Density Operator Expectation Maximization

International Workshop on Quantum Boltzmann Machines 2025

Adit Vishnu [1] Abhay Shastry [1] Dhruva Kashyap [1] Chiranjib Bhattacharyya [1]

[1]Machine Learning Lab, Computer Science and Automation
Indian Institute of Science, Bangalore

Agenda

Probabilistic Latent Variable Models

Density Operator Latent Variable Models

Generating images with QBMs

Agenda

Probabilistic Latent Variable Models

Density Operator Latent Variable Models

Generating images with QBMs

Boltzmann Machines

Boltzmann Machines (BM) are stochastic neural networks that define a probability distribution over binary vectors based on the Ising model [Ackley et al., 1985]

A BM consists of a visible layer $\mathrm{x} \in \{0,1\}^m$ and a hidden layer $\mathrm{z} \in \{0,1\}^n$

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

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

$\mathcal{Z}(\theta)$ makes it difficult to train BMs
BMs are restricted to very small datasets

Restricted Boltzmann Machines

A solution was to restrain the connections in the model [Smolensky, 1986]
The Restricted Boltzmann Machine (RBM) is a BM with a bipartite connection graph

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

RBMs were used for a variety of ML tasks prior to modern DNNs
[Larochelle and Bengio, 2008, Salakhutdinov et al., 2007, Hinton and Salakhutdinov, 2006]
RBMs can scale to image data sets such as MNIST

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

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

Contrastive Divergence

RBM layers are conditionally independent

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

CD-k algorithm [Hinton, 2002, Carreira-Perpiñán and Hinton, 2005]

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

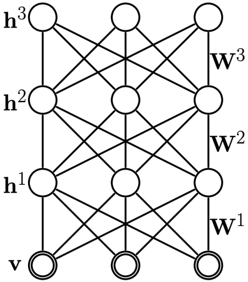
$$\mathrm{x}(t+1)_i \sim p_\theta(\mathrm{x}_i = 1 \mid \mathrm{z}(t)) = \sigma\left(\sum_j \mathbf{W}_{ij}\mathrm{z}(t)_j + \mathbf{a}_i\right)$$

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

Deep Boltzmann Machines

A layered RBM was introduced to capture richer patterns [Salakhutdinov and Hinton, 2009, 2012]

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



Gaussian-Bernoulli RBMs

Gaussian-Bernoulli RBMs (GRBM) extend RBMs to model continuous visible units $\mathrm{x} \in \mathbb{R}^m$ and discrete hidden units [Welling et al., 2004]

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

Contrastive divergence extends to GRBMs

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

Latent Variable Models

BMs are instances of latent variable models (LVM) [Bishop, 2006]

$$\Pr(\mathrm{X}{=}\mathrm{x} \mid \theta) = \sum_{\mathrm{z}} \Pr(X{=}\mathrm{x}, \mathrm{Z}{=}\mathrm{z} \mid \theta)$$

LVMs are the backbone of VAEs and other modern generative models [Kingma and Welling, 2014, Ho et al., 2020]

For a dataset $\mathcal{D} = \{\mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(N)}\}$, the log-likelihood of an LVM is

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

Gradients are usually hard to evaluate due to marginalization

Expectation Maximization Algorithm

The Evidence Lower Bound for the log-likelihood

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

EM algorithm is a two step iterative solution [Baum and Petrie, 1966, Dempster et al., 1977]

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

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

Comments on the EM algorithm

The EM algorithm guarantees monotonic log-likelihood

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

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

ELBO can be seen as a consequence of Shanon's data processing inequality [Shannon, 1948]

The EM algorithm has an information geometric interpretation [Amari, 1995]

Takeaways

Variants of BMs are often more useful in practice

LVMs are very useful in unsupervised learning

Probabilistic LVMs are hard to train; EM algorithm is the answer

Agenda

Probabilistic Latent Variable Models

Density Operator Latent Variable Models

Generating images with QBMs

Quantum Boltzmann Machines

Hamiltonian-based models

$$\mathrm{H}(\theta) = \sum_r \theta_r \mathrm{H}_r \text{ and } \mathcal{Z}(\theta) = \mathrm{Tr}\exp(\mathrm{H}(\theta))$$

$$\rho(\theta) = \frac{\exp(\mathrm{H}(\theta))}{\mathcal{Z}(\theta)} \quad \text{and} \quad \rho_{\mathrm{V}}(\theta) = \mathrm{Tr}_{\mathrm{L}}\rho(\theta)$$

