An em algorithm for quantum Boltzmann machines

arXiv:2507.21569

Takeshi Kimura

Ph. D. Student, Department of Mathematical Informatics, Graduate School of Informatics,

Nagoya University

Joint work with Kohtaro Kato and Masahito Hayashi

Background: Quantum Machine Learning

Goal:
Achieve advantages over classical ML by exploiting quantum resources

Typical approaches:
VQE [Peruzzo+2014]: variational quantum circuits for optimization
QBM [Amin+2018]: quantum extension of Boltzmann machines

Our Perspective: Hybrid Approach

- Quantum model and classical training
    - **Parametrized quantum models** are highly expressive, describing complex quantum states
    - **Classical optimization** updates parameters using measurement outcomes
- Advantage: combine quantum expressivity with classical efficiency

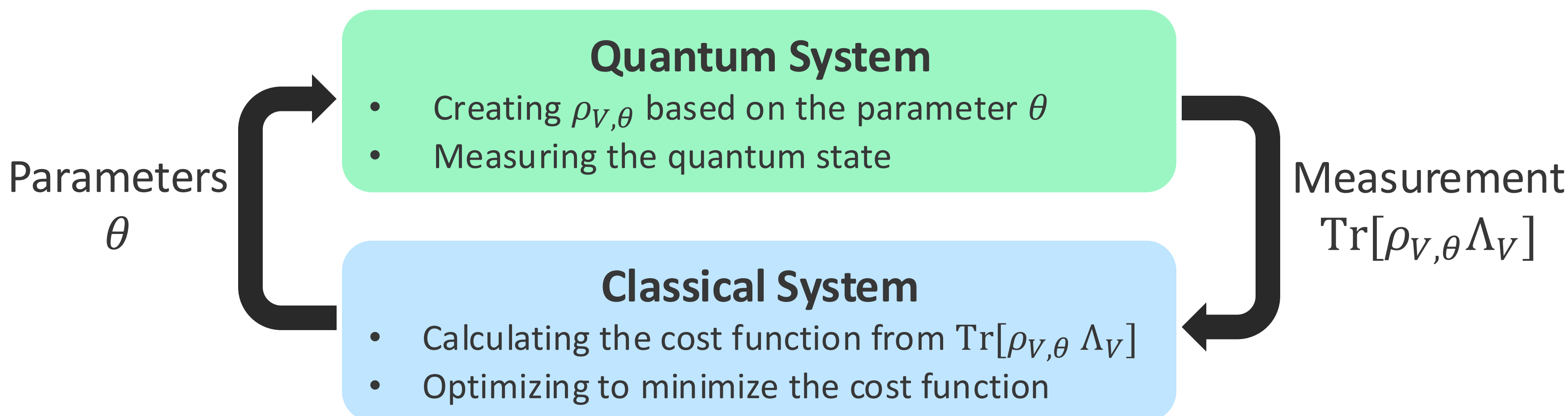
Quantum model and classical training

<u>Quantum model and classical training</u>

1. Creating a quantum state $\rho_{V,\theta}$ based on the parameter θ
2. Measuring $\rho_{V,\theta}$ in computational basis

Classical simulation

3. Calculating the cost function (ex. KL divergence) from the measurement results $\mathrm{Tr}[\rho_{V,\theta}\Lambda_V]$
4. Back to 1 until to minimize the cost function (ex. GD method)



Gradient descent method (GD)

- Conventional method to train parameterized models

Objective

P_V: given data dist.
$P_{V,\theta}$: model dist.

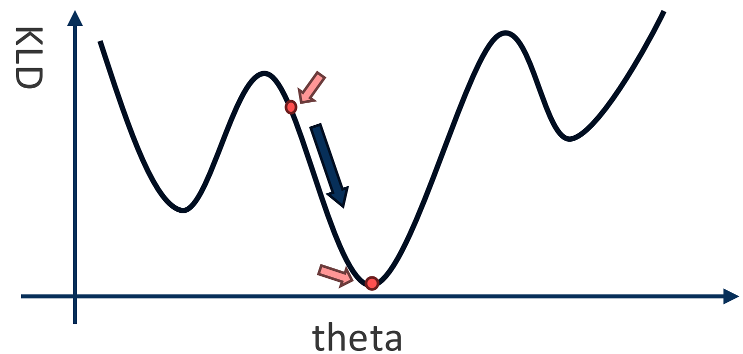
- To minimize KL divergence

$$D_{\mathrm{KL}}(P_V \| P_{V,\theta}) := \sum_{\mathbf{v}} P_V(\mathbf{v}) \left(\log P_V(\mathbf{v}) - \log P_{V,\theta}(\mathbf{v})\right)$$

