Learning Density Operator Latent Variable Models via Quantum Information Projection

Adit Vishnu
Indian Institute of Science, Bengaluru
aditvishnu@iisc.ac.in

Abhay Shastry*
Chanakya University, Bengaluru
abhayshankar.c@chanakyauniversity.edu.in

Dhruva Kashyap*
HP AI Lab, Bengaluru
dhruva.kashyap1@hp.com

Chiranjib Bhattacharyya
Indian Institute of Science, Bengaluru
chiru@iisc.ac.in

Abstract

Latent variable models based on density operators, the mathematical foundation of quantum mechanics, remain far less capable than their probabilistic counterparts. This is primarily due to the lack of an Expectation-Maximization (EM) framework for learning density operator models, since the absence of conditional probability for density operators prevents a direct extension of the classical derivation. This paper addresses this challenge by proposing Density Operator Expectation Maximization (DO-EM), using tenets from quantum information theory. We first derive an operator-theoretic analog of the evidence lower bound from the data processing inequality and study its maximization as a projection onto the data manifold. We solve this information projection problem while also improving existing information recovery results. The DO-EM algorithm guarantees monotone likelihood ascent for a rich class of models and is simpler than direct likelihood gradient ascent. The Expectation step recovers the Petz Recovery Map when the model class can be aligned with the target operator. We specialize these models to Classical-Quantum LVMs which scale to standard image datasets via a per-datapoint decomposition of the variational bound and outperform their classical counterparts under matched computational resources.

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 [1]. Quantum machine learning promises to harness quantum mechanics to enhance learning, either by achieving computational speedups on quantum devices [2–5] or by modeling richer data structures through quantum representations [6–8]. 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 [9, 10]. While significant progress has been made in the use of density operators in kernel methods [6, 11, 12] and optimization [13, 14], there is relatively less progress in unsupervised learning.

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 [15]. 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

*work done while at the Indian Institute of Science, Bengaluru

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

reduction. Classical examples such as Gaussian Mixture Models, Factor Analysis, and Hidden Markov Models [16, 17] 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 [18], Generative Adversarial Networks [19], and Diffusion-based models [20]. The EM algorithm [21, 22] has been instrumental in deriving procedures for learning latent variable models. These algorithms are often preferred over algorithms which directly maximizes likelihood.

The study of Density Operator-based Latent Variable Models (DO-LVM) remains in its early stages, with foundational questions around expressivity, inference, and learning still largely unexplored [8, 23, 24]. Leveraging the modeling power of DO-LVMs on real-world data remains a significant challenge. Existing approaches rarely scale beyond 12 visible units—limited by restricted access to quantum hardware, the exponential cost of simulating quantum systems, and the memory bottlenecks associated with representing and optimizing DO-LVMs on classical devices. As a result, it is currently infeasible to empirically assess whether DO-LVMs offer any practical advantage on real-world datasets in terms of modeling power. EM-based algorithms offer an alternative to direct likelihood optimization in latent variable models. Extending this approach to density operators is difficult because noncommutativity obstructs the classical conditioning argument. We address this through the data processing inequality and constrained relative-entropy minimization.

In this paper we bridge these research gaps by making the following contributions.

A structural condition for analyzing DO-LVMs. Existing density operator LVMs sit at two extremes: general Gibbs state models [8] that are theoretically intractable for structural analysis, and quantum Markov random fields [25] that admit clean analysis but forbid non-commutativity throughout the model, ruling out latent-side non-commutativity central to models of interest. We introduce V-locality, a structural condition that constrains only visible-side commutativity of the Hamiltonian. We show that V-locality for every bipartition recovers the quantum Markov random field condition, establishing V-locality as a strictly weaker condition that suffices for all subsequent results in this paper.

Quantum Information Projection Maximum-likelihood training of probabilistic latent variable models rests on the evidence lower bound (ELBO), but the absence of conditional probability for density operators precludes a direct quantum analog. We derive a quantum ELBO from the data processing inequality for the partial trace, sidestepping conditioning entirely. Tightening this bound is a variational problem we call the quantum information projection; for any Gibbs state, we give the optimum in closed form via a finite-dimensional convex program, recovering the Petz recovery map as a special case under the structural conditions of Section 3. As a consequence, the data processing defect at the optimum admits a closed-form expression, an exact Pythagorean decomposition of the data-processing defect.

An EM algorithm for Density Operators Maximum-likelihood training of density operator LVMs is a longstanding open problem as current methods either cannot update the parameters controlling quantum effects [8] or rely on heuristics without theoretical guarantees [26–28]. The absence of quantum conditional probability blocks the classical EM derivation and the partial trace inside the log-likelihood blocks closed-form gradients. We introduce DO-EM, the density operator analog of the classical Expectation-Maximization algorithm, which addresses both obstructions simultaneously: the E-step is the QIP of Section 4, sidestepping the lack of quantum conditional probability; the M-step admits a closed-form gradient via the Lie–Trotter–Suzuki trick, sidestepping the partial trace. For sufficiently rich model classes, DO-EM achieves strict monotone log-likelihood.

DO-LVMs for Classical Data Prior DO-LVM training methods do not scale to standard image datasets and fail to scale beyond toy data sets of $\sim$12-bit binary strings. We introduce Classical-Quantum LVMs (CQ-LVMs), a new class of DO-LVMs satisfying the V-local assumption, for which DO-EM specializes to a per-datapoint iteration with no dependence on the size of the visible Hilbert space. Concrete instantiations — Quantum Restricted Boltzmann Machines, Quantum Gaussian–Bernoulli RBMs, and Quantum interleaved Deep Boltzmann Machines — are the first DO-LVMs trained on standard image datasets. Under identical hyperparameters and matched compute budgets, our quantum models achieve 20–65% reductions in Fréchet Inception Distance over their

classical counterparts on MNIST, Fashion-MNIST, and CelebA-32 — providing the first empirical evidence of a modeling advantage for density operator formulations on real-world data.

2 Preliminaries

Notation This manuscript works in finite-dimensional Hilbert spaces unless stated otherwise, for the sake of clarity. All results admit natural extensions to the infinite-dimensional setting. 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 $\mathcal{T}(\mathcal{H})$. Completely positive trace preserving (CPTP) maps that take operators to operators are denoted $\mathcal{N}$ and its adjoint is denoted by $\mathcal{N}^\dagger$. 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 $[X, Y] = XY - YX$.

Latent variable models and EM algorithm Latent Variable Models (LVMs) [15] specify the probability distribution of random variables $V{=}[V_1, \ldots, V_{d_\mathrm{V}}]$ through a joint probability model $P(V{=}\,\mathrm{v} \mid \theta) = \sum_h P(V{=}\,\mathrm{v}, H{=}\,\mathrm{h} \mid \theta)$ where $H = [H_1, \ldots, H_{d_\mathrm{L}}]$ are unobserved random variables. Learning an LVM from data, a problem of great interest in Unsupervised Learning [16], refers to estimating the model parameters θ from a dataset $\mathcal{D} = \{\mathrm{v}^{(1)}, \ldots, \mathrm{v}^{(N)}\}$ consisting of i.i.d instances drawn from the LVM. Maximum likelihood-based methods aim to maximize $\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^N \ell_i(\theta)$ where $\ell_i(\theta) = \log P(V{=}\,\mathrm{v}^{(i)} \mid \theta)$. The maximization problem is not only intractable in most cases but even gradient-based algorithms, which can only discover local optima, are difficult to implement because of unwieldy computations in $\ell_i(\theta)$. The EM algorithm [21, 22] is an alternative iterative algorithm with the scheme

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

Markov Random Fields A Markov Random Field (MRF) specifies a joint distribution over random variables $\{X_v\}_{v\in V}$ indexed by an undirected graph $\mathcal{G} = (V, E)$, encoding conditional independence via the graph's separation structure [29]. By the Hammersley–Clifford theorem [30], every strictly positive distribution satisfying the Markov property w.r.t. $\mathcal{G}$ factorizes over its (maximal) cliques $C(\mathcal{G})$:

$$P(\{x_v\}_{v\in V}) = \frac{1}{Z}\prod_{c\in C(\mathcal{G})} \phi_c(\{x_v\}_{v\in c}), \tag{1}$$

where $Z = \sum_{\{x_u\}_{u\in V}} \prod_c \phi_c(\{x_u\}_{u\in c})$ is the normalization constant, called the *partition function*, and ϕ_c are nonnegative *potential functions*. Partitioning $V = O \cup H$ into observed and hidden vertices yields a class of LVMs in which $P(x_O) = \sum_{x_H} P(x_O, x_H)$ can capture dependencies through unobserved structure. This formulation underlies Boltzmann machines (BMs) [31], Restricted Boltzmann Machines (RBMs) [32, 33] and Deep Boltzmann machines (DBMs) [34]. DBMs are usually trained using variants of the Contrastive Divergence algorithm [33, 35, 36]. A detailed discussion on Boltzmann Machines and the Contrastive Divergence algorithm is provided in Appendix C.

2.1 Density operators

A density operator on a Hilbert space $\mathcal{H}$ is a Hermitian, positive semi-definite operator with unit trace [37, 38]. The set of Density operators will be denoted by $\mathcal{P}(\mathcal{H})$, and can be regarded as generalizations of probability distributions. 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 or *sample* λ_i with probability $\Pr(\lambda_i) = \mathrm{Tr}(\rho\Lambda_i)$. A joint density operator $\rho \in \mathcal{P}(\mathcal{H}_A \otimes \mathcal{H}_B)$ can be *marginalized* to $\rho_A \in \mathcal{P}(\mathcal{H}_A)$ by the partial trace operation $\rho_A = \mathrm{Tr}_B(\rho) = \sum_{i=1}^{d_B}(I_A \otimes \mathbf{x}_i^\dagger)\rho(I_A \otimes \mathbf{x}_i)$ where $\{\mathbf{x}_i\}_{i=1}^{d_B}$ is an orthonormal basis of $\mathcal{H}_B$.

Classical information theoretic notions such entropy, relative entropy, and the data processing inequality can be generalized to quantum information theory.

Definition 2.1 (Umegaki [39] Relative Entropy)**.** *Let ω and ρ be density operators in $\mathcal{P}(\mathcal{H}_A \otimes \mathcal{H}_B)$ with $\ker(\rho) \subseteq \ker(\omega)$. Their relative entropy is given by $\mathrm{D_U}(\omega, \rho) = \mathrm{Tr}(\omega \log \omega) - \mathrm{Tr}(\omega \log \rho)$.*

Lindblad [40] showed that the relative entropy does not increase under the action of the parital trace.

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

Petz [41, 42] showed that this inequality is saturated if and only if the Petz Recovery Map reverses the partial trace operation.

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

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

The recoverability literature in quantum information theory addresses how well the partial trace can be inverted. Subsequent work has refined this in several directions — universal recovery maps depending only on the marginal [43], generalization from the partial trace to arbitrary CPTP maps via the pinched Petz recovery map [44], and universal recovery for general channels [45]. These results are approximate; they bound the recovery error in terms of information-theoretic quantities but do not characterize it exactly.

Quantum Markov Random Fields The Hamiltonian is a Hermitian operator $\mathrm{H} \in \mathfrak{L}(\mathcal{H})$ representing the total energy and generalizes the notion of an energy function in classical energy-based models. The model is defined using Gibbs state density matrix analogous to the Boltzmann distribution: $\rho(\theta) = \frac{\exp(\mathrm{H}(\theta))}{Z(\theta)}$ with $Z(\theta)$=$\mathrm{Tr}\exp(\mathrm{H}(\theta))$ and $\mathrm{H}(\theta)$=$\sum_r \theta_r \mathrm{H}_r$, where $\mathrm{H}_r \in \mathcal{T}(\mathcal{H})$ are Hermitian operators and $\theta_r \in \mathbb{R}$ are model parameters. However, general Hamiltonian-based models are quite difficult to analyze for their modeling and algorithmic properties without further assumptions. Probabilistic energy-based models are usually analyzed using the properties of Markov Random Fields. The condition for Hamiltonian-based density operator to be a Quantum Markov Random Field (QMRF) analogous to the classical MRF was identified by Brown and Poulin [25].

Definition 2.4 (Shield Commuting Hamiltonian)**.** *Given a graph $G = (\Lambda, E)$ and a Hamiltonian H acting on the particles located in Λ, we say that H is shield commuting when it can be written as $H = H_{AB} + H_{BC}$ with $[H_{AB}, H_{BC}] = 0$ whenever B shields A from C.*

However, this condition is extremely restrictive as several models of interest such as the Quantum Boltzmann Machine does not satisfy this condition.

The Quantum Boltzmann Machine is a Hamiltonian-based model inspired by the transverse field Ising model [8]. In this paper, $QBM_{m,n}$ denotes a model with m visible and n hidden units with

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

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

Need for an EM algorithm for Density Operators As probabilistic LVMs are a special case of DO-LVMs, the training challenges they face persist in DO-LVMs, which also introduce new operator-theoretic difficulties. Maximizing the log-likelihood of a DO-LVM involves operators that do not commute [23]. The direct computation of gradient in these models is significantly complicated by the partial trace [26]. Due to the difficulty of working with hidden units, recent work on QBMs have focused on models without hidden units [24, 46–48]. Demidik et al. [49] studied a Restricted QBM with 12 visible units and 90 hidden units, the largest model studied in literature so far. Refer

Appendix B for a detailed survey on QBM literature. Hence, training a QBM, the most popular DO-LVM in literature, on real-world data *remains an open challenge*.

Intractability of the gradient of the log-likelihood in probabilistic LVMs is addressed by the EM algorithm. Classical derivations of the EM algorithm fail with density operators since there is no well-defined way to construct conditional density operators [38]. The Conditional Amplitude Operator defined as

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

by Cerf and Adami [50] is a candidate for the E-step. This is insufficient since the CAO does not provide a valid density operators [50] unlike classical conditional probabilities.

Related Work The first EM-like algorithm for density operators was proposed in Warmuth and Kuzmin [51]. They conjectured that in cases where the CAO is trace deficient, there exists a bijection between ρ and $\rho_{B|A}$. They propose an *EM-like algorithm* linking ρ, ρ_A, and $\rho_{B|A}$. This proposal is not an EM algorithm for density operators in the usual sense but an attempt at an algorithmic Bayes rule. However, the map between ρ and $\rho_{B|A}$ is not a bijection as conjectured. The expression for CAO is non-injective when ρ_A is not fixed and one can achieve the same CAO by perturbing $\log \rho$ and $\log \rho_A$ by $M \otimes \mathrm{I}_B$ and M respectively.

Concurrent work by Kimura et al. [52] derive an information-geometric style algorithm for a class of Semi-Quantum Restricted Boltzmann Machines that were first studied in Demidik et al. [49]. These models are special cases of CQ-LVMs introduced in Definition 6.2. The algorithm introduced by Kimura et al. [52] is a special case of Algorithm 2 applied to their models.

