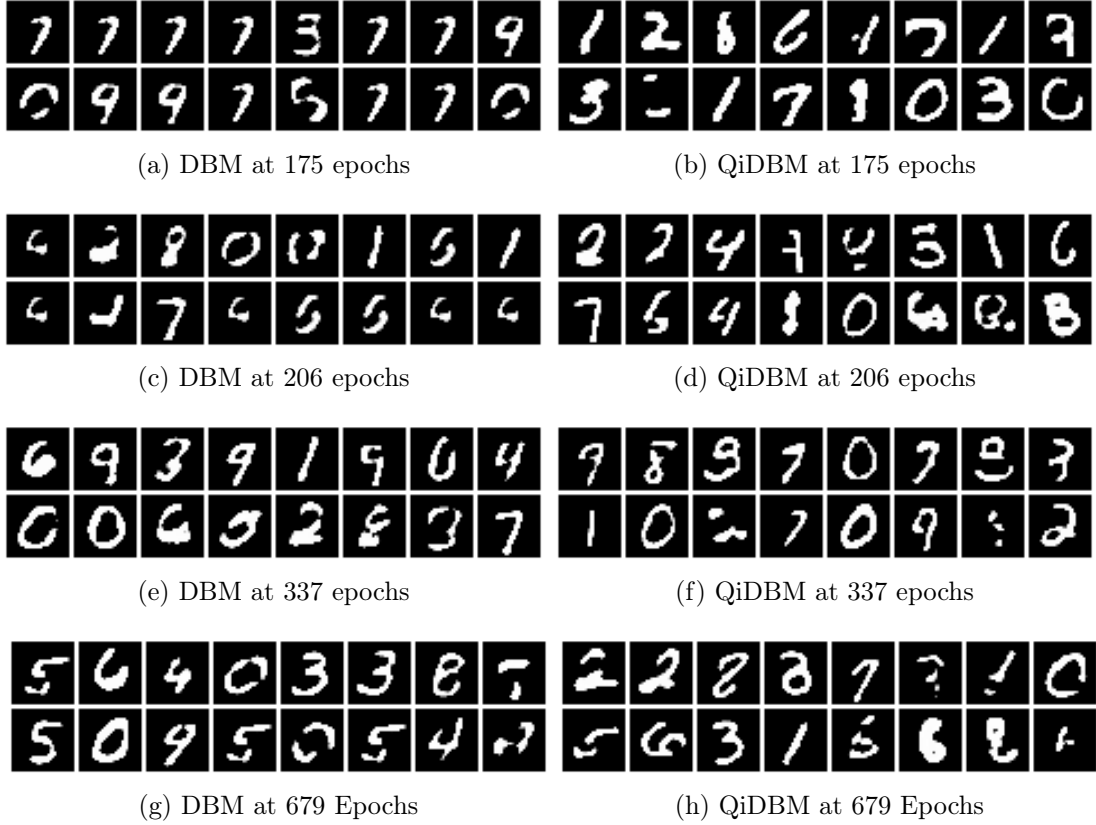
(h) QiDBM at 679 Epochs

Figure 2: Generated samples during training: DBM (left) QiDBM (right).

The FID values of generated samples, recorded every 10 epochs, are plotted in Figure 1b. The proposed method consistently outperforms both classical models of comparable size, achieving a *45% reduction in FID*. The `QiDBM` *converges within 400 epochs*, whereas both DBMs remain unstable beyond 500 epochs. The final FID scores are 62.77 for the `QiDBM`, compared to 111.73 and 99.17 for the two DBMs.

In Figure 2, we present qualitative comparisons of generated samples at randomly selected checkpoints during training. Notably, `QiDBM` samples appear *more coherent than those from DBMs* as early as 250 epochs. These findings show that scaling `Qi-DBMs` to data sets such as MNIST is feasible and leads to *substantial performance gains*.

8.2.2 BINARIZED MNIST

To further assess the impact of the additional quantum bias terms, we compare `Qi-DBMs` and DBMs across different model sizes. `QiDBM` and DBM models with two hidden layers are trained on 1-bit MNIST, with 490, 588, 686, and 784 units in each layer. Training is carried out for 100 epochs using stochastic gradient descent with Contrastive Divergence, a batch size of 600, and a learning rate of 0.001. Each setting is repeated over three random seeds to account for variability.