Parameter update rule

- To use gradient of KL divergence

$$\theta \leftarrow \theta - \eta \cdot \partial_\theta D_{\mathrm{KL}}$$



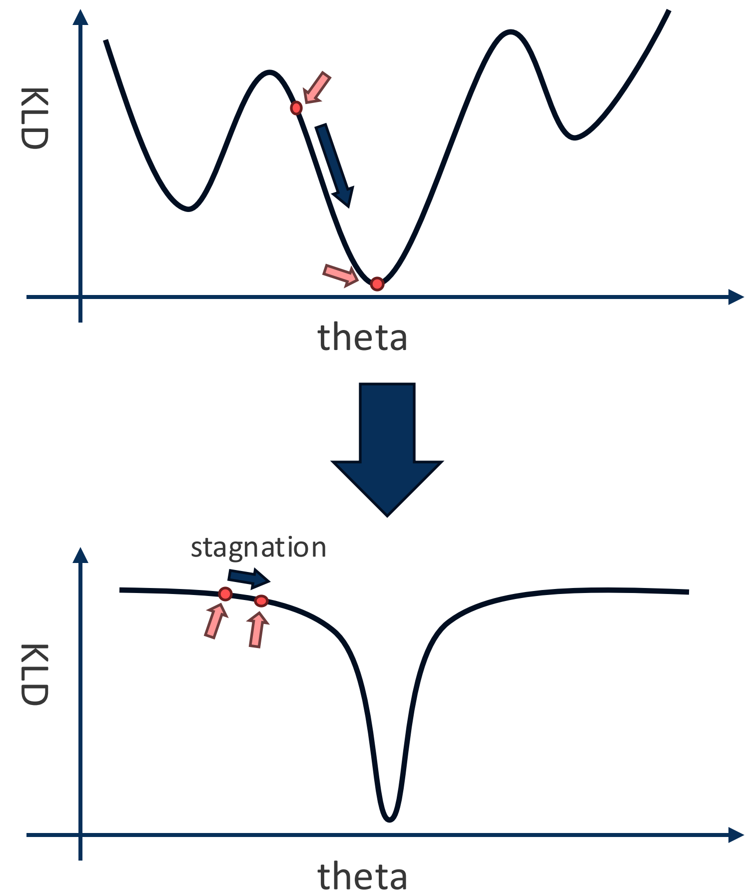
Issues

η: learning rate

- Weak convergence guarantees
  - Sensitive to initialization & learning rate
  - In the case of non-convex functions, convergence to the global optimum is not guaranteed.
- **Susceptible to vanishing gradients (Barren Plateau)**

Barren plateau

- The **vanishing gradient phenomenon** that occurs far from a local minima
- **The gradient variance decreases exponentially** due to the following factors:
  - Deep quantum circuits[McClean+2018]
  - Multi-qubit[McClean+2018]
  - Entanglement[Marrero+2021]
  - Global measurement cost functions[Wang+2021]
- Gradient methods that use gradients for learning will be greatly affected



Typical obstacle of non-convex optimization

Our Proposal: The em Algorithm

GD method

$$\theta \leftarrow \theta + \eta \cdot \partial_\theta D_{\mathrm{KL}}$$

- Since gradients are used for non-convex functions, it is susceptible to the vanishing gradient problem.

Quantum em algorithm (ours)

- We propose to use the *em algorithm* [Amari+1992] instead of the GD method.
- Iteratively performs e-step and m-step
- A mathematical generalization of EM algorithm [Dempster+1977]
- The gradient method is only used for m-step, that is **convex**
- Potential to avoid the Barren Plateau problem

Boltzmann machine (BM) [Ackley+1985]

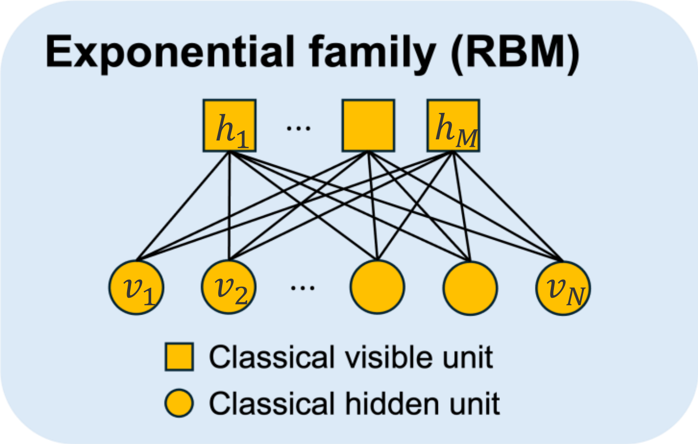
- Energy-based probabilistic generative model defined on an undirected graph.

Components

In quantum case, qubit

- Visible layer: $V = \{v_1, \ldots, v_N\} \quad (v_i = \pm 1, i \in [1, N])$
- Hidden layer: $H = \{h_1, \ldots, h_M\}$
- Parameters:
  - Coupling strength between v_i and h_j: $w_{ij} \in \mathbb{R}$
  - Bias strength: $b_i \in \mathbb{R}$

In this talk, we consider restricted BM definition



Restricted Boltzmann machine (RBM)

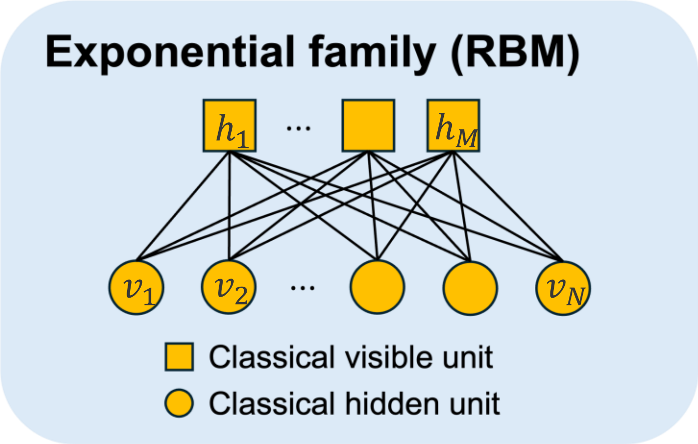
- Energy function of RBM (no connections between visible layers or between hidden units.)