3 Density operator latent variable models

In this section, we undertake a formal study of Density Operator Latent Variable Models and introduce a model class weaker than Quantum Markov Random Fields. We assume a data set is obtained from N arbitrary projective measurements of an underlying density operator ψ_{V} defined on the Hilbert space $\mathcal{H}_{\mathrm{V}}$. As discussed in Section 2.1, 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 $\mathcal{H}_{\mathrm{V}}$. The *empirical data density operator* $\eta_{\mathrm{V}} \in \mathcal{P}(\mathcal{H}_{\mathrm{V}})$ for a data set $\mathcal{D} = \{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(N)}\}$ is defined as $\eta_{\mathrm{V}} = \frac{1}{N}\sum_{i=1}^{N} \Lambda(\mathbf{v}^{(i)})$. We now formally define the model.

Definition 3.1 (Density Operator Latent Variable Model)**.** *A DO-LVM specifies the density operator $\rho_{\mathrm{V}} \in \mathcal{P}(\mathcal{H}_{\mathrm{V}})$ on observables in $\mathcal{H}_{\mathrm{V}}$ through a joint density operator $\rho \in \mathcal{P}(\mathcal{H}_{\mathrm{V}} \otimes \mathcal{H}_{\mathrm{L}})$ as $\rho_{\mathrm{V}} = \mathrm{Tr}_{\mathrm{L}}(\rho(\theta))$ where the space $\mathcal{H}_{\mathrm{L}}$ is not observed.*

These models are often trained using two different objective functions. The first objective, termed the *projective log-likelihood* (PL), is based on projective measurements in the visible Hilbert space [8]. 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

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

The second objective, termed the *quantum log-likelihood* (QL), is based on the Umegaki relative entropy [23]. A data set with inbuilt quantum correlations or a data set with artificially induced quantum correlations [24] is used to build an empirical data operator η_{V} in $\mathcal{P}(\mathcal{H}_{\mathrm{V}})$. The maximization of the quantum log-likelihood

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

is equivalent to the minimization of the Umegaki relative entropy between η_{V} and $\rho_{\mathrm{V}}(\theta)$. We first prove an inequality between the to log-likelihoods.

Proposition 3.2. *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}. Then, $\mathcal{L}_{\mathrm{P}}(\mathcal{D}, \theta) \geq \mathcal{L}_{\mathrm{U}}(\eta_{\mathrm{V}}, \theta)$.*

This result follows from the application of a Jensen-type operator trace inequality proved by Dragomir [53] to the projection operators appearing in PL. Further analysis and optimization of the log-likelihoods are difficult as existing formalisms such as QMRFs are too restrictive on the latent

subspace to include models of interest. We introduce V-locality, a notion weaker than QMRFs that does not constrain the latent space of DO-LVMs.

Definition 3.3 (V-locality)**.** *Let $G = (\Lambda, E)$ be a finite simple graph and let $\Lambda = V \sqcup L$ be a vertex bipartition. Given a Hamiltonian $\mathrm{H} = \sum_{C \in \mathcal{C}(G)} \mathrm{H}_C$ with each H_C Hermitian and supported on the clique C, write $\mathrm{H} = \mathrm{H}_V + \mathrm{H}_L + \mathrm{H}_{VL}$, where $\mathrm{H}_V := \sum_{C \subseteq V} \mathrm{H}_C$ collects clique terms supported in V, $\mathrm{H}_L := \sum_{C \subseteq L} \mathrm{H}_C$ collects clique terms supported in L, and $\mathrm{H}_{VL} := \mathrm{H} - \mathrm{H}_{\mathrm{V}} - \mathrm{H}_{\mathrm{L}}$ collects clique terms intersecting both. We say that H is V-*local *if* $[\mathrm{H}_V, \mathrm{H}_{VL}] = 0$.*

V-locality keeps desirable properties of QMRF Hamiltonians on the visible subspace while not constraining the latent space. We now showcase the connection between V-locality and QMRFs.

Theorem 3.4. *V-locality for every bipartition implies shield-commutation. Conversely, if G is triangle-free and H is shield-commuting, then there exists a clique decomposition of H that is V-local for every bipartition.*

Theorem 3.4 compares the strongest form of V-locality, on every bipartition, to shield commutativity, showing it is no stronger. The downstream results in this paper require only V-locality. We now proceed to derive learning algorithms for DO-LVMs using QL.

4 Quantum ELBO and Information Projection

In this section, we derive a quantum evidence lower bound for QL and maximize it using quantum information projection. The classical ELBO is usually derived for each datapoint using conditional probability and Jensen's inequality. This approach fails for density operators due to the absence well-defined quantum conditional probability [38]. Since the classical ELBO inequality can be seen as a consequence of the data processing inequality for marginalization, we derive a quantum ELBO using Theorem 2.2.

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

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

This inequality is a specialization of Theorem 2.2 to the partial trace. The classical EM algorithm is a consequence of the ELBO being a minorant of the log-likelihood [54, 55] when the variational distribution is given by the conditional probability with respect to the model. In the absence of a well defined notion of conditional probability [38], we look for feasible extensions that saturate QELBO for the partial trace. However, it is well known that Theorem 2.2 is often not saturated for the partial trace [56–60].

Inspired by an information geometric interpretation of the EM algorithm [61, 62], we study an instance of a quantum information projection problem to maximize the QELBO.

Definition 4.2 (Quantum Information Projection)**.** *Given density operators η_{V} in $\mathcal{P}(\mathcal{H}_{\mathrm{V}})$ and ρ faithful in $\mathcal{P}(\mathcal{H}_{\mathrm{V}} \otimes \mathcal{H}_{\mathrm{L}})$, find η^* in $\mathcal{J}(\eta_{\mathrm{V}}) = \{\eta \mid \eta \in \mathcal{P}(\mathcal{H}_{\mathrm{V}} \otimes \mathcal{H}_{\mathrm{L}}) \;\&\; \mathrm{Tr}_{\mathrm{L}}\eta = \eta_{\mathrm{V}}\}$ such that*

$$\eta^*(\rho, \eta_{\mathrm{V}}) = \operatorname*{argmin}_{\mathcal{J}(\eta_{\mathrm{V}})} \mathrm{D}_{\mathrm{U}}(\eta, \rho). \tag{QIP}$$

This constrained optimization problem generalizes the classical notion of information projection [63, 64] and can be seen as Umegaki relative entropy analog of the quantum fidelity-based results in Bai et al. [65]. While both the QIP and CAO reduce to Bayes rule in the classical case, the QIP is always a density operator and hence easier to use as an E-step in an EM algorithm for density operators.

The feasible set $\mathcal{J}(\eta_{\mathrm{V}})$ is non-empty as it contains $\eta_{\mathrm{V}} \otimes \mathrm{I}_{\mathrm{L}}/d_{\mathrm{L}}$, convex, and compact in $\mathcal{P}(\mathcal{H}_{\mathrm{V}} \otimes \mathcal{H}_{\mathrm{L}})$. The objective $\eta \mapsto \mathrm{D}_{\mathrm{U}}(\eta, \rho)$ is strictly convex and lower semicontinuous on this set, with the support condition $\ker(\rho) \subseteq \ker(\eta)$ holding trivially because ρ is faithful. Hence QIP has a unique solution. Let $\mathcal{S} := \mathrm{supp}(\eta_{\mathrm{V}}) \subseteq \mathcal{H}_{\mathrm{V}}$ denote the support of η_{V} and let $\mathrm{P}_{\mathcal{S}}$ denote the orthogonal projector onto $\mathcal{S}$. A standard observation, which we record now and use throughout, is that any feasible η for QIP has support contained in $\mathcal{S} \otimes \mathcal{H}_{\mathrm{L}}$. Hence the QIP problem effectively lives on $\mathcal{S} \otimes \mathcal{H}_{\mathrm{L}}$, and the unique optimum has support contained in this subspace.

Algorithm 1 Density Operator Expectation Maximization

1: **Input:** Data density operator η_{V} and initial model parameters $\theta^{(0)}$
2: **while** not converged **do**
3: **E Step:** $\mathcal{Q}(\theta;\theta^{(t)}) := \mathrm{QELBO}(\eta(\theta^{(t)}),\rho(\theta))$ where $\eta(\theta^{(t)}) = \underset{\eta:\mathrm{Tr}_{\mathrm{L}}\eta=\eta_{\mathrm{V}}}{\mathrm{argmin}}\ \mathrm{D_U}(\eta,\rho(\theta^{(t)}))$
4: **M Step:** $\theta^{(t+1)} = \underset{\theta}{\mathrm{argmax}}\ \mathcal{Q}(\theta;\theta^{(t)}) = \underset{\theta}{\mathrm{argmax}}\ \mathrm{Tr}(\eta(\theta^{(t)})\log\rho(\theta))$

Theorem 4.3. *Let $\eta_{\mathrm{V}} \in \mathcal{P}(\mathcal{H}_{\mathrm{V}})$ have support $\mathcal{S}$ with projector $\mathrm{P}_{\mathcal{S}}$, and let $\rho \in \mathcal{P}(\mathcal{H}_{\mathrm{V}}\otimes\mathcal{H}_{\mathrm{L}})$ be a Gibbs state with Hamiltonian H. Define $\mathrm{H}_{\mathcal{S}} := (P_{\mathcal{S}}\otimes\mathrm{I}_{\mathrm{L}})\,\mathrm{H}\,(P_{\mathcal{S}}\otimes\mathrm{I}_{\mathrm{L}})$. Let Λ^* be the unique minimizer of the strictly convex objective over traceless Hermitian operators in $\mathcal{T}(\mathcal{S})$ $F(\Lambda) := \log\mathrm{Tr}\exp\big(\mathrm{H}_{\mathcal{S}}+\Lambda\otimes\mathrm{I}_{\mathrm{L}}\big) - \mathrm{Tr}(\Lambda\,\eta_{\mathrm{V}})$. Then the solution to QIP is given by*

$$\eta^*(\rho,\eta_{\mathrm{V}}) = \frac{1}{\mathcal{Z}(\Lambda^*)}\exp\big(\mathrm{H}_{\mathcal{S}}+\Lambda^*\otimes\mathrm{I}_{\mathrm{L}}\big), \quad \mathcal{Z}(\Lambda^*) := \mathrm{Tr}\exp\big(\mathrm{H}_{\mathcal{S}}+\Lambda^*\otimes\mathrm{I}_{\mathrm{L}}\big).$$

Theorem 4.3 characterizes this projection via a Lagrange-multiplier formulation: the constraint $\mathrm{Tr}_{\mathrm{L}}\eta = \eta_{\mathrm{V}}$ is enforced by a multiplier $\Lambda^*\otimes\mathrm{I}_{\mathrm{L}}$, which adds to the model's log-density to produce η^*. The projection is the classical-probability I-projection extended to density operators. The QIP problem is closely connected to *recoverability* discussed in Section 2.1, the question of how well the partial trace can be inverted. We now use our solution to the QIP problem to provide an exact data-processing defect decomposition in terms of information-theoretic quantities.

Corollary 4.4. *Let η and ρ be density operators in $\mathcal{P}(\mathcal{H}_{\mathrm{V}}\otimes\mathcal{H}_{\mathrm{L}})$ with ρ faithful and let $\eta_{\mathrm{V}} = \mathrm{Tr}_{\mathrm{L}}\eta$. The data-processing defect is equal to*

$$\mathrm{D_U}(\eta,\rho) - \mathrm{D_U}(\eta_{\mathrm{V}},\rho_{\mathrm{V}}) = \mathrm{D_U}(\eta,\eta^*(\rho,\eta_{\mathrm{V}})) + \mathrm{D_U}(\eta^*(\rho,\eta_{\mathrm{V}}),\rho) - \mathrm{D_U}(\eta_{\mathrm{V}},\rho_{\mathrm{V}}).$$

This is due to an operator-theoretic analog of the classical I-projection Pythagorean identity for KL divergence over linear constraint families [66]. We now give a sufficient condition for when this defect is zero which also leads to closed form expressions for the QIP.

Corollary 4.5. *Let $\rho \in \mathcal{P}(\mathcal{H}_{\mathrm{V}}\otimes\mathcal{H}_{\mathrm{L}})$ be a Gibbs state with V-local Hamiltonian H, and let $\eta_{\mathrm{V}} \in \mathcal{P}(\mathcal{H}_{\mathrm{V}})$ have support $\mathcal{S}$. Suppose $[\rho, \rho_{\mathrm{V}}^{-1/2}\eta_{\mathrm{V}}\rho_{\mathrm{V}}^{-1/2}\otimes\mathrm{I}_{\mathrm{L}}] = 0$. Then the QIP optimum $\eta^*(\rho,\eta_{\mathrm{V}})$ from Theorem 4.3 is given by the Petz Recovery Map $\mathcal{R}_{\rho,\mathrm{Tr}_{\mathrm{L}}}(\eta_{\mathrm{V}})$.*

The Petz Recovery Map was recently shown to be the minimum change operator under the commutativity condition $[\eta_{\mathrm{V}},\rho_{\mathrm{V}}] = 0$ condition [65]. This corollary provides a new interpretation of the Petz Recovery Map as a QIP. Having solved the problem with maximizing the QELBO, we now design our EM algorithm for density operators.

5 The DO-EM framework

In this section, we present the **D**ensity **O**perator **E**xpectation **M**aximization (DO-EM) algorithm prove that it can achieve log-likelihood ascent at every iteration.

For fixed parameters $\theta^{(\mathrm{t})}$, the QELBO is maximized when η is the QIP of $\rho(\theta)$ onto the set of feasible extensions. This allows us to formulate the DO-EM algorithm in Algorithm 1.

The DO-EM algorithm can be considered a density operator analog of the classical EM algorithm as we recover the classical EM algorithm from DO-EM for discrete models if η_{V} and $\rho(\theta)$ are diagonal. The E step can be solved using the convex problem in the previous section. The **M Step** in DO-EM algorithm maximizes an expression without the partial trace. We first show that gradients of the M step are easier to compute than the gradient of QL.

Proposition 5.1. *For a model $\rho(\theta) = \exp(\mathrm{H}(\theta))/\mathcal{Z}(\theta)$ with $\mathrm{H}(\theta) = \sum_r \theta_r \mathrm{H}_r$ and an E-step output $\eta^{(t)}$, the gradient of $\mathcal{Q}(\theta;\theta^{(old)})$ in the M-step is $\frac{\partial}{\partial\theta_r}\mathcal{Q}(\theta;\theta^{(old)}) = \langle\mathrm{H}_r\rangle_{\eta^{(t)}} - \langle\mathrm{H}_r\rangle_{\rho(\theta)}$.*

The log-likelihood of such expressions may have closed-form expressions for the gradients, for example, using Duhamel's formula [24]. In the classical case, this is equivalent to the EM algorithm maximizing a sum of logarithms instead of a logarithm of sums.