The minimum FID over 100 epochs against the number of parameters averaged over the three repeats is graphed in Figure 1c. Models with quantum bias consistently outperform their classical counterparts, achieving *similar FID scores with significantly fewer parameters.* `QiDBMs` outperforms the DBMs in all configurations achieving a minimum FID of 14.77 compared to the DBM's 42.61, *a 65% improvement in generation quality.*

8.3 CQ-LVMs on Continuous Data

To assess the applicability of the DO-EM algorithm to continuous data, we apply Algorithm 2 and Contrastive Divergence to Quantum Gaussian-Bernoulli RBMs (`QGRBM`). We compare our method with Liao et al. (2022). We use results obtained from their official implementation[2] with the hyperparameters recommended in their paper. *No additional hyperparameter tuning* is performed on the quantum variant.

Following Liao et al. (2022), we use a $\{0,1\}$ encoding in the hidden units instead of $\{-1,+1\}$ to perform experiments on Fashion MNIST (Xiao et al., 2017) and CelebA (Liu et al., 2015). The Fashion MNIST data set contains 60,000 training and 10,000 testing images of size 28×28. The CelebA-32 data set contains 162,770 training and 19,962 testing images downsampled to size 32×32. We assess the quality of the generated images using the Fréchet Inception Distance (FID, Seitzer, 2020), computed between generated samples and the test set with the same number of images generated as size of test set.
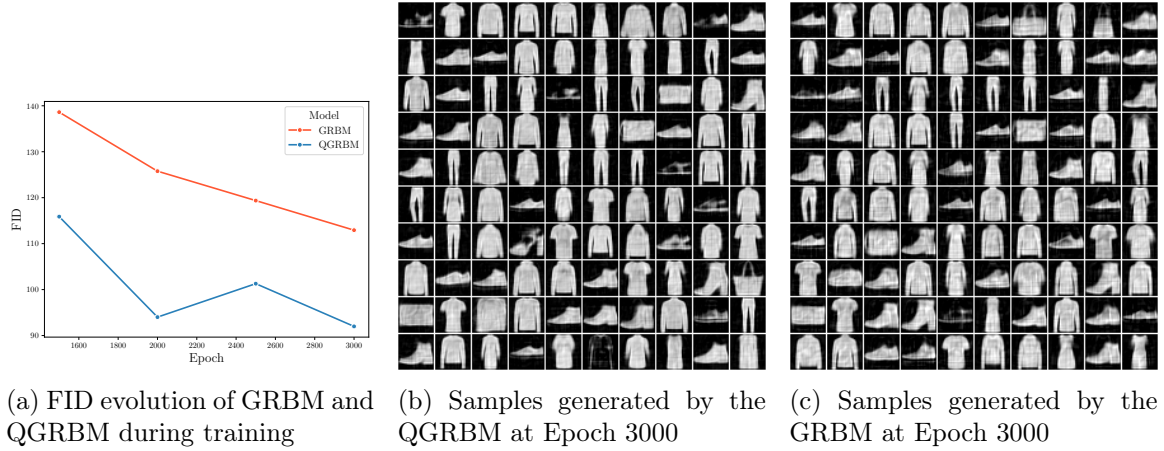

(a) FID evolution of GRBM and QGRBM during training

(b) Samples generated by the QGRBM at Epoch 3000

(c) Samples generated by the GRBM at Epoch 3000

Figure 3: Fashion MNIST

8.3.1 Fashion MNIST

We train a `QGRBM` with 10,000 hidden units on the Fashion-MNIST dataset and compare it to a GRBM trained with Contrastive Divergence under the same conditions following Liao et al. (2022). Training is carried out for 3000 epochs using stochastic gradient with Contrastive Divergence, a batch size of 512, and a learning rate of 0.01. The FID values of generated samples are plotted in Figure 3a. The `QGRBM` achieves *a 20% reduction in FID.*

---

2. `https://github.com/lrjconan/GRBM`

An illustrative comparison of generated samples is provided in Figures 3b and 3c, where we observe noticeably *higher sample diversity* in `QGRBM` relative to GRBM.

(a) FID evolution of GRBM and QGRBM during training

(b) Samples generated by the QGRBM at Epoch 900

(c) Samples generated by the GRBM at Epoch 900

Figure 4: CelebA-32

8.3.2 CELEBA-32

We train a `QGRBM` on the Fashion-MNIST dataset and compare it to a GRBM trained with Contrastive Divergence under the same conditions, following Liao et al. (2022). with 10,000 hidden units for 3000 epochs. Training is carried out for 3000 epochs using stochastic gradient with Contrastive Divergence, a batch size of 512, and a learning rate of 0.01. The FID values of generated samples are plotted in Figure 4a. The `QGRBM` achieves a *30% reduction in FID*. An illustrative comparison of generated samples is provided in Figures 4b and 4c, where we observe *noticeably higher sample diversity* in `QGRBM` relative to GRBM.

8.4 Discussion

The experiments provide clear answers to the three motivating questions.