$$E(\mathbf{v},\mathbf{h}) = -\sum_{v_i \in V} b_i v_i - \sum_{h_j \in H} b_j h_j - \sum_{(i,j)\in E} w_{ij} v_i h_j$$

- Probability distribution

$$P_{VH,\theta}(\mathbf{v},\mathbf{h}) := e^{-E(\mathbf{v},\mathbf{h})}/Z, \quad Z = \sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}$$

$$P_{V,\theta}(\mathbf{v}) := \sum_{\mathbf{h}\in H} P_{VH,\theta}(\mathbf{v},\mathbf{h}),$$



Restricted Quantum Boltzmann machine (RQBM)

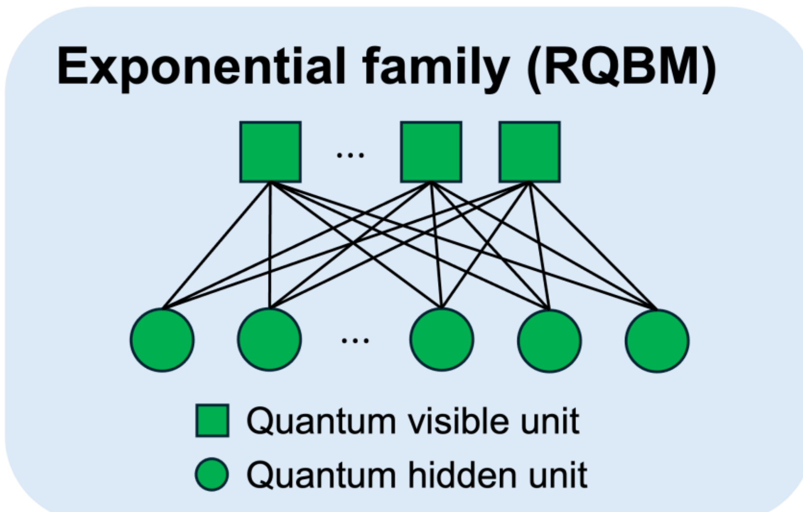
- Hamiltonian

$$H = -\sum_{i \in V} (b_i \sigma_i^z + \Gamma_i \sigma_i^x) - \sum_{j \in H} (b_j \sigma_j^z + \Gamma_j \sigma_j^x) - \sum_{(i,j) \in E} w_{ij} \sigma_i^z \sigma_j^z$$

- Probability distribution

$$P_{V,\theta}(\mathbf{v}) = \mathrm{Tr}[\Lambda_{\mathbf{v}} \rho_{V,\theta}] \quad \Lambda_{\mathbf{v}} = |\mathbf{v}\rangle \langle \mathbf{v}|$$

$$\rho_{VH,\theta} := e^{-H}/Z, \quad Z = \mathrm{Tr}[e^{-H}] \;\; \rho_{V,\theta} = \mathrm{Tr}_H \rho_{VH,\theta}$$



Unsupervised learning in Boltzmann machine

Objective

Fitting model distribution $P_{V,\theta}$ to data distribution P_V.

$$P_{V,\theta} = \mathrm{Tr}[\rho_{V,\theta}\Lambda_V] \; \rho_{V,\theta} = \mathrm{Tr}_H \rho_{VH,\theta}$$

$$\rho_{VH,\theta} := e^{-H}/Z, \quad Z = \mathrm{Tr}[e^{-H}]$$

Training:

Updating the parameter θ to make the distribution of $P_{V,\theta}$ and P_V closer

Evaluation

Using KL divergence between $P_{V,\theta}$ and P_V

$$D_{\mathrm{KL}}(P_V \| P_{V,\theta}) := \sum_{\mathbf{v}} P_V(\mathbf{v}) \left(\log P_V(\mathbf{v}) - \log P_{V,\theta}(\mathbf{v})\right)$$

EM (Expectation Maximization) algorithm

- directly intervenes in the structure of the hidden units and explicitly optimizes them

<u>Objective</u>

- To minimize KL divergence:

$$D_{\mathrm{KL}}(P_V \times P_{H|V} \| P_{VH,\theta}) = \sum_{\mathbf{v}} P_V(\mathbf{v}) D_{\mathrm{KL}}(P_{H|V=\mathbf{v}} \| P_{H|V=\mathbf{v},\theta}) + D_{\mathrm{KL}}(P_V \| P_{V,\theta})$$

<u>Algorithm</u>

- Alternates two steps:
  - E-step: infer hidden variables

    $$P_{VH,\theta} \rightarrow P_{H|V}(h) = P_{H|V,\theta}(h)$$

  - M-step: maximize expected log-likelihood

    $$\theta = \underset{\theta}{\mathrm{argmax}}\, \mathbb{E}_{H \sim P_{H|V=v}}[\log P_{VH,\theta}(v, H)]$$

<u>Benefits</u>

- Convex M-step

EM (Expectation Maximization) algorithm

- directly intervenes in the structure of the hidden units and explicitly optimizes them

<u>Objective</u>

- To minimize KL divergence:

$$D_{\mathrm{KL}}(P_V \times P_{H|V} \| P_{VH,\theta}) = \sum_{\mathbf{v}} P_V(\mathbf{v}) D_{\mathrm{KL}}(P_{H|V=\mathbf{v}} \| P_{H|V=\mathbf{v},\theta}) + D_{\mathrm{KL}}(P_V \| P_{V,\theta})$$