However, the non-saturation of the QELBO makes it difficult to prove the properties of the algorithm. For the convergence analysis, write

$$Q(\theta' \mid \theta) := \mathrm{Tr}(\eta^*(\theta) \log \rho(\theta')),$$

so that $\mathcal{Q}(\theta'; \theta) = Q(\theta' \mid \theta) + \mathrm{S_{VN}}(\eta^*(\theta)) - \mathrm{S_{VN}}(\eta_\mathrm{V})$. The two auxiliary functions have the same M-step maximizers and first-argument gradients. We define a proxy to the likelihood as follows

$$\tilde{\mathcal{L}_\mathrm{U}}(\theta) \;:=\; \max_{\eta : \mathrm{Tr}_L \eta = \eta_\mathrm{V}} \mathrm{QELBO}(\eta, \theta) \;=\; Q(\theta \mid \theta) + \mathrm{S_{VN}}(\eta^*(\theta)) - \mathrm{S_{VN}}(\eta_\mathrm{V})$$

which is attained by the QIP. The proxy $\tilde{\mathcal{L}_\mathrm{U}}$ is a lower bound on the log-likelihood by construction. Inspired by the analysis in Wu [67], we showcase the algorithmic properties of the DO-EM algorithm.

Theorem 5.2. *Let $\rho(\theta) = e^{\mathrm{H}(\theta)}/Z(\theta)$ with $\mathrm{H}(\theta) = \sum_{r=1}^k \theta_r \mathrm{H}_r$ be a Hamiltonian-based DO-LVM on $\mathcal{H}_V \otimes \mathcal{H}_L$, with parameter space $\Theta = \mathbb{R}^k$. Given a target η_V, define the set of stationary points $\tilde{\mathrm{S}} := \{\theta \in \Theta : \nabla \tilde{\mathcal{L}_\mathrm{U}}(\eta_\mathrm{V}, \theta) = 0\}$. Let $M(\theta) := \mathrm{argmax}_{\theta' \in \Theta} \mathcal{Q}(\theta' \mid \theta) \subseteq \Theta$, where M is a point-to-set map; DO-EM produces iterates $\theta^{(t+1)} \in M(\theta^{(t)})$ for $t \geq 0$. Assume that $M(\theta^{(t)})$ is nonempty at every iteration, that each M-step attains its maximum, and that the iterates lie in a compact set $K \subset \Theta$. Then, the following results hold.*

(a) $\nabla_\theta \tilde{\mathcal{L}_\mathrm{U}}(\eta_\mathrm{V}, \theta) = \nabla_{\theta'} \mathcal{Q}(\theta' \mid \theta)|_{\theta'=\theta}$ *for all $\theta \in \Theta$ and $\theta \in M(\theta)$ if and only if $\theta \in \tilde{\mathrm{S}}$.*
(b) $\tilde{\mathcal{L}_\mathrm{U}}(\eta_\mathrm{V}, \theta^{(t)})$ *converges to a stationary value of $\tilde{\mathcal{L}_\mathrm{U}}(\eta_\mathrm{V}, \cdot)$.*
(c) *The iterates $\theta^{(t)}$ converges to a stationary set.*

The proxy log-likelihood differs from the log-likelihood exactly by the data processing defect in Corollary 4.4. This leads us immediately to the following corollary.

Corollary 5.3. *Suppose, in addition to the assumptions of Theorem 5.2, that $\mathrm{H}(\theta)$ is V-local and*

$$\Big[\rho(\theta), \Big(\rho_\mathrm{V}^{-1/2}(\theta) \eta_\mathrm{V} \rho_\mathrm{V}^{-1/2}(\theta)\Big) \otimes \mathrm{I_L}\Big] = 0$$

along the iterate trajectory. Then $\tilde{\mathcal{L}_\mathrm{U}}(\theta) = \mathcal{L}_\mathrm{U}(\eta_\mathrm{V}, \theta)$ along the trajectory and at all limit points. The likelihood values are nondecreasing and converge to a stationary value. Every limit point is stationary for $\mathcal{L}_\mathrm{U}(\eta_\mathrm{V}, \cdot)$, and

$$\mathrm{dist}\Big(\theta^{(t)}, \{\theta \in K : \nabla_\theta \mathcal{L}_\mathrm{U}(\eta_\mathrm{V}, \theta) = 0\}\Big) \longrightarrow 0.$$

Controlling a nonzero data-processing defect to guarantee true-likelihood ascent remains an open challenge. The sufficient target-alignment condition used here holds automatically for CQ-LVMs with a target diagonal in the data basis, as shown next.

Experiments Algorithm 1 is compared to maximizing the projected likelihood using gradient ascent on a QBM with 8 visible units and 2 hidden units [8] in Appendix D. DO-EM achieves the same performance as gradient ascent in fewer steps while taking only *0.2 seconds* per update, while direct gradient ascent on the projective log-likelihood requires approximately *40 seconds*.

6 DO-EM for classical data

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 6.1 (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 [68]. 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.

Algorithm 2 DO-EM for CQ-LVM

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

Definition 6.2 (CQ-LVM)**.** *A DO-LVM is a CQ-LVM with respect to a classical data set generated by the measurement* $\mathcal{X} = \sum_{i=1}^{d_{\mathrm{V}}} \mathrm{x}_i\, \Lambda(\mathbf{u}_i)$ *if it can be expressed as*

$$\rho(\theta) = \sum_{i=1}^{d_{\mathrm{V}}} \Pr(\mathrm{X}{=}\mathrm{x}_i|\theta)\Lambda(\mathbf{u}_i) \otimes \rho_{\mathrm{L}}(i|\theta), \tag{CQ-LVM}$$

where $\rho_{\mathrm{L}}(i|\theta)$ *are in* $\mathcal{P}(\mathcal{H}_{\mathrm{L}})$.

It is easy to check that CQ-LVMs are V-local and satisfy $[\rho, \rho_{\mathrm{V}}^{-1/2}\eta_{\mathrm{V}}\rho_{\mathrm{V}}^{-1/2} \otimes \mathrm{I}_{\mathrm{L}}] = 0$. The QELBO and the DO-EM algorithm can be specialized to each data point for CQ-LVMs.

Theorem 6.3. *The log-likelihood of CQ-LVMs for classical data can be expressed as* $\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \ell_i(\theta)$ *where* $\ell_i(\theta) = \log P(\mathbf{v}^{(i)} \mid \theta)$. *Moreover, this log-likelihood is lower bounded by*

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

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

This result is due to a structural decomposition of the model obtained from the V-locality condition of the Hamiltonian and the commutativity condition in Corollary 4.5. This allows us to specialize Algorithm 1 to Algorithm 2, enabling the implementation of DO-EM without being restricted by the dimension of η_{V}. However, models such as the QBM remain intractable for real-world data due to the normalization term, a problem that exists in classical Boltzmann machines as well.

6.1 Quantum Boltzmann Machine

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

Corollary 6.4. *A* $QBM_{m,n}$ *is a CQ-LVM if and only if quantum terms on the visible units are zero.*

This structural result is obtained by matching the terms in Hamiltonian to the definition in Definition 6.2. The class of semi-quantum models studied in Demidik et al. [49] are CQ-LVMs. Training such a QBM is intractable for real-world data since the free energy term, $-\log Z(\theta)$ is intractable even for classical Boltzmann machines. In Boltzmann machines, this issue is typically side-stepped by the Contrastive Divergence algorithm. 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

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

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.

Proposition 6.5. *In a QRBM, if the visible unit is fixed to* x*, the hidden units are conditionally independent,* $\rho_{\mathrm{L}}(\mathrm{x} \mid \theta) = \otimes_{j=1}^{n} \rho_{\mathrm{L}}(j \mid \mathrm{x}, \theta)$.

The Hamiltonian of the jth unit of the hidden layer L is given by $\mathrm{H}_{\mathrm{L}}(j|\mathrm{x},\theta) = -\mathbf{b}_j^{\mathrm{eff}}\sigma^{(z)} - \Gamma_j\sigma^{(x)}$ with $\mathbf{b}_j^{\mathrm{eff}} = \mathbf{b}_j + \sum_{i=1}^{m} \mathbf{W}_{ij}\mathrm{x}_i$. The Hamiltonian decomposes into n independent 2×2 matrices

and allows for the tractable sampling from the quantum layer using

$$\langle \sigma_j^{(z)} \rangle_{\mathbf{v}} = -\frac{\mathbf{b}_j^{\text{eff}}}{D_j} \tanh D_j \text{ and } \langle \sigma_j^{(x)} \rangle_{\mathbf{v}} = -\frac{\Gamma_j}{D_j} \tanh D_j \tag{2}$$

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 sample along the Pauli-Z operators. This closed-form expression for Gibbs sampling without matrices allows a heuristic CD algorithm to run on a QRBM with the same memory footprint as an RBM. This scheme can be generalized to the models considered in Demidik et al. [49]. However, as is the case with classical RBMs, models trained with CD do not inherit the monotonicity guarantees of the DO-EM algorithm.

Experiments We empirically evaluate our methods with experiments on Classical-Quantum LVMs. To achieve tractable training of QBMs, we introduce the two novel QBMs that can be trained using Contrastive Divergence with a quantum Gibbs sampling step derived here. The first, the **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**. The second, the **Q**uantum **G**aussian-Bernoulli **R**estricted **B**oltzmann **M**achine (QGRBM), is a Gaussian-Bernoulli RBM modified to include the transverse field over the Hidden units. These models scale to standard ML datasets on classical computers and perform $20 - 65\%$ better than their classical counterparts on similar training budget with no additional hyperparameter tuning. Additional experimental details are provided in Appendix D[2].

7 Discussion

This work addresses a longstanding open problem in quantum machine learning: how to train latent variable models defined by density operators without relying on quantum conditional probability, which has no well-defined analog in the operator-theoretic setting. We resolve this by reframing maximum-likelihood training as a quantum information projection problem, yielding the DO-EM algorithm with provable convergence guarantees. We additionally introduce V-locality as a unifying structural condition, CQ-LVMs as a tractable model family for classical data, and provide the first empirical evidence that density operator models offer a modeling advantage on standard image benchmarks. Tractable model classes for the E-step and whether the favorable comparisons against probabilistic models is an improvement in raw representation capacity or due to the optimization dynamics of the geometry of PSD operators is an important question for future work.

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

Acknowledgments and Disclosure of Funding

The authors thank Prof. Apoorva Patel (IISc) and Prof. Shiro Ikeda (ISM) for their valuable suggestions. AV thanks Samarth Hawaldar (ISTA) for many engaging discussions on quantum mechanics. The initial stages of this project was supported by a research grant from Accenture India.

References

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

[2] Iordanis Kerenidis, Jonas Landman, Alessandro Luongo, and Anupam Prakash. q-means: A quantum algorithm for unsupervised machine learning. In *Advances in Neural Information Processing Systems*, pages 4136–4146, 2019.

[3] Iordanis Kerenidis, Alessandro Luongo, and Anupam Prakash. Quantum expectation-maximization for Gaussian mixture models. In *International Conference on Machine Learning*, pages 5187–5197, 7 2020.

[4] Hideyuki Miyahara, Kazuyuki Aihara, and Wolfgang Lechner. Quantum expectation-maximization algorithm. *Phys. Rev. A*, 101(1):012326, 2020.

[2]Code for our experiments are available here.

[5] Youngseok Kim, Andrew Eddins, Sajant Anand, Ken Xuan Wei, Ewout van den Berg, Sami Rosenblatt, Hasan Nayfeh, Yantao Wu, Michael Zaletel, Kristan Temme, and Abhinav Kandala. Evidence for the utility of quantum computing before fault tolerance. *Nature*, 618(7965): 500–505, 2023.

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

[7] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195–202, 9 2017.

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

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

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

[11] M. Schuld and N. Killoran. Quantum machine learning in feature hilbert spaces. *Physical Review Letters*, 122(4):040504, 2019.

[12] Francis Bach. Information theory with kernel methods. *IEEE Transactions on Information Theory*, 69(2):752–775, 2023. doi: 10.1109/TIT.2022.3211077.

[13] Eliza O'Reilly and Venkat Chandrasekaran. Spectrahedral regression. *SIAM Journal on Optimization*, 33(2):553–588, 2023. doi: 10.1137/21M1455899.

[14] Francis Bach. Sum-of-squares relaxations for information theory and variational inference. *Foundations of Computational Mathematics*, 25(3):865–903, 2025.

[15] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

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

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

[18] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

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

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

[22] 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.

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

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

[25] Winton Brown and David Poulin. Quantum markov networks and commuting hamiltonians, 2012.

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

[27] Eric R. Anschuetz and Yudong Cao. Realizing quantum boltzmann machines through eigenstate thermalization. *arXiv preprint arXiv:1903.01359*, 2019.

[28] Christa Zoufal, Aurélien Lucchi, and Stefan Woerner. Variational quantum boltzmann machines. *Quantum Machine Intelligence*, 3(1):7, 2021.

[29] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. Adaptive computation and machine learning. MIT Press, 2009. ISBN 9780262013192.

[30] G. R. Grimmett. A theorem about random fields. *Bulletin of the London Mathematical Society*, 5(1):81–84, 1973.

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

[32] 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.

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

[34] Ruslan Salakhutdinov and Geoffrey Hinton. Deep boltzmann machines. In *International Conference on Artificial Intelligence and Statistics*, pages 448–455, 2009.

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

[36] Shohei Taniguchi, Masahiro Suzuki, Yusuke Iwasawa, and Yutaka Matsuo. End-to-end training of deep boltzmann machines by unbiased contrastive divergence with local mode initialization. In *International Conference on Machine Learning*, pages 33804–33815, 2023.

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

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

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

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

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

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

[43] David Sutter, Omar Fawzi, and Renato Renner. Universal recovery map for approximate Markov chains. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 472(2186):20150623, 2016. doi: 10.1098/rspa.2015.0623.

[44] David Sutter, Marco Tomamichel, and Aram W. Harrow. Strengthened monotonicity of relative entropy via pinched Petz recovery map. *IEEE Transactions on Information Theory*, 62(5): 2907–2913, 2016. doi: 10.1109/TIT.2016.2545680.

[45] Marius Junge, Renato Renner, David Sutter, Mark M. Wilde, and Andreas Winter. Universal recovery maps and approximate sufficiency of quantum relative entropy. *Annales Henri Poincaré*, 19(10):2955–2978, 2018. doi: 10.1007/s00023-018-0716-0.

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

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

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

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

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

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

[52] Takeshi Kimura, Kohtaro Kato, and Masahito Hayashi. Structured quantum learning via em algorithm for boltzmann machines, 2026.

[53] Sever Silvestru 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.

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

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

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

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

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

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

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

[61] Shun-ichi Amari. Information geometry of the em and em algorithms for neural networks. *Neural Networks*, 8(9):1379–1408, 1995.

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

[63] I. Csiszár. i-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975. doi: 10.1214/aop/1176996454.

[64] I. Csiszar and F. Matus. Information projections revisited. *IEEE Transactions on Information Theory*, 49(6):1474–1490, 2003. doi: 10.1109/TIT.2003.810633.

[65] Ge Bai, Francesco Buscemi, and Valerio Scarani. Quantum bayes' rule and petz transpose map from the minimum change principle. *Phys. Rev. Lett.*, 135:090203, Aug 2025.

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

[67] C. F. Jeff Wu. On the convergence properties of the em algorithm. *The Annals of Statistics*, 11 (1):95–103, 1983. ISSN 00905364, 21688966.

[68] John Watrous. *The Theory of Quantum Information*. Cambridge University Press, Cambridge, 2018.

[69] Patrick Hayden, Richard Jozsa, Dénes Petz, and Andreas Winter. Structure of states which satisfy strong subadditivity of quantum entropy with equality. *Communications in Mathematical Physics*, 246(2):359–374, 2004.