QBM Hamiltonian based on the transverse field Ising model [Amin et al., 2018]

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

Projective Log-likelihood

An objective function based on projective measurements

[Amin et al., 2018, Anschuetz and Cao, 2019, Zoufal et al., 2021, Demidik et al., 2025]

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

Talks already discussed: gradients are hard because of the projection operators

Quantum Log-likelihood

An objective based on the relative entropy between density operators
[Kieferová and Wiebe, 2017, Wiebe and Wossnig, 2019, Kappen, 2020]

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

Talks already discussed: gradients are hard because of the partial trace

Density Operators

Definition (Density Operator)

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

The KL divergence can be extended to density operators [Cover and Thomas, 2006, Umegaki, 1962]

Definition (Umegaki Relative Entropy)

Let ω and ρ be density operators in $\mathcal{P}(\mathcal{H})$ with $\ker(\rho) \subseteq \ker(\omega)$. Their relative entropy is

$$\mathrm{D_U}(\omega, \rho) = \mathrm{Tr}(\omega \log \omega) - \mathrm{Tr}(\omega \log \rho).$$

No perfect analog of conditional probability
Proofs from probabilistic LVMs breakdown due to non-commutativity

Petz Recovery Map

Theorem (Monotonicity of Relative Entropy)

For density operators ω *and* ρ *in* $\mathcal{P}(\mathcal{H})$ *such that* $\ker(\omega) \subset \ker(\rho)$, $\mathrm{D}_{\mathrm{U}}(\omega, \rho) \geq \mathrm{D}_{\mathrm{U}}(\mathcal{N}(\omega), \mathcal{N}(\rho))$.

Petz [1986, 1988] proved conditions for when MRE is saturated

Theorem (Petz Recovery Map)

For states ω *and* ρ *in* $\mathcal{P}(\mathcal{H}_A)$ *and a CPTP map* $\mathcal{N} : \mathcal{P}(\mathcal{H}_A) \to \mathcal{P}(\mathcal{H}_B)$,

$$\mathrm{D}_{\mathrm{U}}(\omega, \rho) = \mathrm{D}_{\mathrm{U}}(\mathcal{N}(\omega), \mathcal{N}(\rho))$$

if and only if there exists a CPTP map $\mathcal{R}$ *such that* $\mathcal{R}(\mathcal{N}(\omega)) = \omega$ *and* $\mathcal{R}(\mathcal{N}(\rho)) = \rho$. *Furthermore, on the support of* $\mathcal{N}(\rho)$, $\mathcal{R}$ *is explicity given by the Petz recovery map*

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

Projective Measurement

Definition (Projective Measurement)

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

Assume that data is coming from projective measurements of some ground truth density operator. The empirical target operator is then

$$\eta_{\mathrm{V}} = \frac{1}{N} \sum_{i=1}^{N} \Lambda(\mathbf{v}^{(i)}).$$

Density Operator LVMs

Definition

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

Lemma

For a data set $\mathcal{D} = \{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(N)}\}$ arising out of projective measurements, let the empirical data density operator be η_{V}. Then for a DO-LVM $\rho(\theta)$,

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

Evidence Lower Bound

Lemma (Quantum ELBO)

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

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

By the monotonicity of relative entropy [Lindblad, 1975]

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

Expanding the expression for Umegaki relative entropy and rearranging

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

Deriving DO-EM

The classical EM algorithm is a consequence of the evidence lower bound being a minorant of the log-likelihood.

Monotonicity of relative entropy is often not saturated for the partial trace operation [Lesniewski and Ruskai, 1999, Berta et al., 2015, Wilde, 2015, Carlen and Vershynina, 2020, Cree and Sorce, 2022].

Appeal to the information geometric interpretation of EM.

Quantum Information Projection

Definition (Quantum Information Projection)

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

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

Definition (Sufficient Conditions)

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

$\mathrm{Tr_L}(\rho)$ is faithful

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

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

Quantum Information Projection

Definition (Quantum Information Projection)

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

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

Theorem

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

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

DO-EM

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

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

Algorithm Density Operator Expectation Maximization

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

Properties of DO-EM

DO-EM guarantees log-likelihood ascent under the Sufficient Conditions
DO-EM reduces to the classical EM algorithm if the operators are diagonal
DO-EM solves the problem of computing gradients!