<u>Algorithm</u>

- Alternates two steps:
  - E-step: infer hidden variables

$$P_{VH,\theta} \to P_{H|V}(h) = P_{H|V,\theta}(h)$$

Application to QBM is not easy

  - M-step: maximize expected log-likelihood

$$\theta = \underset{\theta}{\operatorname{argmax}}\, \mathbb{E}_{H \sim P_{H|V=v}}[\log P_{VH,\theta}(v, H)]$$

<u>Benefits</u>

- Convex M-step

- An information geometric reformulation of EM algorithm

<u>Definition</u>

- **Exponential family** $\mathcal{E}$ for a random variable $X = \{x\}$ is a set of probability distributions $p(x;\boldsymbol{\theta})$ given by the exponential form:

$$p(X;\boldsymbol{\theta}) = \exp\left\{\sum_{i=1}^{n} \theta_i r_i(x) + k(x) - \psi(\theta)\right\},$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$ is an n-dimensional vector parameter,

$\{r_i(x)\}_{i=1}^{n}, k(x)$ are functions of x and ψ is a normalization factor.

- **Mixture family** $\mathcal{M}$ is a set of is a set of distributions $q(x)$ formed by a probability mixture of m component distributions $\{q_i(x)\}_{i=1}^{m}$:

$$q(x) = \sum_{i=1}^{m} w_i q_i(x),$$

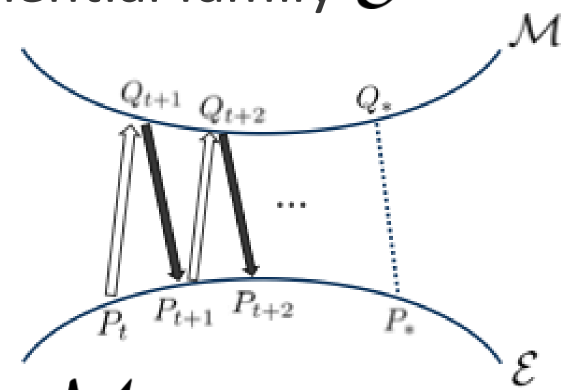
where, $\sum_{i=1}^{m} w_i = 1, \quad w_i \geq 0$

em algorithm

Objective

- To minimize KL divergence between an exponential family $\mathcal{E}$ and a mixture family $\mathcal{M}$

$$\min_{P \in \mathcal{M}, Q \in \mathcal{E}} D_{\mathrm{KL}}(P \| Q)$$



Algorithm

- Alternating projections:
  - e-step (e-projection): a projection of Q_t to $\mathcal{M}$

$$P_t = \operatorname*{argmin}_{P \in \mathcal{M}} D_{\mathrm{KL}}(P \| Q_t)$$

  - m-step (m-projection): a projection of P_t to $\mathcal{E}$

$$Q_{t+1} = \operatorname*{argmin}_{Q \in \mathcal{E}} D_{\mathrm{KL}}(P_t \| Q)$$

Benefits

- Guarantees monotonic decrease of KL divergence.

$$D_{\mathrm{KL}}(P_{t-1} \| Q_t) \geq D_{\mathrm{KL}}(P_t \| Q_t) \geq D_{\mathrm{KL}}(P_t \| Q_{t+1})$$

- Convexity of m-step

Quantum em algorithm

- A quantum expansion of an em algorithm

<u>Definition</u>

- **Exponential family** $\mathcal{E} = \{\rho_\theta \in \mathcal{S}(\mathcal{H})\}$

$$\rho_\theta = \exp(\log \rho + \sum_{i=1}^{k} \theta^i X_i - \phi(\theta))$$

where $\theta = (\theta^1, \ldots, \theta^k)$ is parameters and ϕ is a normalization factor.

- **Mixture family** $\mathcal{M}(\mathbf{a}) = \{\rho \in \mathcal{S}(\mathcal{H}) | \mathrm{Tr}\rho X_i = a_i, i = 1, \ldots, k\}$

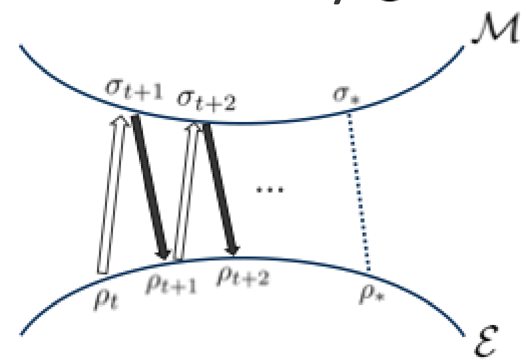
where $\mathcal{H}$ is Hilbert space and $\mathcal{S}(\mathcal{H})$ is set of densities over $\mathcal{H}$.
$X_1, \ldots, X_k$ is linearly independent observables on $\mathcal{H}$.
$\mathbf{a} = (a_1, \ldots, a_k)$ is measurement results.