(Q1) **Effectiveness of DO-EM.** The exact-likelihood studies show that DO-EM accelerates `QBM` training by nearly two orders of magnitude relative to gradient-based optimization, while achieving comparable log-likelihoods. This demonstrates a tangible computational advantage in settings where exact inference is feasible.

(Q2) **Scalability of DO-EM.** The MNIST and Binarized MNIST results confirm that CQ-LVMs scale reliably to models with more than 18,000 units. `QiDBMs` trained under DO-EM remain stable throughout training and attain substantially lower FID scores—up to a 65% improvement—indicating that the method handles standard ML data sets.

(Q3) **Performance of DO-EM.** experiments on both discrete and continuous domains show that DO-EM trained CQ-LVMs outperform classical probabilistic LVMs of comparable size. `QiDBMs` and `QGRBMs` consistently achieve lower FID scores and exhibit higher sample diversity, reflecting a measurable improvement in generative quality.

Taken together, these results demonstrate that DO-EM is computationally effective, scalable to standard ML data sets, and capable of delivering superior generative performance relative to their probabilistic counterparts.

9 Conclusion

This work advances density operator modeling by introducing DO-EM, an EM algorithm for density operator latent variable models that provides a formal guarantee of log-likelihood ascent.We further introduce CQ-LVMs, a broad family of DO-LVMs uniting probabilistic visible variables with quantum latents, for which DO-EM provides a natural training procedure on standard ML datasets. Our experiments show that `QiDBMs` and `QGRBMs`—our newly introduced CQ-LVMs—scale effectively to image datasets, providing the first image generation benchmark for a `QBM`. Across these tasks, they consistently outperform comparable probabilistic models under similar computational budgets and hyperparameter settings, indicating that density-operator formulations can yield tangible performance gains. Finally, the close connection between DO-EM and quantum-information tools like the Petz recovery map positions it as a promising candidate for future quantum-hardware implementations, which we defer to subsequent work.

Appendix A. Matrix Analysis

This appendix summarizes the essential matrix analysis tools used throughout the manuscript. For a more comprehensive treatment, see Bhatia (1997).

We consider only finite-dimensional vector spaces of dimension d. The inner product between two elements of such a vector space is denoted by $\langle u, v \rangle$. Inner products are conjugate linear in the first variable and linear in the second variable. A vector space with an inner product is a finite-dimensional Hilbert space if $\langle u, u \rangle = 0$ if and only if $u = 0$. We denote the space of linear operators from a Hilbert space $\mathcal{H}$ to itself by $\mathcal{T}(\mathcal{H})$. if the basis of $\mathcal{H}$ is fixed, then every operator in $\mathcal{T}(\mathcal{H})$ has a unique matrix associated with it.

An operator A in $\mathcal{T}(\mathcal{H})$ is Hermitian if $A = A^\dagger$. The spectrum of such an operator is denoted $\mathrm{Sp}(A)$. A Hermitian operator A is positive semi-definite if $\langle x, Ax \rangle \geq 0$ for all x in $\mathcal{H}$. The operator is positive definite if $\langle x, Ax \rangle > 0$ for all non-zero x in $\mathcal{H}$. According to the spectral theorem, a Hermitian operator A can be expressed as

$$A = \sum_{i=1}^{d} \lambda_i \Lambda(v_i)$$

where (λ_i, v_i) are eigenvalue-eigenvector pairs of A. A real valued function $f : \mathbb{R} \to \mathbb{R}$ can be naturally extended to act on operators via the spectral theorem,

$$f(A) = \sum_{i=1}^{d} f(\lambda_i) \Lambda(v_i).$$

Several inequalities related to convex functions can be extended to such functions. We provide the finite-dimensional restatement of an inequality in Dragomir (2016).

Theorem 31 (Corollary 2.1 in Dragomir, 2016) *Let A be a Hermitian operator on a Hilbert space $\mathcal{H}$ and assume that $\mathrm{Sp}(A) \subseteq [m, M]$ for some scalars m, M with $m < M$. If f is a continuous convex function on $[m, M]$ and P be a non-zero positive semi-definite operator in $\mathcal{T}(\mathcal{H})$, then $m \leq \frac{\mathrm{Tr}(PA)}{\mathrm{Tr}(P)} \leq M$ and*

$$f\left(\frac{\mathrm{Tr}(PA)}{\mathrm{Tr}(P)}\right) \leq \frac{\mathrm{Tr}(Pf(A))}{\mathrm{Tr}(P)}.$$