[70] Mary Beth Ruskai. Inequalities for quantum entropy: A review with conditions for equality. *Journal of Mathematical Physics*, 43(9):4358–4375, 2002.

[71] Dhrumil Patel, Daniel Koch, Saahil Patel, and Mark M. Wilde. Quantum boltzmann machine learning of ground-state energies. *arXiv preprint arXiv:2410.12935*, 2024.

[72] Ersen Bilgin and Sergio Boixo. Preparing thermal states of quantum systems by dimension reduction. *Phys. Rev. Lett.*, 105:170405, 2010.

[73] Chi-Fang Chen, Michael J. Kastoryano, Fernando G. S. L. Brandão, and András Gilyén. Quantum thermal state preparation. *arXiv preprint arXiv:2303.18224*, 2023.

[74] Zhiyan Ding, Bowen Li, and Lin Lin. Efficient quantum gibbs samplers with kubo–martin–schwinger detailed balance condition. *Communications in Mathematical Physics*, 406(3):67, 2025.

[75] Marcello Benedetti, John Realpe-Gómez, Rupak Biswas, and Alejandro Perdomo-Ortiz. Quantum-assisted learning of hardware-embedded probabilistic graphical models. *Phys. Rev. X*, 7:041052, 2017.

[76] Jon Nelson, Marc Vuffray, Andrey Y. Lokhov, Tameem Albash, and Carleton Coffrin. High-quality thermal gibbs sampling with quantum annealing hardware. *Phys. Rev. Appl.*, 17:044046, 2022.

[77] Ryosuke Shibukawa, Ryo Tamura, and Koji Tsuda. Boltzmann sampling with quantum annealers via fast stein correction. *Phys. Rev. Res.*, 6:043050, 2024.

[78] Rajendra Bhatia. *Matrix Analysis*. Springer, 1997.

[79] Alexander M. Ostrowski. *Solution of Equations and Systems of Equations*, volume 9 of *Pure and Applied Mathematics*. Academic Press, New York, 2 edition, 1966.

[80] Hugo Larochelle and Yoshua Bengio. Classification using discriminative restricted boltzmann machines. In *International Conference on Machine Learning*, page 536–543, 2008.

[81] Ruslan Salakhutdinov, Andriy Mnih, and Geoffrey Hinton. Restricted boltzmann machines for collaborative filtering. In *International Conference on Machine Learning*, page 791–798, 2007.

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

[83] Miguel Á. Carreira-Perpiñán and Geoffrey Hinton. On contrastive divergence learning. In *International Workshop on Artificial Intelligence and Statistics*, pages 33–40, 06–08 Jan 2005.

[84] Max Welling, Michal Rosen-zvi, and Geoffrey E Hinton. Exponential family harmoniums with an application to information retrieval. In *Advances in Neural Information Processing Systems*, 2004.

[85] Renjie Liao, Simon Kornblith, Mengye Ren, David J. Fleet, and Geoffrey Hinton. Gaussian-bernoulli RBMs without tears. *arXiv preprint arXiv:2210.10318*, 2022.

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

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

[88] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[89] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015.

Appendix

In this appendix, we provide the following material.

- In Section A, we provided a discussion on density operators,quantum channels, and related work in training density operator models.
- In Section B, we provide the proofs of the theoretical results proposed in the main body.
- In Section C, we discuss Boltzmann machines and the Contrastive Divergence algorithm.
- In Section D, we provide implementation details and additional experiments with the DO-EM algorithm.
- In Section E, we discuss the computational setup used in our experiments.
- In Section F, we discuss the limitations of this work.

Code

The code to reproduce the experiments in this paper can be found on GitHub.

A Density operators

Current research explores leveraging quantum mechanics for machine learning either by achieving computational speedups with quantum hardware [2–5], or by capturing more complex data structures through quantum formalisms [6, 7]. The use of density operators, quantum generalizations of classical probability distributions, to model data represents an effort in the latter direction.

A density operator on a Hilbert space $\mathcal{H}$ is a Hermitian, positive semi-definite operator with unit trace [37, 38]. The set of Density operators will be denoted by $\mathcal{P}(\mathcal{H})$, and can be regarded as generalizations of probability distributions. We now define quantum channels that act on density operators.

Let $\mathcal{S}(\mathcal{H})$ denote the set of square linear operators on a finite dimensional Hilbert space $\mathcal{H}$. A linear map $\mathcal{N} : \mathcal{S}(\mathcal{H}_A) \to (\mathcal{H}_B)$ is *positive* if $\mathcal{N}(X)$ is positive semi-definite for all positive semi-definite operators X in $\mathcal{S}(\mathcal{H})$. 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 A.1 (Quantum Channel)**.** *A quantum channel $\mathcal{N} : \mathcal{S}(\mathcal{H}_A) \to (\mathcal{H}_B)$ is a linear, completely positive, trace preserving map.*

We now define an inner product on operators.

Definition A.2 (Hilbert-Schmidt Inner Product)**.** *The Hilbert-Schmidt Inner Product between two operators X, Y in $\mathcal{S}(\mathcal{H})$ is defined as $\langle X, Y\rangle = \mathrm{Tr}(X^\dagger Y)$.*

The inner product allows us to define the adjoint of a quantum channel.

Definition A.3. *The adjoint of a quantum channel $\mathcal{N} : \mathcal{S}(\mathcal{H}_A) \to (\mathcal{H}_B)$ is the unique linear map $\mathcal{N}^\dagger : \mathcal{S}(\mathcal{H}_B) \to (\mathcal{H}_A)$ satisfying $\langle Y, \mathcal{N}(X)\rangle = \langle \mathcal{N}^\dagger(Y), X\rangle$ for all X in $\mathcal{S}(\mathcal{H}_A)$ and Y in $\mathcal{S}(\mathcal{H}_B)$.*

Lindblad [40] showed that the relative entropyis monotonic under the action of a quantum channel.

Theorem A.4 (Monotonicity of Relative Entropy)**.** *For density operators ω and ρ in $\mathcal{P}(\mathcal{H}_A \otimes \mathcal{H}_B)$ such that $\ker(\rho) \subset \ker(\omega)$, $\mathrm{D}_{\mathrm{U}}(\omega, \rho) \geq \mathrm{D}_{\mathrm{U}}(\mathcal{N}(\omega), \mathcal{N}(\rho))$.*

Petz [41, 42] showed that `MRE` is saturated if and only if the Petz Recovery Map reverses the quantum channel. We provide a resatement of this result from Hayden et al. [69].

Theorem A.5. *For states ω and ρ, $\mathrm{D}_{\mathrm{U}}(\omega, \rho) = \mathrm{D}_{\mathrm{U}}(\mathcal{N}(\omega), \mathcal{N}(\rho))$ if and only if there exists a quantum channel $\mathcal{R}$ such that $\mathcal{R}(\mathcal{N}(\omega)) = \omega$ and $\mathcal{R}(\mathcal{N}(\rho)) = \rho$. Furthermore, on the support of $\mathcal{N}(\rho)$, $\mathcal{R}$ can be given explicitly by the formula*

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

This lead to the definition of the Petz Recovery Map.

Definition A.6 (Petz Recovery Map)**.** *For a density operator ρ in $\mathcal{P}(\mathcal{H}_A)$, the Petz Recovery Map for a quantum channel $\mathcal{N} : \mathcal{S}(\mathcal{H}_A) \to \mathcal{S}(\mathcal{H}_B)$ is the map $\mathcal{R}_{\mathcal{N},\rho} : \mathcal{S}(\mathcal{H}_B) \to \mathcal{S}(\mathcal{H}_A)$ defined by*

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

The Petz Recovery Map is a completely positive, trace non-increasing map [38] that does not always map density operators to density operators. Ruskai [70] proved an equivalent condition for the saturation of MRE.

Theorem A.7 (Ruskai's condition)**.** *For density operators ω and ρ in $\mathcal{P}(\mathcal{H}_A \otimes \mathcal{H}_B)$ such that $\ker(\rho) \subset \ker(\omega)$, $\mathrm{D_U}(\mathcal{N}(\omega), \mathcal{N}(\rho)) = \mathrm{D_U}(\omega, \rho)$ if and only if*

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

.

In this work, we focus on the quantum channel defined by the partial trace operation $\mathcal{N}(\rho) = \mathrm{Tr}_B\rho$ and its adjoint $\mathcal{N}^\dagger(\rho_A) = \rho_A \otimes \mathrm{I}_B$. All our results are specialized to this case.

A.1 Related work on Density Operator Models

Most recent works on density operator models (**DOMs**) consider settings without hidden units [24, 46, 47, 71]due to operator-theoretic difficulties brought about by the partial trace operation during gradient evaluation [26]. Other studies considered hidden variable **DOMs**, where the gradient evaluation was accomplished using heuristic methods [8, 27, 28]. As such, these methods rely on the approximate (variational) preparation of the quantum Gibbs state density matrix [72–74]. Sampling from a Gibbs state distribution can also be performed using a quantum annealer, which may offer advantages even for classical machine learning models [8, 75–77].

These existing approaches rarely scale beyond 12 visible units—limited by restricted access to quantum hardware, the exponential cost of simulating quantum systems, and the memory bottlenecks associated with representing and optimizing DO-LVMs on classical devices. EM based algorithms can provide a simpler alternative to existing learning algorithms for DO-LVMs which directly maximizes the likelihood. However deriving such algorithms in the density operator theoretic setup is extremely challenging for a variety of reasons. *Most notably operator theoretic analogues of classical inequalities, such as the Jensen Inequality, cannot be directly applied to derive an Evidence lower bound (ELBO) for DO-LVMs.*

B Proofs

B.1 Proofs from Section 3

Several inequalities related to convex functions can be extended to such functions. We provide the finite-dimensional restatement of an inequality in Dragomir [53].

Theorem B.1 (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*

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

We now prove the inequality on log-likelihoods.

Proof of Proposition 3.2. 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 $\mathcal{T}(\mathcal{H}_\mathrm{v})$. Then by Theorem B.1, we have the inequality

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

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. □

We now prove that V-locality on all partitions implies shield commutativity.

Proof of Theorem 3.4. We prove this theorem in two parts.
V-locality to shield commutativity. Fix a partition $\Lambda = A \sqcup B \sqcup C$ such that B shields A from C. Apply the hypothesis to the bipartition

$$V := A \cup B, \qquad L := C,$$

and decompose $H = H_V + H_L + H_{VL}$ as in Definition 3.3. By assumption, $[H_V, H_{VL}] = 0$. Define

$$X := H_V, \qquad Y := H_L + H_{VL}.$$

We verify the three requirements for shield commutation.

Decomposition. By construction $X + Y = H_V + H_L + H_{VL} = H$.

Supports. The operator $X = H_V$ is a sum of clique terms contained in $V = A \cup B$, so $\mathrm{supp}(X) \subseteq A \cup B$. The term H_L is a sum of clique terms in $L = C$, hence supported in $B \cup C$. We claim that H_{VL} is also supported in $B \cup C$, equivalently that no clique contributing to H_{VL} contains a vertex of A. Suppose for contradiction that a clique Q contributing to H_{VL} contains a vertex $a \in A$; since Q intersects $L = C$, it also contains some $c \in C$. By the clique property, a and c are adjacent in G, so the edge $\{a, c\}$ is a path from A to C that avoids B, contradicting the shielding hypothesis. Hence $\mathrm{supp}(H_{VL}) \subseteq B \cup C$, and therefore $\mathrm{supp}(Y) \subseteq B \cup C$.

Commutation. We have

$$[X, Y] = [H_V, H_L] + [H_V, H_{VL}].$$

The first commutator vanishes because H_V is supported in V and H_L in L, with $V \cap L = \emptyset$. The second vanishes by the V-local commutation condition. Hence $[X, Y] = 0$.

Since $A|B|C$ was arbitrary, H is shield-commuting.

Shield commutation implies V-local for all bipartitions when G is triangle-free. The argument proceeds in two steps. We first invoke Brown–Poulin to extract a clique decomposition with pairwise commuting clique terms, then show that any such decomposition is V-local on every bipartition.

Step 1: Extracting a local-commuting decomposition. On a triangle-free graph, every clique of G is either a single vertex or a single edge. By Brown–Poulin [25, Theorem 4], every shield-commuting Hamiltonian admits a clique decomposition

$$H = \sum_{e \in E} h_e + \sum_{v \in \Lambda} h_v, \qquad [h_Q, h_{Q'}] = 0 \text{ for all } Q, Q' \in \mathcal{C}(G),$$

where h_e is supported on edge e, h_v is supported on vertex v, and any two clique terms commute. We work with this commuting decomposition for the remainder of the proof; in particular, H_V, H_L, and H_{VL} in what follows refer to the decomposition of H relative to $\{h_e, h_v\}$.

Step 2: Local commutation gives V-locality on every bipartition. Fix any bipartition $\Lambda = V \sqcup L$. By Definition 3.3,

$$H_V = \sum_{Q \subseteq V} h_Q, \qquad H_L = \sum_{Q \subseteq L} h_Q, \qquad H_{VL} = \sum_{\substack{Q \in \mathcal{C}(G) \\ Q \cap V \neq \emptyset,\ Q \cap L \neq \emptyset}} h_Q.$$

Both H_V and H_{VL} are linear combinations of clique terms h_Q. By local commutation from Step 1, every clique term commutes with every other clique term. Linearity of the commutator gives

$$[H_V, H_{VL}] = \sum_{\substack{Q \subseteq V \\ Q' : Q' \cap V \neq \emptyset,\ Q' \cap L \neq \emptyset}} [h_Q, h_{Q'}] = 0.$$

Hence H is V-local on this bipartition. Since the bipartition was arbitrary, H is V-local on every bipartition. □

B.2 Proofs from Section 4

Establishing the quantum evidence lower bound.

Proof of Proposition 4.1. Since $\mathrm{Tr}_{\mathrm{L}} \eta = \eta_{\mathrm{V}}$, by Theorem 2.2 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

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

which is the required inequality. □

This proof reduces to a proof of the classical ELBO from the chain rule of KL divergence [66] when $\rho(\theta)$ is diagonal.