Quantum em algorithm

<u>Objective</u>

- To minimize KL divergence between an exponential family $\mathcal{E}$ and a mixture family $\mathcal{M}$

$$\min_{\rho\in\mathcal{M},\sigma\in\mathcal{E}} D_{\mathrm{KL}}(\rho||\sigma)$$



<u>Algorithm</u>

- Alternating projections:
  - e-step (e-projection): a projection of σ_t to $\mathcal{M}$

$$\rho_t = \operatorname*{argmin}_{\rho\in\mathcal{M}} D_{\mathrm{KL}}(\rho||\sigma_t)$$

  - m-step (m-projection): a projection of ρ_t to $\mathcal{E}$

$$\sigma_{t+1} = \operatorname*{argmin}_{\sigma\in\mathcal{E}} D_{\mathrm{KL}}(\rho_t||\sigma)$$

Semi-quantum Restricted Boltzmann machine (sqRBM) [Demidik+2025]

- Hamiltonian

$$H = -\sum_{i\in V}(b_i\sigma_i^z + \cancel{\Gamma_i\sigma_i^x}) - \sum_{j\in H}(b_j\sigma_j^z + \Gamma_j\sigma_j^x) - \sum_{(i,j)\in E} w_{ij}\sigma_i^z\sigma_j^z$$

- Probability distribution

$$\rho_{VH,\theta} := e^{-H}/Z, \quad Z = \mathrm{Tr}[e^{-H}]$$

$$P_{V,\theta}(\mathbf{v}) = \mathrm{Tr}[\Lambda_{\mathbf{v}}\rho_{V,\theta}] \quad \Lambda_{\mathbf{v}} = |\mathbf{v}\rangle\langle\mathbf{v}|$$

$$\rho_{V,\theta} = \mathrm{Tr}_H\rho_{VH,\theta}$$

semi-quantum Boltzmann machine (sqRBM)

Advantage

- **Analytical tractability:**
  - Allows closed-form output probabilities & gradients
  - Efficient gradient estimation avoids costly QRBM training.
- **Practical benefits:**
  - Mitigates barren plateaus (no entanglement across visible–hidden cut).
  - Demonstrated strong performance across multiple datasets.

em algorithm for sqRBM

Definition

- **Exponential family** $\mathcal{E} = \{\rho_{VH,\theta} | \theta \in \mathbb{R}^{|\theta|}\}$
- **Mixture family** $\mathcal{M} = \{\rho_{VH} | \langle \mathbf{v}| \rho_V |\mathbf{v}\rangle = P_V(\mathbf{v}), \mathbf{v} \in V\}$

$$\rho_{VH} = P_V \times \rho_{H|V} := \sum_{\mathbf{v}} P_V(\mathbf{v}) |\mathbf{v}\rangle \langle \mathbf{v}| \otimes \rho_{H|V=\mathbf{v}} \qquad \rho_{H|V=\mathbf{v}} = \frac{\langle \mathbf{v}| \rho_{VH} |\mathbf{v}\rangle}{\langle \mathbf{v}| \rho_V |\mathbf{v}\rangle}$$

Objective

- To minimize KL divergence:

$$\min_{P_V \times \rho_{H|V} \in \mathcal{E}} \min_{\rho_{VH,\theta} \in \mathcal{M}} D(P_V \times \rho_{H|V} || \rho_{VH,\theta})$$

- KL divergence

$$D_{\mathrm{KL}}(P_V \times \rho_{H|V} \| \rho_{VH,\theta}) = \sum_{\mathbf{v}} P_V(\mathbf{v}) D_{\mathrm{KL}}(\rho_{H|V=\mathbf{v}} \| \rho_{H|V=\mathbf{v},\theta}) + D_{\mathrm{KL}}(P_V \| P_{V,\theta})$$

em algorithm for sqRBM

Algorithm 1 The em algorithm for sqRBM

Input Initial value of parameters $\theta^{(0)}$
Output Parameters θ

1: $\theta = \theta^{(0)}$
2: **for** $t = 0, 1, \ldots$ **do**
3: e-step:

$$\rho_{H|V}^{(t)} := \operatorname*{argmin}_{\rho_{H|V}} \sum_{\mathbf{v}} P_V(\mathbf{v}) D_{\mathrm{KL}}(\rho_{H|V=\mathbf{v}} \| \rho_{H|V=\mathbf{v},\theta^{(t)}})$$

Since D_{KL} ≥ 0, the minimum is achieved when

$$\rho_{H|V}^{(t)} = \rho_{H|V,\theta^{(t)}}$$

4: m-step:

convex optimization

$$\theta^{(t+1)} = \operatorname*{argmin}_{\theta} D_{\mathrm{KL}}(P_V \times \rho_{H|V,\theta^{(t)}} \| \rho_{VH,\theta})$$

$$\theta = \theta^{(t+1)}$$

Using GD method

$$\theta \leftarrow \theta + \eta\big(\partial_\theta Z + \mathrm{Tr}(P_V \times \rho_{H|V,\theta^{(t)}})\partial_\theta H\big)$$