A.1 Matrix exponential and logarithm

In this appendix, we specialize to the matrix exponential and logarithm functions, the most common functions used in this manuscript. The exponential function extended to all Hermitian operators is given by

$$\exp(A) = \sum_{k=0}^{\infty} \frac{1}{k} A^k.$$

The exponential of a Hermitian operator A is positive definite. The inverse of the exponential function, the logarithm, can be defined on positive definite matrices as

$$\log(A) = -\sum_{k=1}^{\infty} \frac{1}{k} (I - A)^k.$$

Let X be in $\mathcal{T}(\mathcal{H}_A)$ and Y be in $\mathcal{T}(\mathcal{H}_B)$. We state a few binary matrix operations and their relationship with the exponential function. The direct sum of X and Y is given by
$X \oplus Y = \begin{pmatrix} X & 0 \\ 0 & Y \end{pmatrix}$ The exponential of $X \oplus Y$ is given by $\exp(X) \oplus \exp(Y)$. The tensor or

Kronecker product is given by $X \otimes Y = \begin{pmatrix} X_{11}Y & \dots & X_{1d_A}Y \\ \vdots & \vdots & \vdots \\ X_{d_A 1}Y & \dots & X_{d_A d_A}Y \end{pmatrix}$ where X_{ij} are entries of

the matrix X. The following lemma can be proved using the spectral theorem applied to the Kronecker product.

Lemma 32 *If X is a Hermitian operator in $\mathcal{T}(\mathcal{H}_A)$, $\exp(A) \otimes \mathrm{I}_B = \exp(A \otimes \mathrm{I}_B)$. Similarly, if M is positive definite operator in $\mathcal{T}(\mathcal{H}_A)$, $\log(M) \otimes \mathrm{I}_B = \log(M \otimes \mathrm{I}_B)$.*

A.2 M Step Gradient

In this appendix, we provide a proof of Equation (6) presented in Kappen (2020).

Corollary 33 *For a Hamiltonian-based model $\rho(\theta) = \exp(\mathrm{H}(\theta))/Z(\theta)$ with $\mathrm{H}(\theta) = \sum_r \theta_r \mathrm{H}_r$ and and E step solution $\eta^{(t)}$, the gradient of $\mathcal{Q}(\theta; \theta^{(old)})$ in the M-step of Algorithm 1 with respect to the parameter θ_r is given by*

$$\frac{\partial}{\partial \theta_r} \mathcal{Q}(\theta; \theta^{(old)}) = \langle \mathrm{H}_r \rangle_{\eta^{(t)}} - \langle \mathrm{H}_r \rangle_{\rho(\theta)} .$$

Proof The expression for $\mathcal{Q}(\theta; \theta^{(\mathrm{old})})$ is given by

$$\mathcal{Q}(\theta; \theta^{(\mathrm{old})}) = \mathrm{Tr}(\eta^{(\mathrm{old})} \log \rho(\theta)) + S(\eta^{(\mathrm{old})}) - S(\eta_{\mathrm{V}}).$$

We note that the latter two terms in the expression are not dependent on θ_r. The logarithm of the model is given by $\log \rho(\theta) = \sum_r \theta_r \mathrm{H}_r - \log(Z(\theta))\mathrm{I}$. The gradients can then be written as

$$\frac{\partial}{\partial \theta_r} \mathrm{Tr}(\eta^{(\mathrm{old})} \sum_r \theta_r \mathrm{H}_r) = \mathrm{Tr}(\eta^{(\mathrm{old})} \mathrm{H}_r), \text{ and}$$

$$\frac{\partial}{\partial \theta_r} \mathrm{Tr}(\eta^{(\mathrm{old})} \log Z(\theta)\mathrm{I}) = \frac{\partial}{\partial \theta_r} \log \mathrm{Tr} \exp(\mathrm{H}(\theta)).$$

We know by the Lie-Trotter-Suzuki formula that $e^{\mathrm{H}(\theta)} = \lim_{t \to \infty} \left(e^{\mathrm{H}(\theta)/m} \right)^m$. Hence,

$$\frac{\partial}{\partial \theta_r} \log \mathrm{Tr} \exp(\mathrm{H}(\theta)) = \frac{1}{\mathrm{Tr} \exp(\mathrm{H}(\theta))} \mathrm{Tr} \left(\int_0^1 dt e^{\mathrm{H}(\theta)t} \mathrm{H}_r e^{\mathrm{H}(\theta)(1-t)} \right),$$

$$= \frac{1}{\mathrm{Tr} \exp(\mathrm{H}(\theta))} \mathrm{Tr}(e^{\mathrm{H}(\theta)} \mathrm{H}_r),$$

$$= \mathrm{Tr}(\rho(\theta) \mathrm{H}_r).$$

Collecting together the two partial derivatives, we obtain

$$\frac{\partial}{\partial \theta_r} \mathcal{Q}(\theta; \theta^{(\mathrm{old})}) = \mathrm{Tr}(\eta^{(t)} \mathrm{H}_r) - \mathrm{Tr}(\rho(\theta) \mathrm{H}_r)$$

which is the required result. ■

A.3 Chain Rule for Classical-Quantum States

In this appendix, we prove a chain rule for the Umegaki relative entropy of two classical-quantum states with respect to the same orthonormal basis.