For a Hamiltonian-based model

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

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

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

Classical Data

Theorem (CQ-LVM)

A DO-LVM that satisfies the Sufficient Conditions for a classical dataset generated by the measurement $\mathcal{X} = \sum_{i=1}^{d_V} \mathrm{x}_i \, \Lambda(\mathbf{u}_i)$ *if it can be expressed as*

$$\rho(\theta) = \sum_{i=1}^{d_V} \Pr(\mathrm{X}{=}\mathrm{x}_i|\theta)\Lambda(\mathbf{u}_i) \otimes \rho_{\mathrm{L}}(i|\theta) \quad \textit{where} \quad \rho_{\mathrm{L}}(i|\theta) \in \mathcal{P}(\mathcal{H}_{\mathrm{L}}). \qquad \text{(CQ-LVM)}$$

Algorithm DO-EM for CQ-LVM[1]

1: **Input:** $\mathcal{D} = \{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(N)}\}$ and $\theta^{(0)}$

2: **while** not converged **do**

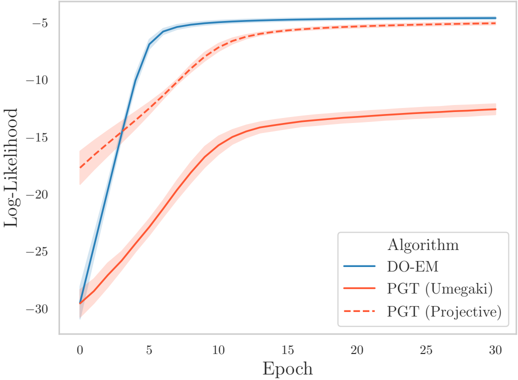
3: **for** $i = 1$ to N **do**

4: $\mathcal{Q}_i(\theta; \theta^{(t)}) = \mathrm{Tr}\left(\rho_{\mathrm{L}}(\mathrm{x}^{(i)} \mid \theta^{(t)}) \log P(\mathrm{x}^{(i)} \mid \theta)\rho_{\mathrm{L}}(\mathrm{x}^{(i)} \mid \theta)\right) + \mathrm{S}_{\mathrm{VN}}(\rho_{\mathrm{L}}(\mathrm{x}^{(i)} \mid \theta^{(t)}))$

5: $\theta^{(t+1)} = \mathrm{argmax}_\theta \frac{1}{N} \sum_{i=1}^{N} \mathcal{Q}_i(\theta; \theta^{(t)})$

[1] Hayashi's em-algorithm for sq-QBMs is a special case of this theorem, as sq-QBMs are instances of CQ-LVMs

QBMs on Mixture of Bernoulli datasets



8+2 SR QBM with projective log-likelihood gradient-based training (PGT) (40s/step)[2]
8+2 QBM satisfying the Sufficient Conditions trained using DO-EM (0.2s/step)

[2] Amin et al. [2018]

Agenda

Probabilistic Latent Variable Models

Density Operator Latent Variable Models

Generating images with QBMs

Contrastive Divergence

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

Conditioned on visible layer, Hamiltonian of each hidden qubit is

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

Leading to Gibbs sampling scheme[3]

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

[3] details in paper; suffice to see one $2^n \times 2^n$ matrix becones n tractable 2×2 matrices

QiDBM

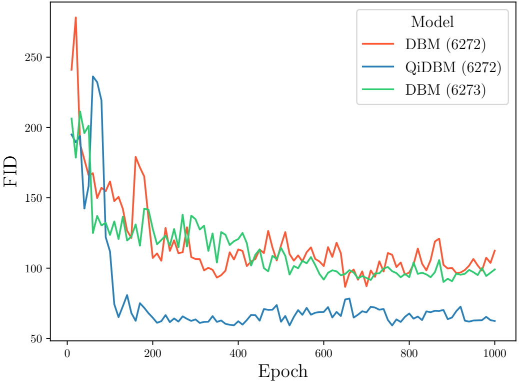
Quantum interleaved Deep Boltzmann Machine

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

Conducive to CD like QRBMs

Compare generated samples using the Fréchet Inception Distance [Seitzer, 2020].

QiDBMs on MNIST



QiDBM with 18,000 units compared against DBM in Taniguchi et al. [2023].