5: End if convergence conditions are met
6: **end for**

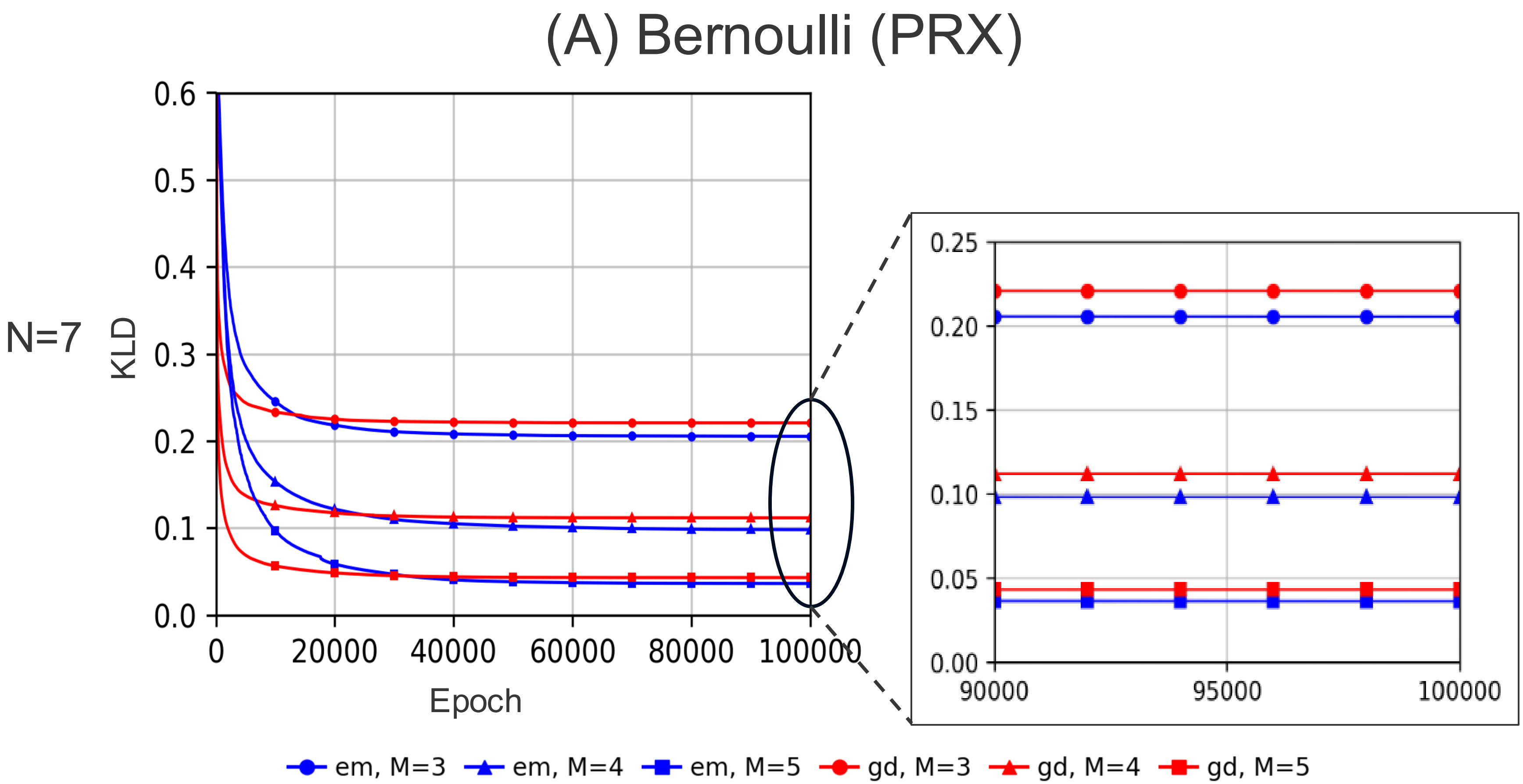
em algorithm for sqRBM

Mitigating the Barren Plateau Problem

- **From the Model Aspect (sqRBM):**
  - The hybrid structure (Classical Visible + Quantum Hidden) prevents entanglement between the visible and hidden layers.
  - This structurally avoids the exponential vanishing of gradients, as the gradient calculations remain localized.
- **From the Learning Method Aspect (em algorithm):**
  - The m-step is a **convex optimization problem**.
  - This guarantees a well-defined optimization path, allowing the model to escape flat landscapes and ensuring stable learning.

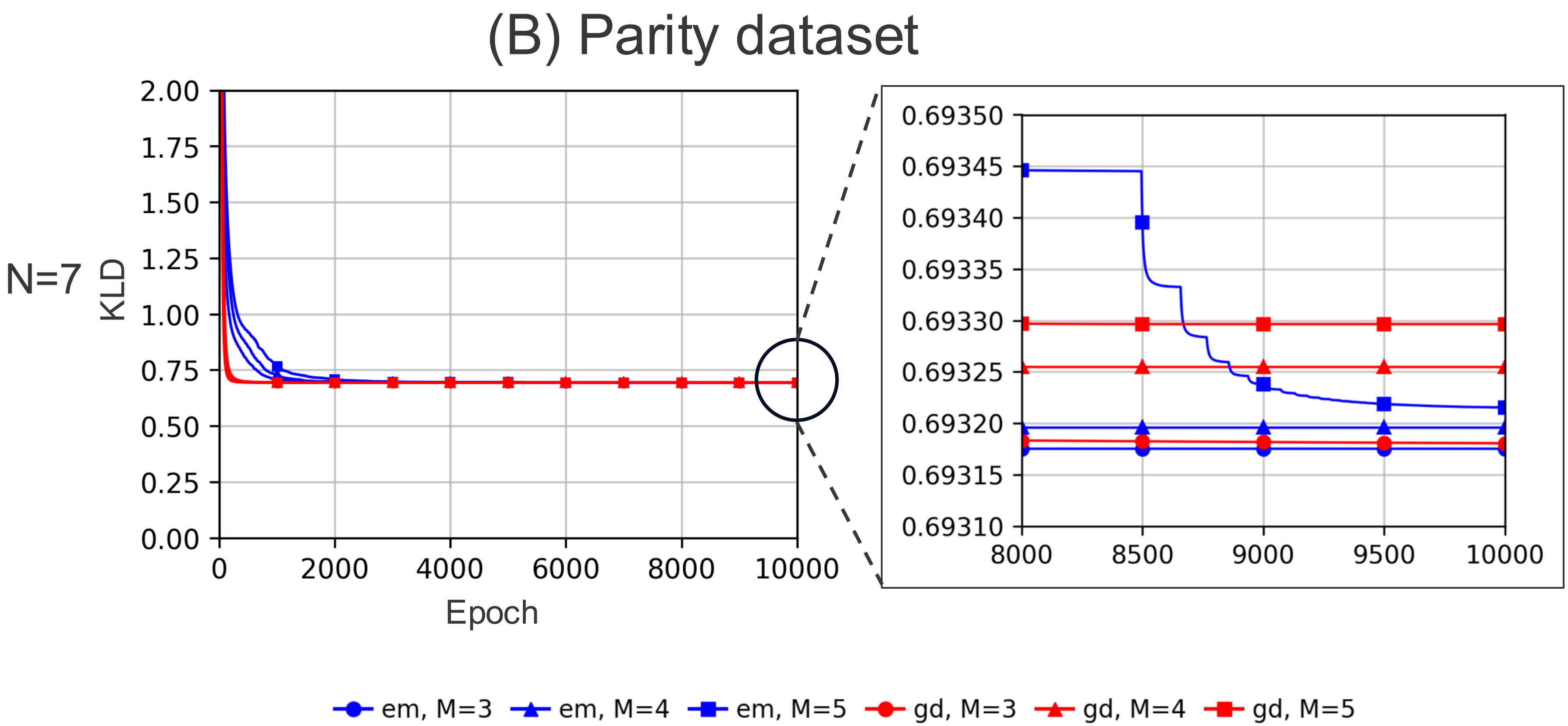
em algorithm vs GD (KL divergence)