Theorem 34 (Chain Rule for Classical-Quantum States) *Suppose η and ρ are two classical-quantum states on $\mathcal{H}_A \otimes \mathcal{H}_B$ with respect to the same orthonormal basis of $\mathcal{H}_A$ given by*

$$\eta = \sum_{i=1}^{d_A} q(X = x_i)\Lambda(\mathbf{u}_i) \otimes \eta_L(x_i) \ \text{and} \ \rho = \sum_{i=1}^{d_A} p(X = x_i)\Lambda(\mathbf{u}_i) \otimes \rho_L(x_i).$$

Then,

$$D_U(\eta, \rho) = D_{KL}(q(X), p(X)) + \sum_{x_i} q(X = x_i)D_U(\eta_L(x_i), \rho_L(x_i)).$$

Proof Rewrite the density operators as

$$\eta = \sum_{i=1}^{d_A} \Lambda(\mathbf{u}_i) \otimes q(X = x_i)\eta_L(x_i) \ \text{and} \ \rho = \sum_{i=1}^{d_A} \Lambda(\mathbf{u}_i) \otimes p(X = x_i)\rho_L(x_i).$$

For any positive real c and Hermitian operator $\mathcal{O}$, $\log(c\mathcal{O}) = \log(c)I + \log(\mathcal{O})$. Hence,

$$\mathrm{Tr}(\eta \log \eta) = \sum_{x_i} q(X = x_i) \log q(X = x_i) + q(X = x_i) \log \eta_L(x_i),$$

$$\mathrm{Tr}(\eta \log \rho) = \sum_{x_i} q(X = x_i) \log p(X = x_i) + q(X = x_i) \log \rho_L(x_i).$$

Combining the two expressions and using the definitions of the KL divergence and the Umegaki relative entropy, we obtain

$$D_U(\eta, \rho) = D_{KL}(q(X), p(X)) + \sum_{x_i} q(X = x_i)D_U(\eta_L(x_i), \rho_L(x_i)),$$

which is the required expression. ∎

The chain rule provides a condition for when Theorem 12 is saturated under the partial trace for classical-quantum states.

A.4 Gibbs Sampling

In this appendix, we provide a justification for the Gibbs sampling step in Equation (8). We first state a useful result. The Kronecker sum of two matrices X in $\mathcal{T}(\mathcal{H}_A)$ and Y in $\mathcal{T}(\mathcal{H}_B)$ is given by $X \otimes I_B + I_A \otimes Y$. Since $[X \otimes I_B, I_A \otimes Y] = 0$, we have from Lemma 32 that

$$\exp(X \otimes I_B + I_A \otimes Y) = \exp(X) \otimes \exp(Y). \tag{9}$$

Consider a Quantum RBM (`QRBM`) with m visible and n hidden units given by the Hamiltonian

$$\mathrm{H}(\theta) = -\sum_{i=1}^{m} \mathbf{a}_i \sigma_i^{(z)} - \sum_{i=1}^{n} \mathbf{b}_i \sigma_{m+i}^{(z)} - \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{W}_{ij} \sigma_i^{(z)} \sigma_{m+j}^{(z)} - \sum_{i=1}^{n} \Gamma_i \sigma_{m+i}^{(z)}.$$

The model satisfies the Sufficient Conditions for classical data since it only has the transverse terms on the hidden units. For a given visible sample x, the condition density operator for the hidden layer(well-defined due to the model being a CQ-LVM) is given by

$$\mathrm{H}_{\mathrm{L}}(\mathrm{x}, \theta) = -\sum_{j=1}^{n} \mathbf{b}_j^{\mathrm{eff}} \sigma_j^{(z)} - \sum_{j=1}^{n} \Gamma_j \sigma_j^{(x)},$$

$$\rho_{\mathrm{L}}(\mathrm{x}|\theta) = \frac{1}{\mathcal{Z}_{\mathrm{L}}(\mathrm{x}, \theta)} \exp(\mathrm{H}_{\mathrm{L}}(\mathrm{x}, \theta)) \quad \text{where} \quad \mathcal{Z}_{\mathrm{L}}(\mathrm{x}, \theta) = \operatorname{Tr} \exp(\mathrm{H}^{\mathrm{L}}(\mathrm{x}, \theta)).$$

where $\mathbf{b}_j^{\mathrm{eff}} = \mathbf{b}_j + \sum_{i=1}^{m} \mathbf{W}_{ij} \mathrm{x}$. Using the definition of the Pauli Operators in Equation (2), we can rewrite this Hamiltonian as

$$\mathrm{H}_{\mathrm{L}}(j|\mathrm{x}, \theta) = -\mathbf{b}_j^{\mathrm{eff}} \sigma^{(z)} - \Gamma_j \sigma^{(x)},$$

$$\mathrm{H}_{\mathrm{L}}(\mathrm{x}, \theta) = \sum_{j=1}^{n} \mathrm{I}_{2^{j-1}} \otimes \mathrm{H}_{\mathrm{L}}(j|\mathrm{x}, \theta) \otimes \mathrm{I}_{2^{n-i}}.$$