Solving QIP

Lemma B.2 (Support reduction)**.** *Let $\eta_{\mathrm{V}} \in \mathcal{P}(\mathcal{H}_{\mathrm{V}})$ have support $\mathcal{S}$ with projector $P_{\mathcal{S}}$. Every feasible η for the QIP problem satisfies*

$$(P_{\mathcal{S}} \otimes \mathrm{I}_{\mathrm{L}})\, \eta\, (P_{\mathcal{S}} \otimes \mathrm{I}_{\mathrm{L}}) = \eta.$$

Proof. Let $|v_0\rangle \in \ker(\eta_{\mathrm{V}})$ and let $\{|e_k\rangle\}$ be an orthonormal basis of $\mathcal{H}_{\mathrm{L}}$. Then

$$0 = \langle v_0|\, \eta_{\mathrm{V}}\, |v_0\rangle = \langle v_0|\, \mathrm{Tr}_{\mathrm{L}} \eta\, |v_0\rangle = \sum_k \langle v_0 \otimes e_k|\, \eta\, |v_0 \otimes e_k\rangle.$$

Each summand is non-negative since $\eta \succeq 0$, so each vanishes. Combined with $\eta \succeq 0$, this forces $\eta\,(|v_0\rangle \otimes |e_k\rangle) = 0$ for all k, hence η vanishes on $|v_0\rangle \otimes \mathcal{H}_{\mathrm{L}}$. Since $|v_0\rangle$ was arbitrary in $\ker(\eta_{\mathrm{V}}) = \mathcal{S}^{\perp}$, the operator η vanishes on $\mathcal{S}^{\perp} \otimes \mathcal{H}_{\mathrm{L}}$, equivalent to the stated identity. □

Let $\mathcal{T}_0(\mathcal{S}) := \{\Lambda \in \mathcal{T}(\mathcal{S}) : \mathrm{Tr}\Lambda = 0\}$ denote the real vector space of traceless Hermitian operators on $\mathcal{S}$. Let $r = \mathrm{rank}(\eta_{\mathrm{V}}) > 0$. Since $\mathrm{Tr}\, \mathrm{P}_{\mathcal{S}} = r > 0$, the decomposition

$$\mathcal{T}(\mathcal{S}) = \mathcal{T}_0(\mathcal{S}) \oplus \mathbb{R}\, P_{\mathcal{S}}, \qquad \Lambda = \Big(\Lambda - \tfrac{\mathrm{Tr}\Lambda}{r}\, P_{\mathcal{S}}\Big) + \tfrac{\mathrm{Tr}\Lambda}{r}\, P_{\mathcal{S}}, \tag{3}$$

is direct, and $\dim \mathcal{T}_0(\mathcal{S}) = r^2 - 1$.

For any Hermitian operator $\Lambda \in \mathcal{T}(\mathcal{S})$, define

$$\eta_\Lambda := \frac{1}{\mathcal{Z}(\Lambda)} \exp\big(\mathrm{H}_{\mathcal{S}} + \Lambda \otimes \mathrm{I}_{\mathrm{L}}\big), \qquad \mathcal{Z}(\Lambda) := \mathrm{Tr} \exp\big(\mathrm{H}_{\mathcal{S}} + \Lambda \otimes \mathrm{I}_{\mathrm{L}}\big), \tag{4}$$

where the exponential and trace are taken on $\mathcal{S} \otimes \mathcal{H}_{\mathrm{L}}$ and $\Lambda \otimes \mathrm{I}_{\mathrm{L}}$ is regarded as an operator on this space via the inclusion $\mathcal{S} \hookrightarrow \mathcal{H}_{\mathrm{V}}$. Then $\eta_\Lambda \in \mathcal{P}(\mathcal{S} \otimes \mathcal{H}_{\mathrm{L}})$ is faithful for every $\Lambda \in \mathcal{T}(\mathcal{S})$, and we identify it with an element of $\mathcal{P}(\mathcal{H}_{\mathrm{V}} \otimes \mathcal{H}_{\mathrm{L}})$ supported on $\mathcal{S} \otimes \mathcal{H}_{\mathrm{L}}$. Note that η_Λ is invariant under $\Lambda \mapsto \Lambda + c\, P_{\mathcal{S}}$ for any $c \in \mathbb{R}$: such a shift multiplies the exponential by e^c, which cancels in the normalization. Hence η_Λ depends only on the trace-zero component of Λ in the decomposition (3).

Define the log-partition function

$$\Psi : \mathcal{T}(\mathcal{S}) \to \mathbb{R}, \qquad \Psi(\Lambda) := \log \mathrm{Tr} \exp\big(\mathrm{H}_{\mathcal{S}} + \Lambda \otimes \mathrm{I}_{\mathrm{L}}\big). \tag{5}$$

Ψ is real-analytic. We record two properties used in the proof of Theorem 4.3.

Lemma B.3 (Strict convexity)**.** Ψ *is strictly convex on* $\mathcal{T}_0(\mathcal{S})$.

Proof. The Peirels-Bogoliubov inequality states that the function $A \mapsto \log \mathrm{Tr} e^A$ is convex on Hermitian operators, with strict convexity except along directions of the form $A + c\,\mathrm{I}$ [78]. The map $\Lambda \mapsto \mathrm{H}_{\mathcal{S}} + \Lambda \otimes \mathrm{I}_{\mathrm{L}}$ is affine in Λ, so Ψ inherits convexity, with strict convexity along directions $\delta\Lambda$ such that $\delta\Lambda \otimes \mathrm{I}_{\mathrm{L}}$ is not a scalar multiple of $\mathrm{I}_{\mathcal{S}\otimes\mathcal{H}_{\mathrm{L}}}$. The condition $\delta\Lambda \otimes \mathrm{I}_{\mathrm{L}} = c\,\mathrm{I}_{\mathcal{S}\otimes\mathcal{H}_{\mathrm{L}}}$ is equivalent to $\delta\Lambda = c\,P_{\mathcal{S}}$, and $\mathcal{T}_0(\mathcal{S}) \cap \mathbb{R}\,P_{\mathcal{S}} = \{0\}$ since $\mathrm{Tr} P_{\mathcal{S}} = r \neq 0$. □

Lemma B.4 (Gradient)**.** $\nabla\Psi(\Lambda) = \mathrm{Tr}_{\mathrm{L}}\eta_\Lambda$ *under the Hilbert–Schmidt inner product on* $\mathcal{T}(\mathcal{S})$.

Proof. For $\delta\Lambda \in \mathcal{T}(\mathcal{S})$, Duhamel's formula gives

$$\frac{d}{dt}\Big|_{t=0} \exp\big(\mathrm{H}_{\mathcal{S}} + (\Lambda + t\,\delta\Lambda) \otimes \mathrm{I}_{\mathrm{L}}\big) = \int_0^1 e^{sA(\Lambda)}\,(\delta\Lambda \otimes \mathrm{I}_{\mathrm{L}})\,e^{(1-s)A(\Lambda)}\,ds,$$

where $A(\Lambda) := \mathrm{H}_{\mathcal{S}} + \Lambda \otimes \mathrm{I}_{\mathrm{L}}$. Taking the trace, applying cyclicity to combine the exponentials, and dividing by $\mathrm{Tr} e^{A(\Lambda)}$,

$$\frac{d}{dt}\Big|_{t=0} \Psi(\Lambda + t\,\delta\Lambda) = \mathrm{Tr}\big((\delta\Lambda \otimes \mathrm{I}_{\mathrm{L}})\,\eta_\Lambda\big) = \mathrm{Tr}(\delta\Lambda \cdot \mathrm{Tr}_{\mathrm{L}}\eta_\Lambda),$$

the second equality by the definition of partial trace. Since this holds for all $\delta\Lambda \in \mathcal{T}(\mathcal{S})$, the right-hand side identifies $\nabla\Psi(\Lambda) = \mathrm{Tr}_{\mathrm{L}}\eta_\Lambda$ via the defining identity $\frac{d}{dt}|_{t=0}\Psi(\Lambda + t\,\delta\Lambda) = \langle\nabla\Psi(\Lambda),\, \delta\Lambda\rangle_{\mathrm{HS}}$ of the Hilbert–Schmidt gradient. □

Proof of Theorem 4.3. The proof proceeds in three steps. Step 1 establishes existence and uniqueness of the QIP minimizer η^* by compactness and strict convexity. Step 2 identifies its form via Lagrangian stationarity. Step 3 shows that the resulting Λ^* is the unique minimizer of the convex program F stated in the theorem.

Step 1: Existence and uniqueness of η^*. The QIP feasible set

$$\mathcal{F} := \{\eta \in \mathcal{P}(\mathcal{H}_{\mathrm{V}} \otimes \mathcal{H}_{\mathrm{L}}) : \mathrm{Tr}_{\mathrm{L}}\eta = \eta_{\mathrm{V}}\}$$

is convex (intersection of the convex set $\mathcal{P}(\mathcal{H}_{\mathrm{V}}\otimes\mathcal{H}_{\mathrm{L}})$ with the affine constraint $\mathrm{Tr}_{\mathrm{L}}\eta = \eta_{\mathrm{V}}$), nonempty (it contains $\eta_{\mathrm{V}} \otimes \mathrm{I}_{\mathrm{L}}/d_{\mathrm{L}}$), and compact in finite dimensions (closed, since Tr_{L} is continuous, and bounded, since density operators have unit trace). The objective $\mathrm{D}_{\mathrm{U}}(\cdot, \rho)$ is continuous and strictly convex on its domain. Hence D_{U} attains a unique minimum on $\mathcal{F}$ at some $\eta^* \in \mathcal{F}$.

Step 2: Lagrangian form of η^*. By Lemma B.2, every $\eta \in \mathcal{F}$ is supported on $\mathcal{S} \otimes \mathcal{H}_{\mathrm{L}}$, and on this subspace $\mathrm{Tr}(\eta\,\mathrm{H}) = \mathrm{Tr}(\eta\,\mathrm{H}_{\mathcal{S}})$. With $\rho = e^{\mathrm{H}}/Z$, $\log\rho = \mathrm{H} - (\log Z)\,\mathrm{I}$, so

$$\mathrm{D}_{\mathrm{U}}(\eta, \rho) = -\mathrm{S}_{\mathrm{VN}}(\eta) - \mathrm{Tr}(\eta\,\mathrm{H}_{\mathcal{S}}) + \log Z$$

for η supported on $\mathcal{S} \otimes \mathcal{H}_{\mathrm{L}}$.

We first show that η^* is positive definite on $\mathcal{S} \otimes \mathcal{H}_{\mathrm{L}}$. The feasible state $\eta_0 := \eta_{\mathrm{V}} \otimes \mathrm{I}_{\mathrm{L}}/d_{\mathrm{L}}$ is positive definite on this space. If η^* were singular there, let P_0 project onto its kernel and set $\eta_t = (1-t)\eta^* + t\eta_0$. The eigenvalues emerging from the kernel contribute

$$\mathrm{Tr}(\eta_t \log\eta_t) - \mathrm{Tr}(\eta^* \log\eta^*) = c\,t\log t + O(t), \qquad c = \mathrm{Tr}(P_0\eta_0) > 0.$$

The remaining term in the objective is affine in η_t, hence changes by $O(t)$. Therefore $\mathrm{D}_{\mathrm{U}}(\eta_t, \rho) - \mathrm{D}_{\mathrm{U}}(\eta^*, \rho) = c\,t\log t + O(t) < 0$ for sufficiently small $t > 0$, contradicting optimality. Thus η^* is positive definite on the reduced space, and Lagrangian stationarity applies there.

The Lagrangian for the partial-trace constraint $\mathrm{Tr}_{\mathrm{L}}\eta = \eta_{\mathrm{V}}$, with multiplier $\widetilde{\Lambda} \in \mathcal{T}(\mathcal{S})$ (only the $\mathcal{S}$-component of the multiplier is active by Lemma B.2), is

$$\mathcal{L}(\eta, \widetilde{\Lambda}) = \mathrm{D}_{\mathrm{U}}(\eta, \rho) - \mathrm{Tr}\big(\widetilde{\Lambda}\,(\mathrm{Tr}_{\mathrm{L}}\eta - \eta_{\mathrm{V}})\big).$$

Using the adjoint of the partial trace, $\mathrm{Tr}(\widetilde{\Lambda}\,\mathrm{Tr}_{\mathrm{L}}\eta) = \mathrm{Tr}((\widetilde{\Lambda} \otimes \mathrm{I}_{\mathrm{L}})\eta)$. Stationarity on the affine space of unit-trace operators therefore gives, for some scalar c,

$$\log\eta^* - \mathrm{H}_{\mathcal{S}} - \widetilde{\Lambda} \otimes \mathrm{I}_{\mathrm{L}} = c\,\mathrm{I}_{\mathcal{S}\otimes\mathcal{H}_{\mathrm{L}}}.$$

Exponentiating and imposing unit trace yields

$$\eta^* = \frac{\exp(\mathrm{H}_{\mathcal{S}} + \widetilde{\Lambda} \otimes \mathrm{I}_{\mathrm{L}})}{\mathcal{Z}(\widetilde{\Lambda})} = \eta_{\widetilde{\Lambda}}.$$

Here the exponential and trace are taken on $\mathcal{S} \otimes \mathcal{H}_{\mathrm{L}}$, and the resulting state is extended by zero to the full joint space.

Define

$$\Lambda^* := \widetilde{\Lambda} - \tfrac{\mathrm{Tr}\widetilde{\Lambda}}{r} P_{\mathcal{S}} \ \in \ \mathcal{T}_0(\mathcal{S}).$$

By Lemma B.2, $\eta_{\Lambda^*} = \eta_{\widetilde{\Lambda}} = \eta^*$, so $\eta^*(\rho, \eta_{\mathrm{V}}) = \eta_{\Lambda^*}$ with $\Lambda^* \in \mathcal{T}_0(\mathcal{S})$. Feasibility $\mathrm{Tr}_{\mathrm{L}}\eta^* = \eta_{\mathrm{V}}$ gives

$$\mathrm{Tr}_{\mathrm{L}}\eta_{\Lambda^*} = \eta_{\mathrm{V}}. \tag{6}$$

Step 3: Convex-program characterization. Let

$$F : \mathcal{T}_0(\mathcal{S}) \to \mathbb{R}, \qquad F(\Lambda) := \Psi(\Lambda) - \mathrm{Tr}(\Lambda\, \eta_{\mathrm{V}}).$$

F is strictly convex on $\mathcal{T}_0(\mathcal{S})$ as the sum of a strictly convex function (Lemma B.3) and a linear function. By Lemma B.4, its gradient is

$$\nabla F(\Lambda) = \mathrm{Tr}_{\mathrm{L}}\eta_{\Lambda} - \eta_{\mathrm{V}}.$$

The feasibility condition (6) therefore reads $\nabla F(\Lambda^*) = 0$. By strict convexity of F, this critical point is the unique global minimizer of F over $\mathcal{T}_0(\mathcal{S})$. This proves the convex-program characterization of Λ^* stated in the theorem. □

Proving data processing defect

Proof of Corollary 4.4. Write $\eta^* := \eta^*(\rho, \eta_{\mathrm{V}})$ and let $\mathcal{S} := \mathrm{supp}(\eta_{\mathrm{V}}) \subseteq \mathcal{H}_{\mathrm{V}}$ with projector $P_{\mathcal{S}}$. Since ρ is faithful, write $\rho = e^{\mathrm{H}}/Z$ with H Hermitian and $Z = \mathrm{Tr}e^{\mathrm{H}}$, and set $\mathrm{H}_{\mathcal{S}} := (P_{\mathcal{S}} \otimes \mathrm{I}_{\mathrm{L}})\, \mathrm{H}\, (P_{\mathcal{S}} \otimes \mathrm{I}_{\mathrm{L}})$. By Theorem 4.3, η^* has support $\mathcal{S} \otimes \mathcal{H}_{\mathrm{L}}$ and on this support

$$\log \eta^* = \mathrm{H}_{\mathcal{S}} + \Lambda^* \otimes \mathrm{I}_{\mathrm{L}} - \log \mathcal{Z}(\Lambda^*)\, \mathrm{I}, \tag{7}$$

with $\Lambda^* \in \mathcal{T}_0(\mathcal{S})$ the unique minimizer from Theorem 4.3 and $\mathrm{Tr}_{\mathrm{L}}\eta^* = \eta_{\mathrm{V}}$.

