Quantum advantage in variational Bayes inference

Hideyuki Miyahara and Vwani Roychowdhury

Edited by Eli Yablonovitch, University of California, Berkeley, CA; received July 27, 2022; accepted February 14, 2023

Variational Bayes (VB) inference algorithm is used widely to estimate both the parameters and the unobserved hidden variables in generative statistical models. The algorithm—inspired by variational methods used in computational physics—is iterative and can get easily stuck in local minima, even when classical techniques, such as deterministic annealing (DA), are used. We study a VB inference algorithm based on a nontraditional quantum annealing approach—referred to as quantum annealing variational Bayes (QAVB) inference—and show that there is indeed a quantum advantage to QAVB over its classical counterparts. In particular, we show that such better performance is rooted in key quantum mechanics concepts: i) The ground state of the Hamiltonian of a quantum system—defined from the given data—corresponds to an optimal solution for the minimization problem of the variational free energy at very low temperatures; ii) such a ground state can be achieved by a technique paralleling the quantum annealing process; and iii) starting from this ground state, the optimal solution to the VB problem can be achieved by increasing the heat bath temperature to unity, and thereby avoiding local minima introduced by spontaneous symmetry breaking observed in classical physics based VB algorithms. We also show that the update equations of QAVB can be potentially implemented using $\lceil \log K \rceil$ qubits and $O(K)$ operations per step, where $K$ is the number of values hidden categorical variables can take. Thus, QAVB can match the time complexity of existing VB algorithms, while delivering higher performance.

Quantum machine learning (QML) primarily deals with quantum algorithms and quantum-inspired algorithms for data analysis and is an emerging research field that is forming new bridges between the traditional fields of physics and machine learning. Several QML frameworks, such as quantum principal component analysis (qPCA) (1) and quantum recommendation systems (2), have been introduced that show significant quantum speedups while achieving the same performance as the corresponding classical algorithms. These quantum algorithms, in turn, were later shown to have classical counterparts, and randomized algorithms with the same time complexity were derived (3, 4). This discovery process showed an encouraging synergy where the principles of quantum mechanics can also facilitate the design of better classical algorithms. Interest in QML has also been fueled by the emergence of noisy intermediate-scale quantum (NISQ) devices. The low fidelity and limited scale of such devices prevent implementations of well-known algorithms such as the Shor’s factorization algorithm or combinatorial optimization algorithms based on quantum annealing. However, efficient QML algorithms for conventional machine learning (ML) tasks, such as dimensionality reduction, clustering, classification, and Bayesian inference, could likely be implemented on NISQ devices and show potential quantum advantages in speed or accuracy. For example, variational quantum classifiers (VQC) and quantum circuit learning (QCL) frameworks have been proposed that hold the promise of time and hardware efficient training and realizations of conventional classifiers (5–7). Recent results, however, show that simple kernel method-based classical classifiers are guaranteed to have better performance than their quantum counterparts. Furthermore, there is no analytical guarantee or numerical evidence suggesting that the variational quantum algorithms will even have reasonable performance, especially for high-dimensional datasets where these algorithms are expected to have speedup advantages.

The above-mentioned examples underscore the general trend in the QML field: Existing algorithms can potentially speed up classical algorithms, but QML algorithms that outperform their classical ML counterparts are very rare or nonexistent. Thus, the search for QML algorithms that either perform better than any classical algorithm without incurring significant computational overheads or exhibit significant speedups

Significance

Quantum machine learning (QML) is an emerging research field that deals with quantum algorithms for data analysis. It is hoped that QML will yield practical demonstrations of quantum advantage by exploiting the emerging noisy intermediate-scale quantum (NISQ) devices, which cannot yet implement large-scale quantum algorithms. Most of the proposed QML frameworks, such as quantum principal component analysis, quantum circuits, and quantum recommendation systems, provide potential quantum speedups of corresponding classical algorithms. These algorithms, thus, do not improve the quality of the solutions, and QML algorithms that outperform classical ML schemes are rare or nonexistent. This paper shows an example of how quantum mechanics can lead to better solutions for machine learning (ML) problems without incurring increased time complexity overheads.
(for the same performance) continues to remain an active area of interest. More interestingly, there is a report of utilizing NISQ devices for ML (8).