QiDBMs on MNIST

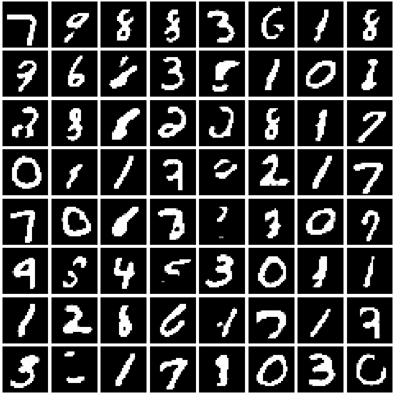


Figure: QiDBMs after 175 epochs

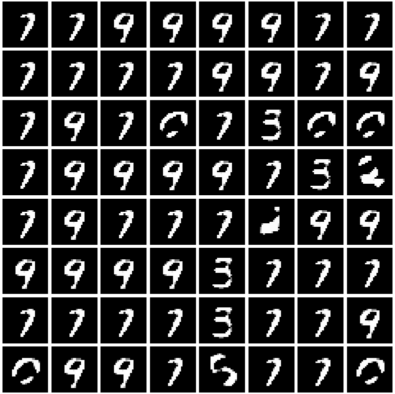
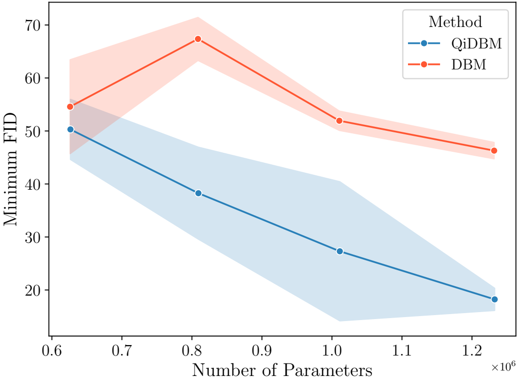


Figure: DBMs after 175 epochs

QiDBMs on Binary MNIST



QGRBM

Quantum Gaussian-Bernoulli Restricted Boltzmann Machine

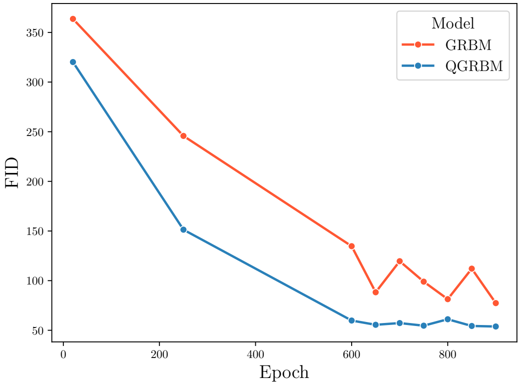
$$\mathrm{H}(\mathrm{x},\theta) = -\sum_{i=1}^{m} \frac{(\mathrm{x}_i - \mathbf{a}_i)^2}{\mathbf{s}_i^2} \mathrm{I}_{\mathrm{L}} - \sum_{j=1}^{n} \mathbf{b}_j^{\mathrm{eff}} \sigma_j^{(z)} - \sum_{j=1}^{n} \Gamma_j \sigma_j^{(x)}, \qquad \text{(QGRBM)}$$

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

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

Infinite dimensional density operator but can still do CD!

QiDBMs on MNIST



QGRBM with 11,000 units compared against GRBM in Liao et al. [2022].

QGRBMs on CelebA-32

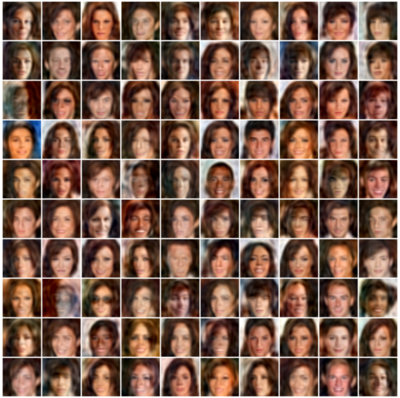


Figure: QGRBM samples

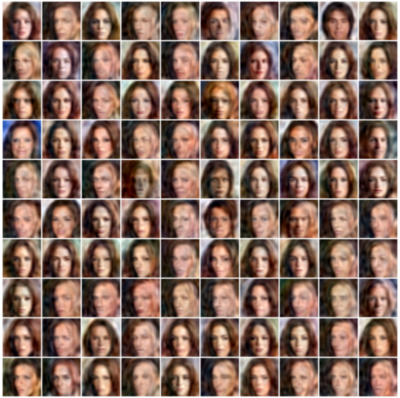


Figure: GRBM samples

Final Thoughts

No hyperparameter tuning on quantum models and similar computational resources

CQ-LVMs may be useful for classical data too

DO-EM is ready for any DO-LVM

Paper and Code

arXiv:2507.22786 (long version out today!)
All code will be documented and released by January 1st, 2026.

References I

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

References II

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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