Step 1: η is supported on $\mathcal{S} \otimes \mathcal{H}_{\mathrm{L}}$. Since $\eta_{\mathrm{V}} = \mathrm{Tr}_{\mathrm{L}}\eta$ has support $\mathcal{S}$, Lemma B.2 gives $(P_{\mathcal{S}} \otimes \mathrm{I}_{\mathrm{L}})\, \eta\, (P_{\mathcal{S}} \otimes \mathrm{I}_{\mathrm{L}}) = \eta$. In particular, both $\mathrm{Tr}(\eta \log \eta^*)$ and $\mathrm{Tr}(\eta \log \rho)$ are finite.

Step 2: A trace identity depending only on ρ and η_{V}. By Step 1 and cyclicity of the trace,

$$\mathrm{Tr}(\eta\, \mathrm{H}) = \mathrm{Tr}\big((P_{\mathcal{S}} \otimes \mathrm{I}_{\mathrm{L}})\, \eta\, (P_{\mathcal{S}} \otimes \mathrm{I}_{\mathrm{L}})\, \mathrm{H}\big) = \mathrm{Tr}\big(\eta\, (P_{\mathcal{S}} \otimes \mathrm{I}_{\mathrm{L}})\, \mathrm{H}\, (P_{\mathcal{S}} \otimes \mathrm{I}_{\mathrm{L}})\big) = \mathrm{Tr}(\eta\, \mathrm{H}_{\mathcal{S}}).$$

Combined with $\log \rho = \mathrm{H} - (\log Z)\, \mathrm{I}$,

$$\mathrm{Tr}(\eta\ \log \rho) = \mathrm{Tr}(\eta\, \mathrm{H}_{\mathcal{S}}) - \log Z. \tag{8}$$

From (7) and (8),

$$\begin{aligned} \mathrm{Tr}(\eta\ \log \eta^*) - \mathrm{Tr}(\eta\ \log \rho) &= \mathrm{Tr}(\eta\, \mathrm{H}_{\mathcal{S}}) + \mathrm{Tr}\big((\Lambda^* \otimes \mathrm{I}_{\mathrm{L}})\, \eta\big) - \log \mathcal{Z}(\Lambda^*) - \mathrm{Tr}(\eta\, \mathrm{H}_{\mathcal{S}}) + \log Z \\ &= \mathrm{Tr}(\Lambda^*\, \eta_{\mathrm{V}}) + \log Z - \log \mathcal{Z}(\Lambda^*), \end{aligned} \tag{9}$$

where the last step uses $\mathrm{Tr}\big((\Lambda^* \otimes \mathrm{I}_{\mathrm{L}})\, \eta\big) = \mathrm{Tr}(\Lambda^* \cdot \mathrm{Tr}_{\mathrm{L}}\eta) = \mathrm{Tr}(\Lambda^*\, \eta_{\mathrm{V}})$ (adjoint of the partial trace). The right-hand side of (9) depends only on ρ and η_{V}.

Step 3: Pythagorean identity. Since $\mathrm{Tr}_{\mathrm{L}}\eta^* = \eta_{\mathrm{V}}$, we may apply (9) with η replaced by η^*:

$$\mathrm{Tr}(\eta^*\ \log \eta^*) - \mathrm{Tr}(\eta^*\ \log \rho) = \mathrm{Tr}(\Lambda^*\, \eta_{\mathrm{V}}) + \log Z - \log \mathcal{Z}(\Lambda^*). \tag{10}$$

Equating (9) and (10),

$$\mathrm{Tr}(\eta\ \log \eta^*) - \mathrm{Tr}(\eta\ \log \rho) = \mathrm{Tr}(\eta^*\ \log \eta^*) - \mathrm{Tr}(\eta^*\ \log \rho). \tag{11}$$

Adding $\mathrm{Tr}(\eta \log \eta) - \mathrm{Tr}(\eta \log \eta^*)$ to both sides of (11) and rearranging gives, by definition of Umegaki relative entropy,

$$\mathrm{D}_{\mathrm{U}}(\eta, \rho) = \mathrm{D}_{\mathrm{U}}(\eta, \eta^*) + \mathrm{D}_{\mathrm{U}}(\eta^*, \rho). \tag{12}$$

Subtracting $\mathrm{D}_{\mathrm{U}}(\eta_{\mathrm{V}}, \rho_{\mathrm{V}})$ from both sides yields (4.4). □

Recovering the Petz Recovery Map We first prove a lemma for V-locality.

Lemma B.5 (Visible block structure under V-locality)**.** *Let $\rho = e^{\mathrm{H}}/Z$ be a faithful Gibbs state on $\mathcal{H}_{\mathrm{V}} \otimes \mathcal{H}_{\mathrm{L}}$, with Hamiltonian*

$$\mathrm{H} = \mathrm{H}_V \otimes \mathrm{I}_{\mathrm{L}} + \mathrm{I}_{\mathrm{V}} \otimes \mathrm{H}_L + \mathrm{H}_{VL}$$

that is V-local in the sense of Definition 3.3. Let $\mathrm{H}_V = \sum_k h_k \Pi_k$ be its spectral decomposition over distinct eigenvalues, where Π_k projects onto the corresponding eigenspace V_k. Then:

(a) $[\mathrm{H}_V \otimes \mathrm{I}_{\mathrm{L}}, \rho] = 0$ *and* $[\mathrm{H}_V, \rho_{\mathrm{V}}] = 0$.

(b) The states ρ and ρ_{V} are block-diagonal with respect to the eigenspace decomposition $\mathcal{H}_{\mathrm{V}} = \bigoplus_k V_k$:

$$\rho = \sum_k (\Pi_k \otimes \mathrm{I}_{\mathrm{L}}) \rho (\Pi_k \otimes \mathrm{I}_{\mathrm{L}}), \qquad \rho_{\mathrm{V}} = \sum_k \Pi_k \rho_{\mathrm{V}} \Pi_k.$$

The projectors Π_k need not have rank one.

Proof. By V-locality, $[\mathrm{H}_V \otimes \mathrm{I}_{\mathrm{L}}, \mathrm{H}_{VL}] = 0$. Moreover, $\mathrm{H}_V \otimes \mathrm{I}_{\mathrm{L}}$ commutes with $\mathrm{I}_{\mathrm{V}} \otimes \mathrm{H}_L$ by disjoint tensor supports. Consequently,

$$[\mathrm{H}_V \otimes \mathrm{I}_{\mathrm{L}}, \mathrm{H}] = 0.$$

Functional calculus therefore gives $[\mathrm{H}_V \otimes \mathrm{I}_{\mathrm{L}}, \rho] = 0$. Taking the partial trace yields

$$[\mathrm{H}_V, \rho_{\mathrm{V}}] = \mathrm{Tr}_{\mathrm{L}}[\mathrm{H}_V \otimes \mathrm{I}_{\mathrm{L}}, \rho] = 0,$$

proving (a). Each spectral projector Π_k is a polynomial in H_V, so

$$[\Pi_k \otimes \mathrm{I}_{\mathrm{L}}, \rho] = 0, \qquad [\Pi_k, \rho_{\mathrm{V}}] = 0.$$

Inserting $\sum_k \Pi_k = \mathrm{I}_{\mathrm{V}}$ on both sides of each state, orthogonality of the spectral projectors eliminates all blocks with distinct indices. This gives the decompositions in (b). □

Proof of Corollary 4.5. Set

$$K := \rho_{\mathrm{V}}^{-1/2} \eta_{\mathrm{V}} \rho_{\mathrm{V}}^{-1/2}, \qquad \mathcal{R} := \mathcal{R}_{\rho, \mathrm{Tr}_{\mathrm{L}}}, \qquad \omega := \mathcal{R}(\eta_{\mathrm{V}}).$$

By the assumed commutation relation,

$$\omega = \rho^{1/2} (K \otimes \mathrm{I}_{\mathrm{L}}) \rho^{1/2} = \rho (K \otimes \mathrm{I}_{\mathrm{L}}).$$

Partial tracing $[\rho, K \otimes \mathrm{I}_{\mathrm{L}}] = 0$ gives $[\rho_{\mathrm{V}}, K] = 0$. Consequently,

$$\mathrm{Tr}_{\mathrm{L}} \omega = \rho_{\mathrm{V}} K = \rho_{\mathrm{V}}^{1/2} K \rho_{\mathrm{V}}^{1/2} = \eta_{\mathrm{V}}.$$

Thus ω is a feasible extension of η_{V}. Since ρ is faithful, $\mathcal{R}$ is CPTP and $\mathcal{R}(\rho_{\mathrm{V}}) = \rho$. Applying the data processing inequality first to $\mathcal{R}$ and then to the partial trace yields

$$\begin{aligned} \mathrm{D_U}(\eta_{\mathrm{V}}, \rho_{\mathrm{V}}) &\geq \mathrm{D_U}(\omega, \rho) \\ &\geq \mathrm{D_U}(\mathrm{Tr}_{\mathrm{L}} \omega, \rho_{\mathrm{V}}) \\ &= \mathrm{D_U}(\eta_{\mathrm{V}}, \rho_{\mathrm{V}}). \end{aligned}$$

Hence ω attains the data-processing lower bound on the QIP objective. By uniqueness of the QIP optimum,

$$\eta^*(\rho, \eta_{\mathrm{V}}) = \mathcal{R}_{\rho, \mathrm{Tr}_{\mathrm{L}}}(\eta_{\mathrm{V}}), \qquad \mathrm{D_U}(\eta^*, \rho) = \mathrm{D_U}(\eta_{\mathrm{V}}, \rho_{\mathrm{V}}).$$

□

B.3 Proofs from Section 5

Proof of Proposition 5.1. The expression for $\mathcal{Q}(\theta;\theta^{(\text{old})})$ is given by

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

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

$$\frac{\partial}{\partial\theta_r}\text{Tr}(\eta^{(\text{old})}\sum_r \theta_r \text{H}_r) = \text{Tr}(\eta^{(\text{old})}\text{H}_r), \text{ and}$$
$$\frac{\partial}{\partial\theta_r}\text{Tr}(\eta^{(\text{old})}\log Z(\theta)\text{I}) = \frac{\partial}{\partial\theta_r}\log\text{Tr}\exp(\text{H}(\theta)).$$

By Duhamel's formula,

$$\begin{aligned}\frac{\partial}{\partial\theta_r}\log\text{Tr}\exp(\text{H}(\theta)) &= \frac{1}{\text{Tr}\exp(\text{H}(\theta))}\text{Tr}\left(\int_0^1 \text{d}t e^{\text{H}(\theta)t}\text{H}_r e^{\text{H}(\theta)(1-t)}\right),\\ &= \frac{1}{\text{Tr}\exp(\text{H}(\theta))}\text{Tr}(e^{\text{H}(\theta)}\text{H}_r),\\ &= \text{Tr}(\rho(\theta)\text{H}_r).\end{aligned}$$

Collecting together the two partial derivatives, we obtain

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

which is the required result. □

Proof of Likelihood ascent We first state a useful fact.

Proposition B.6 (Ostrowski [79])**.** *Let $\{x_t\}_{t\geq 0}$ be a bounded sequence in a finite-dimensional normed space with $\|x_{t+1} - x_t\| \to 0$. The set of subsequential limits of $\{x_t\}$ is non-empty, compact, and connected.*

Proof of Theorem 5.2. **(a) Gradient identity and fixed-point characterization.** Let $\mathcal{S} := \text{supp}(\eta_{\text{v}})$, $P_{\mathcal{S}}$ its projector, and $\text{H}_{\mathcal{S}}(\theta) := (P_{\mathcal{S}} \otimes \text{I}_L)\text{H}(\theta)(P_{\mathcal{S}} \otimes \text{I}_L)$. By Theorem 4.3, $\eta^*(\theta) = e^{\text{H}_{\mathcal{S}}(\theta)+\Lambda^*(\theta)\otimes\text{I}_L}/\mathcal{Z}(\theta)$ on $\mathcal{S}\otimes\mathcal{H}_L$, where $\Lambda^*(\theta) \in \mathcal{T}(\mathcal{S})$ is the unique traceless Hermitian multiplier with $\text{Tr}_L\eta^*(\theta) = \eta_{\text{v}}$. The implicit function theorem applied to this constraint shows that $\eta^*(\theta)$ is C^1 in θ.

Differentiate $\tilde{\mathcal{L}_{\text{U}}}(\theta) = Q(\theta \mid \theta) + \text{S}_{\text{VN}}(\eta^*(\theta)) - \text{S}_{\text{VN}}(\eta_{\text{v}})$ via the chain rule. Writing $F(\theta_1,\theta_2) := Q(\theta_1 \mid \theta_2)$,

$$\nabla_\theta\tilde{\mathcal{L}_{\text{U}}} \;=\; \partial_{\theta_1}F\big|_{\theta_1=\theta_2=\theta} \;+\; \partial_{\theta_2}F\big|_{\theta_1=\theta_2=\theta} \;+\; \nabla_\theta\text{S}_{\text{VN}}(\eta^*(\theta)).$$

The first term, computed using $\partial_\theta\log\text{Tr}e^{X(\theta)} = \text{Tr}(\sigma\,\partial_\theta X)$ with $X = \text{H}(\theta_1)$, gives $\partial_{\theta_1,r}F\big|_{\text{diag}} = \text{Tr}(\eta^*(\theta)\text{H}_r) - \text{Tr}(\rho(\theta)\text{H}_r) = g_r(\theta)$.

For the remaining two terms we collect the contributions of $\partial_\theta\eta^*$. Using $\text{Tr}\partial_\theta\eta^* = \partial_\theta\text{Tr}\eta^* = 0$ and $\text{Tr}(\partial_\theta\eta^*\cdot\text{H}) = \text{Tr}(\partial_\theta\eta^*\cdot\text{H}_{\mathcal{S}})$ (support reduction),

$$\partial_{\theta_2,r}F\big|_{\text{diag}} = \text{Tr}\big(\partial_{\theta_r}\eta^*(\theta)\cdot\log\rho(\theta)\big) = \text{Tr}\big(\partial_{\theta_r}\eta^*\cdot\text{H}_{\mathcal{S}}\big).$$

For the entropy gradient, $\nabla_\theta\text{S}_{\text{VN}}(\eta^*) = -\text{Tr}(\partial_\theta\eta^*\cdot\log\eta^*)$, and

$$\text{Tr}(\partial_\theta\eta^*\cdot\log\eta^*) = \text{Tr}(\partial_\theta\eta^*\cdot\text{H}_{\mathcal{S}})+\text{Tr}(\partial_\theta\eta^*\cdot\Lambda^*\otimes\text{I}_L) = \text{Tr}(\partial_\theta\eta^*\cdot\text{H}_{\mathcal{S}})+\text{Tr}(\partial_\theta\eta_{\text{v}}\cdot\Lambda^*) = \text{Tr}(\partial_\theta\eta^*\cdot\text{H}_{\mathcal{S}}),$$

the last equality from $\partial_\theta\eta_{\text{v}} = 0$. Hence $\partial_{\theta_2}F|_{\text{diag}} + \nabla_\theta\text{S}_{\text{VN}}(\eta^*)$ contributes $\text{Tr}(\partial_\theta\eta^*\cdot\text{H}_{\mathcal{S}}) - \text{Tr}(\partial_\theta\eta^*\cdot\text{H}_{\mathcal{S}}) = 0$, and $\nabla\tilde{\mathcal{L}_{\text{U}}} = g$.