Using Equation (9) recursively, the density operator $\rho_{\mathrm{L}}(\mathrm{x}_i|\theta)$ is given by

$$\rho_{\mathrm{L}}(\mathrm{x}|\theta) = \bigotimes_{i=1}^{n} \frac{1}{\mathcal{Z}_{\mathrm{L}}(j|\mathrm{x}, \theta)} \exp(\mathrm{H}_{\mathrm{L}}(j|\mathrm{x}, \theta)) \quad \text{where} \quad \mathcal{Z}_{\mathrm{L}}(j|\mathrm{x}, \theta) = \operatorname{Tr} \exp(\mathrm{H}_{\mathrm{L}}(j|\mathrm{x}, \theta)).$$

This gives us the desired sampling expression in Equation (8).

Appendix B. Compute Platform

In this section, we provide additional details related to our experiments that we do not provide in the main body of the paper.

Implementation Details We implement our proposed methods in PyTorch 2.5.1.

Hardware Tables 1 and 2 details the hardware we use to conduct our experiments. Values in (*) indicated reported values obtained from `https://www.amd.com/en/products/accelerators/instinct/mi200/mi210.html`. Machine 1 runs Ubuntu 22.04.3 LTS with kernel 6.8.0-40-generic with the hardware in Table 1. Machine 2 runs Ubuntu 22.04.1 LTS with kernel 5.15.0-127-generic with the hardware in Table 2. Our software stack comprises of 3.12.8, PyTorch 2.5.1, torchvision version 0.20.1.

Table 1: Machine 1: Specifications of GPU hardware used for computation

CPU Model Name	AMD EPYC 9654 96-Core Processor
CPU(s)	192
Thread(s) per core	1
Core(s) per socket	96
Socket(s)	2
NUMA node(s)	2
CPU MHz(Max)	3707.8120
L1d & L1i cache	6 MiB
L2 cache	192 MiB
L3 cache	768 MiB
RAM	1.48 TiB (DDR5, 4800 MT/s)
GPU Model name	Instinct MI210
GPU(s)	4
GPU Architecture	AMD Aldebaran
Dedicated Memory Size(per GPU)	64 GB
ROCm Version	6.0.2
Peak FP32 Performance*	22.6 TFLOPs
Peak FP64 Performance*	22.6 TFLOPs
Memory Clock*	1.6 GHz
Peak Memory Bandwidth*	1.6 TB/s

Table 2: Machine 2: Specifications of GPU hardware used for computation

CPU Model Name	Intel(R) Xeon(R) Silver 4216 CPU
CPU(s)	64
Thread(s) per core	2
Core(s) per socket	16
Socket(s)	2
NUMA node(s)	2
CPU MHz(Max)	3200.0000
L1d & L1i cache	1 MiB
L2 cache	32 MiB
L3 cache	44 MiB
RAM	62 GiB
GPU Model name	NVIDIA GeForce RTX 2080 Ti
GPU(s)	8
Dedicated Memory Size(per GPU)	11 GB
CUDA Version	12.9

References

D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. A learning algorithm for boltzmann machines. *Cognitive Science*, 9(1):147–169, 1985.

S.-i. Amari. Information geometry of the em and em algorithms for neural networks. *Neural Networks*, 8(9):1379–1408, 1995.

M. H. Amin, E. Andriyash, J. Rolfe, B. Kulchytskyy, and R. Melko. Quantum boltzmann machine. *Phys. Rev. X*, 8:021050, 2018.

E. R. Anschuetz and Y. Cao. Realizing quantum boltzmann machines through eigenstate thermalization. *arXiv preprint arXiv:1903.01359*, 2019.

A. Basilevsky. *Statistical Factor Analysis and Related Methods: Theory and Applications*. Wiley, 2009.

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

A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

M. Benedetti, J. Realpe-Gómez, R. Biswas, and A. Perdomo-Ortiz. Quantum-assisted learning of hardware-embedded probabilistic graphical models. *Phys. Rev. X*, 7:041052, 2017.