The blue lines consistently achieve a lower final KLD than the red lines for different numbers of hidden units.



em algorithm vs GD (KL divergence)

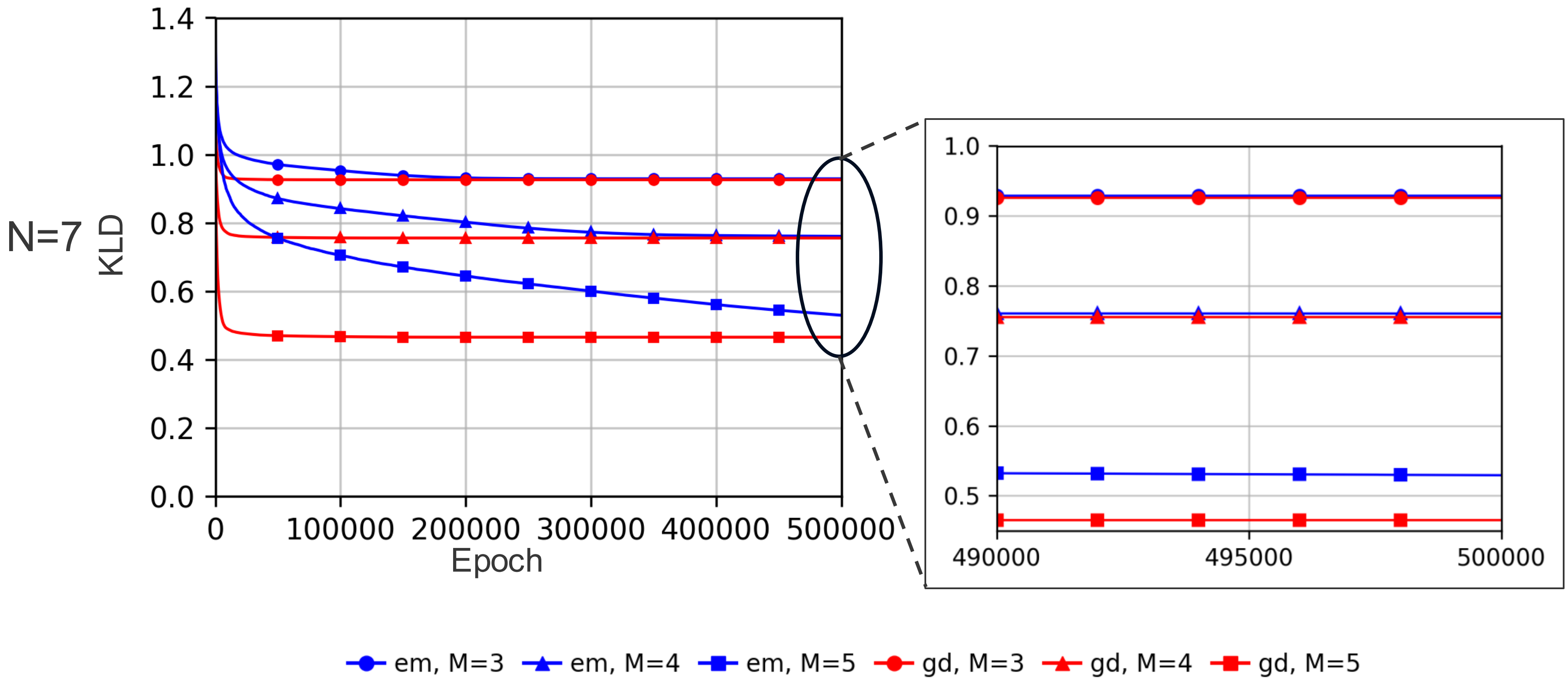
The blue lines consistently achieve a lower final KLD than the red lines for different numbers of hidden units.



em algorithm vs GD (KL divergence)