The map $\theta' \mapsto Q(\theta' \mid \theta) = \sum_r \theta'_r\text{Tr}(\eta^*(\theta)\text{H}_r) - \log Z(\theta')$ is concave for every θ since $\log Z$ is convex. By first-order conditions for concave functions, $\theta' \in M(\theta) \iff \nabla_{\theta'}Q(\theta' \mid \theta) = 0$. Setting $\theta' = \theta$, $\theta \in M(\theta) \iff g(\theta) = 0 \iff \nabla\tilde{\mathcal{L}_{\text{U}}}(\theta) = 0 \iff \theta \in \tilde{\text{S}}$.

(b) Monotone $\tilde{\mathcal{L}}_{\mathrm{U}}$-ascent and limit set. The QELBO inequality applied at $\eta = \eta^*(\theta^{(t)})$ with target $\theta^{(t+1)}$ gives

$$\tilde{\mathcal{L}}_{\mathrm{U}}(\theta^{(t+1)}) \;\geq\; \mathrm{QELBO}\big(\eta^*(\theta^{(t)}), \theta^{(t+1)}\big) \;=\; Q(\theta^{(t+1)} \mid \theta^{(t)}) + \mathrm{S}_{\mathrm{VN}}(\eta^*(\theta^{(t)})) - \mathrm{S}_{\mathrm{VN}}(\eta_{\mathrm{V}}),$$

since $\eta^*(\theta^{(t+1)})$ optimises QELBO at $\theta^{(t+1)}$. The M-step ascent $Q(\theta^{(t+1)} \mid \theta^{(t)}) \geq Q(\theta^{(t)} \mid \theta^{(t)})$ then yields

$$\tilde{\mathcal{L}}_{\mathrm{U}}(\theta^{(t+1)}) - \tilde{\mathcal{L}}_{\mathrm{U}}(\theta^{(t)}) \;\geq\; Q(\theta^{(t+1)} \mid \theta^{(t)}) - Q(\theta^{(t)} \mid \theta^{(t)}) \;\geq\; 0.$$

The function $\tilde{\mathcal{L}}_{\mathrm{U}}$ is continuous on Θ (from (a)), hence bounded on the compact K. Therefore $\tilde{\mathcal{L}}_{\mathrm{U}}(\theta^{(t)})$ is monotone non-decreasing and bounded above, and converges to some real value, which we denote $\tilde{\mathcal{L}}_{\mathrm{U}}{}^*$.

By compactness of K the set of subsequential limit points of $\{\theta^{(t)}\}$ is non-empty and compact. Let $\theta^\dagger$ be any limit point, with $\theta^{(t_j)} \to \theta^\dagger$. By compactness of K pass to a sub-subsequence along which $\theta^{(t_{j_k}+1)} \to \theta'$ for some $\theta' \in K$.

Closed graph of M [67]. For every $\xi \in \Theta$, $\theta^{(t_{j_k}+1)} \in M(\theta^{(t_{j_k})})$ gives $Q(\theta^{(t_{j_k}+1)} \mid \theta^{(t_{j_k})}) \geq Q(\xi \mid \theta^{(t_{j_k})})$. Joint continuity of Q in both arguments lets us send $k \to \infty$ on both sides, yielding $Q(\theta' \mid \theta^\dagger) \geq Q(\xi \mid \theta^\dagger)$ for every ξ, i.e. $\theta' \in M(\theta^\dagger)$.

Limit-value identification. Continuity of $\tilde{\mathcal{L}}_{\mathrm{U}}$ and the monotone convergence $\tilde{\mathcal{L}}_{\mathrm{U}}(\theta^{(t)}) \to \tilde{\mathcal{L}}_{\mathrm{U}}{}^*$ give $\tilde{\mathcal{L}}_{\mathrm{U}}(\theta^\dagger) = \tilde{\mathcal{L}}_{\mathrm{U}}{}^* = \tilde{\mathcal{L}}_{\mathrm{U}}(\theta')$. Apply QELBO at $\eta = \eta^*(\theta^\dagger)$ with target θ':

$$0 \;=\; \tilde{\mathcal{L}}_{\mathrm{U}}(\theta') - \tilde{\mathcal{L}}_{\mathrm{U}}(\theta^\dagger) \;\geq\; Q(\theta' \mid \theta^\dagger) - Q(\theta^\dagger \mid \theta^\dagger).$$

Combined with $Q(\theta' \mid \theta^\dagger) \geq Q(\theta^\dagger \mid \theta^\dagger)$ from $\theta' \in M(\theta^\dagger)$, equality holds, so $\theta^\dagger$ also achieves the maximum: $\theta^\dagger \in M(\theta^\dagger)$. By (a), $\theta^\dagger \in \tilde{\mathrm{S}}$.

Every limit point therefore lies in $\tilde{\mathrm{S}} \cap K$ and takes the value $\tilde{\mathcal{L}}_{\mathrm{U}}{}^*$, and any single limit point $\theta^\dagger$ exhibits $\tilde{\mathcal{L}}_{\mathrm{U}}{}^* = \tilde{\mathcal{L}}_{\mathrm{U}}(\theta^\dagger)$ as a stationary value of $\tilde{\mathcal{L}}_{\mathrm{U}}$. This proves (b).

(c) Approach to the stationary set. The iterates approach the stationary set:

$$\mathrm{dist}\big(\theta^{(t)}, \tilde{\mathrm{S}} \cap K\big) \longrightarrow 0.$$

Suppose otherwise. Then some subsequence remains at distance at least $\delta > 0$ from $\tilde{\mathrm{S}} \cap K$. By compactness of K, it has a further subsequence converging to a point $\bar{\theta} \in K$. By the proof of part (b), $\bar{\theta} \in \tilde{\mathrm{S}}$. Along this subsequence,

$$\mathrm{dist}\big(\theta^{(t_j)}, \tilde{\mathrm{S}} \cap K\big) \leq \|\theta^{(t_j)} - \bar{\theta}\| \longrightarrow 0,$$

a contradiction. □

Proof of Corollary 5.3. Define the nonnegative, differentiable defect

$$\varepsilon(\theta) := \mathcal{L}_{\mathrm{U}}(\eta_{\mathrm{V}}, \theta) - \tilde{\mathcal{L}}_{\mathrm{U}}(\theta) \geq 0.$$

By Corollary 4.5, alignment gives $\varepsilon(\theta^{(t)}) = 0$ at every iterate. Continuity implies that it also vanishes at every limit point. Since a differentiable nonnegative function has zero gradient wherever it vanishes, the proxy and likelihood have equal gradients at these points. The likelihood values equal the proxy values along the trajectory, so monotonicity and convergence follow from Theorem 5.2. Every limit point is therefore stationary for the likelihood. Compactness of K then gives the stated distance-to-set conclusion. □

B.4 Proof from Section 6

Proof of Theorem 6.3. For a data point $\mathbf{v}^{(i)}$, set $p_i = P(\mathbf{v}^{(i)} \mid \theta)$ and $\tau_i = \rho_{\mathrm{L}}^{(i)}(\theta)$. Then

$$\begin{aligned}\mathrm{Tr}(\eta_{\mathrm{L}} \log(p_i \tau_i)) + \mathrm{S}_{\mathrm{VN}}(\eta_{\mathrm{L}}) &= \log p_i - \mathrm{D}_{\mathrm{U}}(\eta_{\mathrm{L}}, \tau_i) \\ &\leq \log p_i,\end{aligned}$$

with equality exactly at $\eta_{\mathrm{L}} = \tau_i$. Averaging gives the stated bound. Writing the CQ model and target in their common data basis as $\rho = \sum_j p_j \Pi_j \otimes \tau_j$ and $\eta_{\mathrm{V}} = \sum_j q_j \Pi_j$, the Petz formula gives

$$\mathcal{R}_{\rho,\mathrm{Tr}_{\mathrm{L}}}(\eta_{\mathrm{V}}) = \rho^{1/2} \left(\sum_j \frac{q_j}{p_j} \Pi_j \otimes \mathrm{I}_{\mathrm{L}} \right) \rho^{1/2} = \sum_j q_j \Pi_j \otimes \tau_j,$$

which is the required recovery expression. □

Proof of Corollary 6.4. Proof is immediate by matching the terms in the $\sigma_i^{(z)}$ terms. □

Proof of Proposition 6.5. Fix a visible Pauli-Z configuration $\mathrm{x} \in \{-1, +1\}^m$. Its visible-only contribution to the Hamiltonian is scalar on the hidden space and cancels when the hidden state is normalized. The remaining Hamiltonian is

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

where

$$h_j(\mathrm{x}, \theta) = -\mathbf{b}_j^{\mathrm{eff}} \sigma^{(z)} - \Gamma_j \sigma^{(x)}, \qquad \mathbf{b}_j^{\mathrm{eff}} = \mathbf{b}_j + \sum_{i=1}^{m} \mathbf{W}_{ij} \mathrm{x}_i.$$

The summands act on disjoint hidden qubits and therefore commute. Consequently,

$$e^{\mathrm{H}_{\mathrm{L}}(\mathrm{x},\theta)} = \bigotimes_{j=1}^{n} e^{h_j(\mathrm{x},\theta)}, \qquad \mathcal{Z}_{\mathrm{L}}(\mathrm{x}, \theta) = \prod_{j=1}^{n} \mathrm{Tr} e^{h_j(\mathrm{x},\theta)}.$$

Normalizing gives

$$\rho_{\mathrm{L}}(\mathrm{x} \mid \theta) = \bigotimes_{j=1}^{n} \frac{e^{h_j(\mathrm{x},\theta)}}{\mathrm{Tr} e^{h_j(\mathrm{x},\theta)}}.$$

This is the required product decomposition. □

C Boltzmann machines and Contrastive Divergence

This section discusses variants of Boltzmann Machines that played a key role in the development of modern deep learning [80–82], and the contrastive divergence algorithm for learning them.

C.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 [31]. 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

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

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

Restricted Boltzmann Machines (RBM) were introduced with no intra-layer connections due to the intractability of learning in fully connected BMs [32, 33, 83]. 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 $\mathrm{z} \in \{0,1\}^n$ [84, 85]. It is additionally parameterized by a variance term $\mathbf{s} \in \mathbb{R}^m_{>0}$. The energy and partition functions are given by

$$\begin{aligned} E_\theta(\mathrm{x},\mathrm{z}) &= -\frac{1}{2}\sum_{i=1}^{m}\frac{(\mathrm{x}_i-\mathbf{a}_i)^2}{s_i} - \sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{W}_{ij}\frac{\mathrm{x}_i}{s_i}\mathrm{z}_j - \sum_{j=1}^{n}\mathbf{b}_j\mathrm{z}_j, \\ \mathcal{Z}(\theta) &= \int_{-\infty}^{\infty}\sum_{\mathrm{z}}\exp(E_\theta(\mathrm{x},\mathrm{z}))\mathrm{dx}. \end{aligned} \tag{GRBM}$$

Deep Boltzmann Machines (DBM) stack RBMs using undirected connections and allow for joint training of all layers [34]. 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.

C.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 [33, 83]. 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

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

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

$$\begin{aligned} \mathrm{z}_j(t) \sim p_\theta(\mathrm{z}_j = 1 \mid \mathrm{x}(t)) &= \sigma\left(\sum_i \mathbf{W}_{ij}\mathrm{x}(t)_i + \mathbf{b}_j\right), \\ \mathrm{x}(t+1)_i \sim p_\theta(\mathrm{x}_i = 1 \mid \mathrm{z}(t)) &= \sigma\left(\sum_j \mathbf{W}_{ij}\mathrm{z}(t)_j + \mathbf{a}_i\right), \\ \mathbb{E}_{p_\theta(\mathrm{Z}|\mathrm{X}=\mathrm{x})}(\mathrm{x}\mathrm{z}^\top) &\approx \mathrm{x}\mathrm{z}(0)^\top, \\ \mathbb{E}_{p_\theta(\mathrm{X},\mathrm{Z})}(\mathrm{x}\mathrm{z}^\top) &\approx \mathrm{x}(k)\mathrm{z}(k)^\top, \end{aligned}$$

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 [34, 35]. In a GRBM, the conditional probabilities for the layers are given by

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

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

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.

C.3 Novel QBM Variants

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

Quantum Interleaved DBM A **Q**uantum **I**nterleaved **D**eep **B**oltzmann **M**achine (QiDBM) is a DBM 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.

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

The Gibbs sampling scheme described in Equation (2) can be extended to QiDBMs.

Quantum Gaussian–Bernoulli RBM The **Q**uantum **G**aussian–Bernoulli **R**estricted **B**oltzmann **M**achine (QGRBM) extends the GRBM by introducing transverse fields on the hidden units. For continuous visible observations $\mathrm{x} \in \mathbb{R}^m$, the CQ model is described on the algebra $L^\infty(\mathbb{R}^m, \mathrm{d}\mathrm{x}) \bar{\otimes} \mathcal{T}(\mathcal{H}_{\mathrm{L}})$, where $\mathcal{H}_{\mathrm{L}}$ is finite-dimensional. The visible variable is classical, so the model carries no coherence between distinct visible values.

For a fixed visible value x, the Hamiltonian on $\mathcal{H}_{\mathrm{L}}$ is

$$\begin{aligned} \mathrm{H}(\mathrm{x}, \theta) = &-\sum_{i=1}^{m} \frac{(\mathrm{x}_i - \mathbf{a}_i)^2}{s_i^2} \mathrm{I}_{\mathrm{L}} \\ &-\sum_{j=1}^{n} \mathbf{b}_j^{\mathrm{eff}}(\mathrm{x}) \sigma_j^{(z)} - \sum_{j=1}^{n} \Gamma_j \sigma_j^{(x)}, \end{aligned} \tag{QGRBM}$$

where

$$\mathbf{b}_j^{\mathrm{eff}}(\mathrm{x}) = \mathbf{b}_j + \sum_{i=1}^{m} \mathbf{W}_{ij} \frac{\mathrm{x}_i}{s_i}.$$

Here $s_i > 0$ is a scale parameter; with this convention, the uncoupled visible variance is $s_i^2/2$.