M. Berta, M. Lemm, and M. M. Wilde. Monotonicity of quantum relative entropy and recoverability. *Quantum Info. Comput.*, 15(15–16):1333–1354, 2015.

R. Bhatia. *Matrix Analysis*. Springer, 1997.

J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd. Quantum machine learning. *Nature*, 549(7671):195–202, 9 2017.

E. Bilgin and S. Boixo. Preparing thermal states of quantum systems by dimension reduction. *Phys. Rev. Lett.*, 105:170405, 2010.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

O. Bratteli and D. W. Robinson. *Operator Algebras and Quantum Statistical Mechanics I*. Theoretical and Mathematical Physics. Springer, Berlin, Heidelberg, 2 edition, 1987.

O. Bratteli and D. W. Robinson. *Operator Algebras and Quantum Statistical Mechanics II*. Theoretical and Mathematical Physics. Springer, Berlin, Heidelberg, 2 edition, 1997.

E. A. Carlen and A. Vershynina. Recovery map stability for the data processing inequality. *Journal of Physics A: Mathematical and Theoretical*, 53(3):035204, 2020.

M. A. Carreira-Perpiñán and G. Hinton. On contrastive divergence learning. In *International Workshop on Artificial Intelligence and Statistics*, pages 33–40, 06–08 Jan 2005.

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

N. J. Cerf and C. Adami. Quantum extension of conditional probability. *Phys. Rev. A*, 60: 893–897, Aug 1999.

C.-F. Chen, M. J. Kastoryano, F. G. S. L. Brandão, and A. Gilyén. Quantum thermal state preparation. *arXiv preprint arXiv:2303.18224*, 2023.

J. B. Conway. *A Course in Functional Analysis.* Springer, New York, 2 edition, 1990.

L. Coopmans and M. Benedetti. On the sample complexity of quantum boltzmann machine learning. *Communications Physics*, 7(1):274, 2024.

T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley-Interscience, USA, 2006.

S. S. Cree and J. Sorce. Geometric conditions for saturating the data processing inequality. *Journal of Physics A: Mathematical and Theoretical*, 55(13):135301, 2022.

J. de Leeuw. Block-relaxation algorithms in statistics. In *Information Systems and Data Analysis*, pages 308–324, Berlin, Heidelberg, 1994. Springer.

M. Demidik, C. Tüysüz, N. Piatkowski, M. Grossi, and K. Jansen. Expressive equivalence of classical and quantum restricted boltzmann machines. *arXiv preprint arXiv:2502.17562*, 2025.

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.

L. 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.

Z. Ding, B. Li, and L. Lin. Efficient quantum gibbs samplers with kubo–martin–schwinger detailed balance condition. *Communications in Mathematical Physics*, 406(3):67, 2025.

S. S. Dragomir. Jensen's type trace inequalities for convex functions of selfadjoint operators in hilbert spaces. *Facta Universitatis, Series Mathematics and Informatics*, 31(5):981–998, 2016.

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

A. M. Gleason. Measures on the closed subspaces of a hilbert space. *Journal of Mathematics and Mechanics*, 6(4):885–893, 1957.

Y. Gujju, A. Matsuo, and R. 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, 2024.

P. Hayden, R. Jozsa, D. Petz, and A. Winter. Structure of states which satisfy strong sub-additivity of quantum entropy with equality. *Communications in Mathematical Physics*, 246(2):359–374, 2004.

H. Hino, S. Akaho, and N. Murata. Geometry of em and related iterative algorithms. *Information Geometry*, 7(1):39–77, 2024.

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

G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851, 2020.

A. Holevo. The capacity of the quantum channel with general signal states. *IEEE Transactions on Information Theory*, 44(1):269–273, 1998.

Z. Hradil. Quantum-state estimation. *Phys. Rev. A*, 55:R1561–R1564, 1997.

O. Huijgen, L. Coopmans, P. Najafi, M. Benedetti, and H. J. Kappen. Training quantum boltzmann machines with the β-variational quantum eigensolver. *arXiv preprint arXiv:2304.08631*, 2024.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

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

D. R. H. Kenneth Lange and I. Yang. Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, 9(1):1–20, 2000.

I. Kerenidis, J. Landman, A. Luongo, and A. Prakash. q-means: A quantum algorithm for unsupervised machine learning. In *Advances in Neural Information Processing Systems*, pages 4136–4146, 2019.

I. Kerenidis, A. Luongo, and A. Prakash. Quantum expectation-maximization for Gaussian mixture models. In *International Conference on Machine Learning*, pages 5187–5197, 7 2020.

M. Kieferová and N. Wiebe. Tomography and generative training with quantum boltzmann machines. *Phys. Rev. A*, 96:062327, 12 2017.