The standard GD method performed slightly better

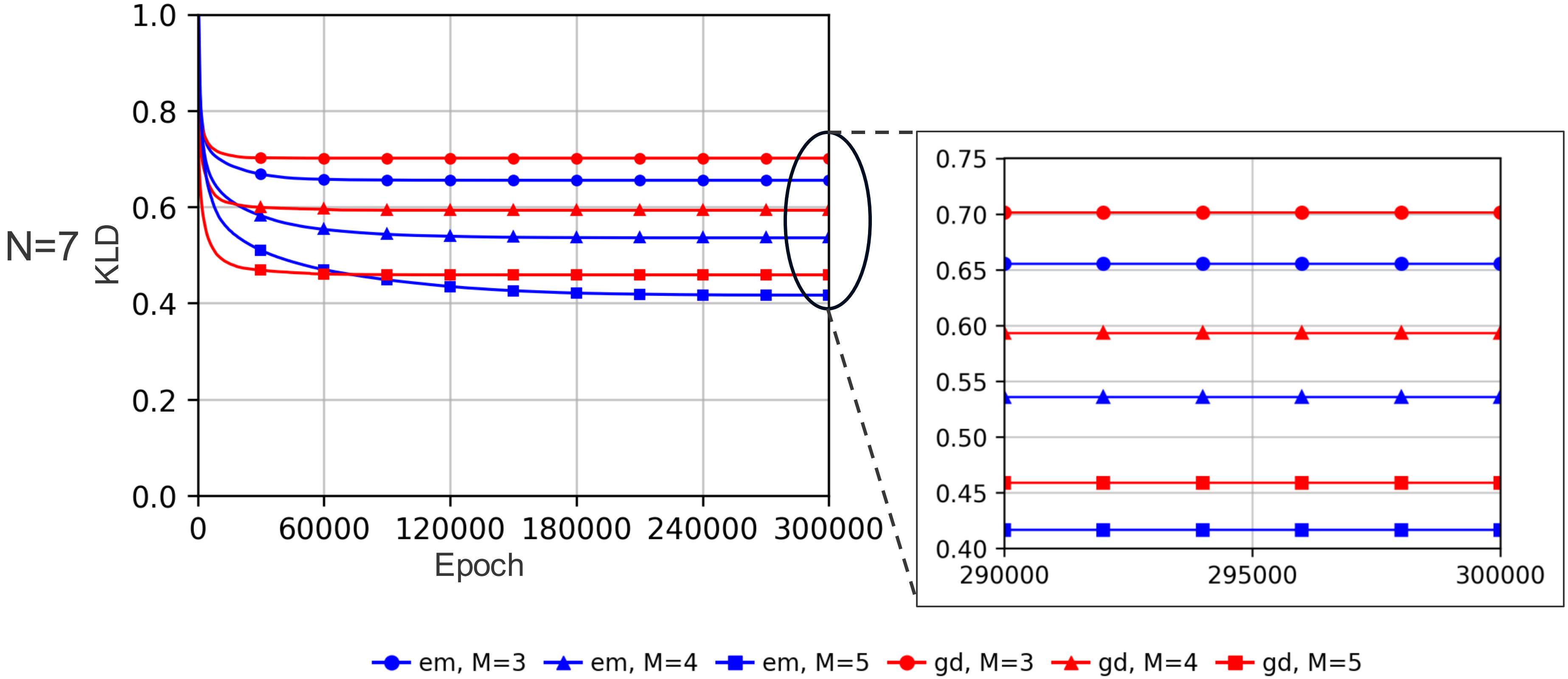
(C) Cardinality dataset



em algorithm vs GD (KL divergence)

The blue lines consistently achieve a lower final KLD than the red lines for different numbers of hidden units.

(D) $\mathcal{O}(n^2)$ dataset



Conclusion

- Contributions:
  - Proposed em algorithm for QBMs.
  - Analytical update rules in sqRBM.
  - Demonstrated stable and effective learning.
  - Potential to avoid barren plateau
- Experimental results:
  - em > GD in 3/4 datasets
  - GD better on Cardinality dataset
- Limitations:
  - Convergence speed.
- Future Work:
  - Faster optimization (accelerated GD) to utilize convexity of m-step.
  - Extension to fully quantum RBMs.

References

[Peruzzo+2014] Peruzzo, Alberto, et al. "A variational eigenvalue solver on a photonic quantum processor." Nature communications 5.1 (2014): 4213.

[Amin+2018] Amin, Mohammad H., et al. "Quantum Boltzmann machine." Physical Review X 8.2 (2018): 021050.

[McClean+2018] McClean, Jarrod R., et al. "Barren plateaus in quantum neural network training landscapes." Nature communications 9.1 (2018): 4812.

[Marrero+2021] Ortiz Marrero, Carlos, Mária Kieferová, and Nathan Wiebe. "Entanglement-induced barren plateaus." PRX quantum 2.4 (2021): 040316.

[Wang+2021] Wang, Samson, et al. "Noise-induced barren plateaus in variational quantum algorithms." Nature communications 12.1 (2021): 6961.

[Amari+1992] Amari, Shun-ichi, Koji Kurata, and Hiroshi Nagaoka. "Information geometry of Boltzmann machines." IEEE Transactions on neural networks 3.2 (1992): 260-271.

[Dempster+1977] Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm." Journal of the royal statistical society: series B (methodological) 39.1 (1977): 1-22.

[Ackley+1985] Ackley, David H., Geoffrey E. Hinton, and Terrence J. Sejnowski. "A learning algorithm for Boltzmann machines." Cognitive science 9.1 (1985): 147-169.