The joint model is represented by the positive operator-valued density

$$\varrho_\theta(\mathrm{x}) = \frac{\exp(\mathrm{H}(\mathrm{x}, \theta))}{\mathcal{Z}(\theta)}, \qquad \mathcal{Z}(\theta) = \int_{\mathbb{R}^m} \mathrm{Tr} \exp(\mathrm{H}(\mathrm{x}, \theta))\, \mathrm{d}\mathrm{x},$$

with $0 < \mathcal{Z}(\theta) < \infty$ and $\int_{\mathbb{R}^m} \mathrm{Tr}\varrho_\theta(\mathrm{x})\, \mathrm{d}\mathrm{x} = 1$. The visible probability density and clamped hidden state are

$$p_\theta(\mathrm{x}) = \mathrm{Tr}\varrho_\theta(\mathrm{x}), \qquad \tau_\theta(\mathrm{x}) = \frac{\exp(\mathrm{H}(\mathrm{x}, \theta))}{\mathrm{Tr}\exp(\mathrm{H}(\mathrm{x}, \theta))}.$$

When the visible value is fixed to x, the hidden state factorizes into single-qubit states. The Hamiltonian of the jth hidden qubit is

$$\mathrm{H_L}(j \mid \mathrm{x}, \theta) = -\mathbf{b}_j^{\mathrm{eff}}(\mathrm{x})\sigma^{(z)} - \Gamma_j \sigma^{(x)}.$$

Its Pauli expectations can therefore be evaluated efficiently as in Equation (2).

For an observation $\mathrm{x}^{(i)}$ and any hidden density operator τ_i, we have

$$\log p_\theta(\mathrm{x}^{(i)}) = \mathrm{Tr}\Big(\tau_i \mathrm{H}(\mathrm{x}^{(i)}, \theta)\Big) + \mathrm{S_{VN}}(\tau_i) - \log \mathcal{Z}(\theta) + \mathrm{D_U}\Big(\tau_i, \tau_\theta(\mathrm{x}^{(i)})\Big).$$

Nonnegativity of relative entropy gives a variational lower bound, with equality exactly when $\tau_i = \tau_\theta(\mathrm{x}^{(i)})$. Consequently, exact E-steps and M-steps that increase the averaged bound yield nondecreasing empirical log-likelihood. This argument does not establish parameter convergence or guarantees for heuristic sampling-based updates.

Similar to the QRBM, sampling over these qubits remains efficient, allowing CD to approximate the gradients. The visible units can be drawn from the Gaussian distribution similar to a GRBM, using samples obtained in the Pauli-Z basis.

D Additional Experimental Details

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. In this section, we provide additional details related to the experimental results presented in the main body. Details of the computing infrastructure used for all experiments are provided in Appendix E.

(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 1 scale to standard ML data sets such as MNIST?

(Q3) **Performance of DO-EM.** Do CQ-LVMs trained with Algorithm 1 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 the literature. Details of the computing infrastructure used for all experiments are provided in Appendix E.

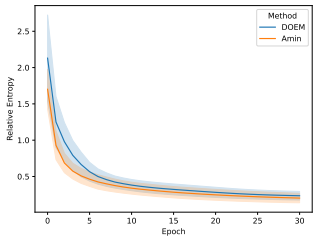
D.1 DO-EM vs Gradient-based Maximization of Projective Log-Likelihood

We conduct experiments using *exact log-likelihood computation* to compare DO-EM with gradient-based maximization of the projective log-likelihood. A QBM that is a CQ-LVM 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. [8].

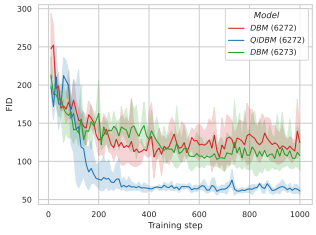
The QBMs are trained on a mixture of Bernoulli data set introduced in Amin et al. [8]. A Bernoulli distribution is built on a randomly selected mode and the average probability distribution over M such modes determines the mixture of Bernoulli distribution. The distribution can be written as

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

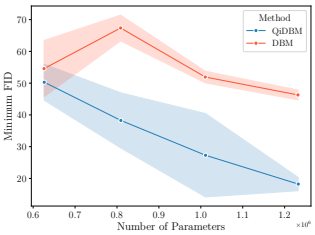
where m is the dimension of data, p is the probability of success, and d_v^k is the Hamming distance between v and the randomly selected mode. We use 100 randomly sampled data sets of 1000 samples each with $N = 1000$, $M = 8$, $m = 8$, and $p = 0.9$. A model with $m = 8$ visible units and $n = 2$ hidden 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.



(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 Projective log-likelihood gradient-based training (PGT) with exact computation on mixture of Bernoulli data set. (b,c) QiDBM vs DBM on MNIST with CD.

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 1 to train the QBM. As shown in Figure 1a, the model trained with DO-EM *converges in roughly half the number of steps* 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.

D.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. The theoretical results proved in this work do not directly apply to Contrastive Divergence, but provide a tractable practical technique for experimentation. We compare our method with Taniguchi et al. [36], 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[3] with the hyperparameters recommended in their paper. *No additional hyperparameter tuning* is performed on the quantum variant.

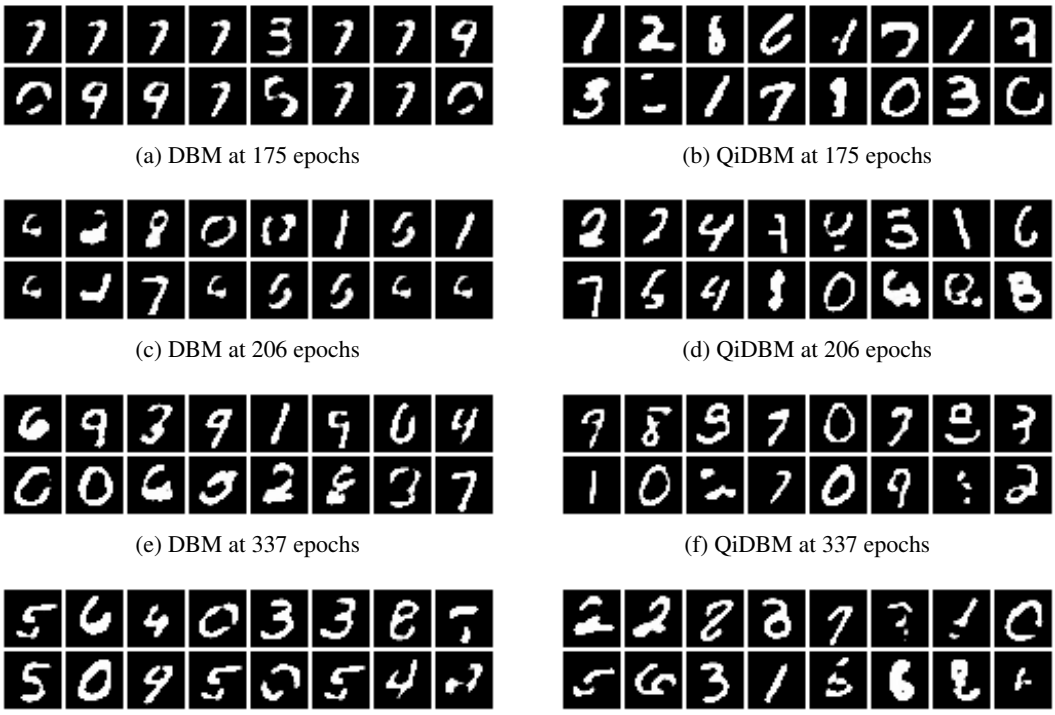
Following Taniguchi et al. [36], we use a $\{0, 1\}$ encoding instead of $\{-1, +1\}$ to perform experiments on MNIST and Binarized MNIST [86]. The data sets contain 60,000 training and 10,000 testing images of size 28×28. We evaluate the quality of generated images using the Fréchet Inception Distance [FID, 87], computed between 10,000 generated samples and the MNIST test set.

D.2.1 MNIST

To demonstrate the scalability of CQ-LVMs on standard ML data sets under the DO-EM scheme, we train *QiDBMs* on the MNIST dataset. Following Taniguchi et al. [36], 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.

The FID values of generated samples, recorded every 10 epochs across 3 training runs with seeds 0, 1, and 2, 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.

[3] `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) and QiDBM (right).

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 *QiDBMs* to data sets such as MNIST is feasible and leads to *substantial performance gains*.

D.2.2 Binarized MNIST

To further assess the impact of the additional quantum bias terms, we compare *QiDBM* 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*.

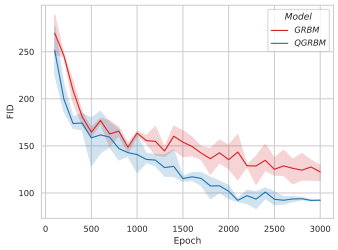
D.3 CQ-LVMs on Continuous Data

To assess the applicability of the DO-EM algorithm to continuous data, we apply Algorithm 1 and Contrastive Divergence to Quantum Gaussian-Bernoulli RBMs (*QGRBM*). We compare our method with Liao et al. [85]. We use results obtained from their official implementation[4] with the hyperparameters recommended in their paper. *No additional hyperparameter tuning* is performed on the quantum variant.

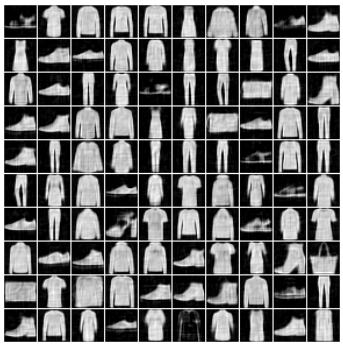
Following Liao et al. [85], we use a $\{0, 1\}$ encoding in the hidden units instead of $\{-1, +1\}$ to perform experiments on Fashion MNIST [88] and CelebA [89]. The Fashion MNIST data set

[4] `https://github.com/lrjconan/GRBM`

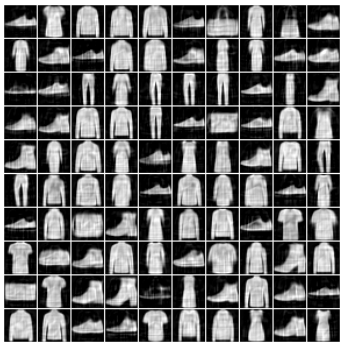
contains 60,000 training and 10,000 testing images of size 28×28. The CelebA-32 data set contains 162,770 training and 19,962 testing images downsampled to size 32×32. We assess the quality of the generated images using the Fréchet Inception Distance [FID, 87], computed between generated samples and 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

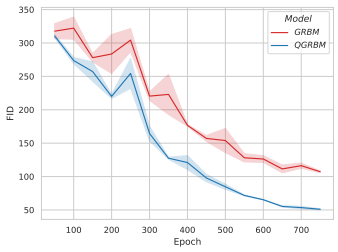


(c) Samples generated by the GRBM at Epoch 3000

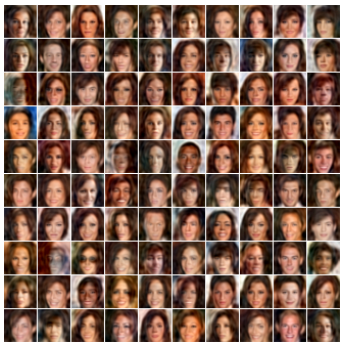
Figure 3: Fashion MNIST

D.3.1 Fashion MNIST

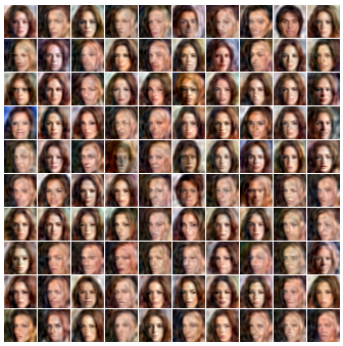
We train a *QGRBM* with 10,000 hidden units on the Fashion-MNIST dataset and compare it to a GRBM trained with Contrastive Divergence under the same conditions following Liao et al. [85]. Training is carried out for 3000 epochs using stochastic gradient with Contrastive Divergence, a batch size of 512, and a learning rate of 0.01. The FID values of generated samples across 3 random seeds (1234,1235,1236) are plotted in Figure 3a. The *QGRBM* achieves *a 20% reduction in FID*. 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 750



(c) Samples generated by the GRBM at Epoch 750

Figure 4: CelebA-32

D.3.2 CelebA-32

We train a *QGRBM* on the CelebA-32 dataset and compare it to a GRBM trained with Contrastive Divergence under the same conditions with 10,000 hidden units, following Liao et al. [85]. Training is carried out for 750 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 from 2 training runs, with seed 1235 and 1236, are plotted in Figure 4a. The *QGRBM* achieves around a *50% 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.

Table 1: Effect of the fraction of middle-layer units with a transverse field on sample quality. FID is averaged over 3 seeds (lower is better).

% transverse-field units	FID ↓
0% (DBM with diagonal density operator)	42.61
50%	17.62
100% (Figure 1c)	**14.77**

D.4 Effectiveness of transverse field parameter

To measure the effect of introducing the transverse field parameter γ in our experiments, we consider an ablation in which we set a fixed percentage of parameters to zero. We additionally trained a 3-layer QiDBM with (784, 784, 784) hidden units on the Binarized MNIST dataset, with only 50% of the middle layer having a transverse field term, using the same hyperparameters as in Figure 1c. This is an intermediate setting between the DBM and QiDBM with 784 hidden units. We report the average minimum FID across 100 epochs over 3 random seeds in Table 1. This ablation shows that the FID gain is a function of the number of hidden units with transverse fields. Whether this is an improvement in raw representation capacity or due to the optimization dynamics of the geometry of PSD operators is an important question for future work.

D.5 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.

E Compute Platform

Implementation Details We implement our proposed methods in PyTorch 2.5.1.

Hardware Tables 2 and 3 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 2. Machine 2 runs Ubuntu 22.04.1 LTS with kernel 5.15.0-127-generic with the hardware in Table 3. Our software stack comprises of 3.12.8, PyTorch 2.5.1, torchvision version 0.20.1.

Table 2: 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 3: Machine 2: Specifications of GPU hardware used for computation

CPU Model Name	Intel(R) Core(TM) i9-10980XE
CPU(s)	36
Thread(s) per core	2
Core(s) per socket	18
Socket(s)	1
NUMA node(s)	1
CPU MHz(Max)	4800.0000
L1d & L1i cache	576 KiB
L2 cache	18 MiB
L3 cache	24.8 MiB
RAM	125GiB
GPU Model name	NVIDIA GeForce RTX 3090
GPU(s)	2
Dedicated Memory Size(per GPU)	24 GB
CUDA Version	12.2

F Limitations

This work only addresses finite dimensional Hilbert spaces, analogues of discrete probability spaces. While most of the results in this work generalize to the continuous case, we leave their analysis for future work. Although the DO EM algorithm is amenable to implementation on a quantum computer using Gibbs state preparation its precise design and analysis needs further research.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

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

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Limitations are dicussed in the Appendix.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: The paper provides a full set of assumptions and complete proofs of all stated theorems.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: All code is provided in an anonymous github link.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: All details are provided in the anonymous github link.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Full details are provided in the main body, appendix, and code provided in the anonymous github.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: We provide all results with 2-sigma error bars.

   Guidelines:

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

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Details are provided in the Appendix.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research conforms to the code of ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: This paper discusses theoretical aspects of expectation maximization for density operator latent variable models. We do not believe there is any societal impact in this work.

    Guidelines:

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

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: As a theoretical paper on density operator latent variable models, we believe the paper poses no such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All original owners or creators are credited.

    Guidelines:

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

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

13. **New assets**

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

    Answer: [Yes]

    Justification: Code with details are provided in the anonymous github link provided.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.