Y. Kim, A. Eddins, S. Anand, K. X. Wei, E. van den Berg, S. Rosenblatt, H. Nayfeh, Y. Wu, M. Zaletel, K. Temme, and A. Kandala. Evidence for the utility of quantum computing before fault tolerance. *Nature*, 618(7965):500–505, 2023.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

H. Larochelle and Y. Bengio. Classification using discriminative restricted boltzmann machines. In *International Conference on Machine Learning*, page 536–543, 2008.

A. Lesniewski and M. B. Ruskai. Monotone riemannian metrics and relative entropy on noncommutative probability spaces. *Journal of Mathematical Physics*, 40(11):5702–5724, 11 1999.

R. Liao, S. Kornblith, M. Ren, D. J. Fleet, and G. Hinton. Gaussian-bernoulli RBMs without tears. *arXiv preprint arXiv:2210.10318*, 2022.

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

Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015.

G. Mahan. *Many-Particle Physics*. Physics of Solids and Liquids. Springer US, 2013.

H. Miyahara, K. Aihara, and W. Lechner. Quantum expectation-maximization algorithm. *Phys. Rev. A*, 101(1):012326, 2020.

J. Nelson, M. Vuffray, A. Y. Lokhov, T. Albash, and C. Coffrin. High-quality thermal gibbs sampling with quantum annealing hardware. *Phys. Rev. Appl.*, 17:044046, 2022.

M. A. Nielsen and I. L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, 2010.

D. Patel and M. M. Wilde. Natural gradient and parameter estimation for quantum boltzmann machines. *arXiv preprint arXiv:2410.24058*, 2024.

D. Patel, D. Koch, S. Patel, and M. M. Wilde. Quantum boltzmann machine learning of ground-state energies. *arXiv preprint arXiv:2410.12935*, 2024.

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

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

D. Petz. Monotonicity of quantum relative entropy revisited. *Reviews in Mathematical Physics*, 15(01):79–91, 2003.

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

J. Rehácek, Z. Hradil, E. Knill, and A. I. Lvovsky. Diluted maximum-likelihood algorithm for quantum tomography. *Phys. Rev. A*, 75:042108, 2007.

M. B. Ruskai. Inequalities for quantum entropy: A review with conditions for equality. *Journal of Mathematical Physics*, 43(9):4358–4375, 2002.

R. Salakhutdinov and G. Hinton. Deep boltzmann machines. In *International Conference on Artificial Intelligence and Statistics*, pages 448–455, 2009.

R. Salakhutdinov and G. E. Hinton. An efficient learning procedure for deep boltzmann machines. *Neural Computation*, 24(8):1967–2006, 2012.

R. Salakhutdinov, A. Mnih, and G. Hinton. Restricted boltzmann machines for collaborative filtering. In *International Conference on Machine Learning*, page 791–798, 2007.

M. Schuld, I. Sinayskiy, and F. Petruccione. An introduction to quantum machine learning. *Contemporary Physics*, 56(2):172–185, 2015.

M. Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, 2020. Version 0.3.0.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

R. Shibukawa, R. Tamura, and K. Tsuda. Boltzmann sampling with quantum annealers via fast stein correction. *Phys. Rev. Res.*, 6:043050, 2024.

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, 1986.

S. Taniguchi, M. Suzuki, Y. Iwasawa, and Y. Matsuo. End-to-end training of deep boltzmann machines by unbiased contrastive divergence with local mode initialization. In *International Conference on Machine Learning*, pages 33804–33815, 2023.

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

J. von Neumann. *Mathematical Foundations of Quantum Mechanics*. Princeton University Press, Princeton, NJ, 1955. English translation of *Mathematische Grundlagen der Quantenmechanik*.

M. K. Warmuth and D. Kuzmin. Bayesian generalized probability calculus for density matrices. *Mach. Learn.*, 78(1–2):63–101, 2010.

M. K. K. Warmuth. A bayes rule for density matrices. In *Advances in Neural Information Processing Systems*, 2005.

J. Watrous. *The Theory of Quantum Information*. Cambridge University Press, Cambridge, 2018.

M. Welling, M. Rosen-zvi, and G. E. Hinton. Exponential family harmoniums with an application to information retrieval. In *Advances in Neural Information Processing Systems*, 2004.

N. Wiebe and L. Wossnig. Generative training of quantum boltzmann machines with hidden units. *arXiv preprint arXiv:1905.09902*, 2019.

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

M. M. Wilde. *Quantum Information Theory*. Cambridge University Press, 2016.

H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

C. Zoufal, A. Lucchi, and S. Woerner. Variational quantum boltzmann machines. *Quantum Machine Intelligence*, 3(1):7, 2021.