In this paper, we address the problem of variational Bayes (VB) inference, which is a popular technique in ML, and explore how quantum mechanics can help design an algorithm with better performance than the existing classical techniques. In fact, principles from classical statistical physics have already inspired a genre of algorithms for VB. The history of optimization algorithms motivated by physics dates back to simulated annealing (SA) (9), which utilizes a thermostat to overcome the local optima problem in optimization, and SA has been applied to several ML tasks (10, 11). SA approaches, however, have a well-known drawback in that they require an infinitely long annealing schedule to guarantee the global optimum of an optimization problem or at least a very long annealing schedule to reach its equilibrium state at a finite temperature. To fix such drawbacks of SA, deterministic annealing (DA) was developed and applied to several machine learning problems (12). For example, by applying DA to variational Bayes (VB) inference (13, 14), deterministic annealing variational Bayes (DAVB) inference (15) was proposed. However, it has been shown that DA and DAVB find them with low probability. Moreover, these numerical results show that the QA part of QAVB is critically important for optimal parameter estimation and is the key to obtaining better performance than classical algorithms. For results on higher dimensional datasets where QAVB outperforms VB and DAVB, please refer to ref. 16. Then, we show that the QAVB update steps are completely positive and trace-preserving (CPTP) maps. Since it is known that a CPTP map can be implemented on quantum systems, we thus show that QAVB can be implemented using NISQ devices, comprising only \( \log K \) qubits.

Variational Bayes Inference

Suppose that we have \( N \) observable data points \( Z_{\text{obs}} := \{z_{i,\text{obs}}\}_{i=1}^{N} \) that are the output of an unknown generative model \( p_{\text{gen}}^{\theta}(z) \); \( z_{i,\text{obs}} \sim p_{\text{gen}}^{\theta}(\cdot) \) with additional dynamics that are not necessarily observed. One of the important approaches in ML is to assume that the generative model can be well approximated by a parameterized model that outputs both the observable data points \( Z_{\text{obs}} \) as well as an associated set of unobservable or hidden data points, \( \Sigma := \{\sigma_{j}\}_{j=1}^{K} \), where \( \sigma_{j} \in \{1, 2, \ldots, K\} \) is a categorical variable with \( K \) outcomes. These hidden variables often have interpretable meanings and can be used to predict other outcomes associated with the dataset. The task then is to estimate the parameters of the generative model and the posterior distributions of the unobservable variables from the observable data.

More specifically, we first introduce the underlying model via a distribution \( p_{\text{gen}}^{\theta}(z, \sigma | \theta) \), which is the conditional probability distribution of \( z \) and \( \sigma \) when \( \theta \) is given, and \( p_{\text{pr}}(\theta) \), which is the prior distribution of \( \theta \). Here, \( \theta \) is the set of parameters that characterize the conditional distribution and \( \Sigma := \{\sigma_{j}\}_{j=1}^{K} \) is the set of unobservable variables. For the above modeling to be successful, \( p_{\text{gen}}^{\theta}(z, \sigma | \theta) \) and \( p_{\text{gen}}(z) \) have to satisfy \( p_{\text{gen}}^{\theta}(\cdot) \approx \sum_{\sigma} p_{\text{gen}}^{\theta}(z, \sigma | \theta) = p_{\text{gen}}(z) \).

Then, VB is an algorithm to compute the posterior distribution of \( \theta \) and the hidden variables \( \Sigma \) in the above setup. In particular, the posterior distribution \( p_{\Sigma,\theta|Z}(Z, \Sigma, \theta | Z_{\text{obs}}) := \frac{p_{\Sigma,\theta|Z}(Z, \Sigma, \theta | Z_{\text{obs}}) p_{\text{pr}}(\theta)}{p(Z_{\text{obs}})} \) is computationally intractable, as it is difficult to compute \( p(Z_{\text{obs}}) \). Note that \( p(Z_{\text{obs}}) := \sum_{\Sigma,\theta} p_{\Sigma,\theta|Z}(Z_{\text{obs}}, \Sigma, \theta) p_{\text{pr}}(\theta) \). Then, in VB, we try to approximate \( p_{\Sigma,\theta|Z}(Z, \Sigma, \theta | Z_{\text{obs}}) \) by introducing a variational function \( q_{\theta}(\Sigma, \theta) \) and minimizing the Kullback–Leibler (KL) divergence between \( q_{\theta}(\Sigma, \theta) \) and \( p_{\Sigma,\theta|Z}(Z, \Sigma, \theta | Z_{\text{obs}}) \). Specifically, we solve

\[
q_{\theta}(\Sigma, \theta) = \underset{q_{\theta}(\Sigma, \theta)}{\arg \min} \text{KL} \left( q_{\theta}(\Sigma, \theta) \left\| \frac{p_{\Sigma,\theta|Z}(Z, \Sigma, \theta | Z_{\text{obs}})}{p(Z_{\text{obs}})} \right\right)
\]

[1]
where the KL divergence between \( p(x) \) and \( q(x) \), defined over their domain \( S^x \), is given by
\[
\text{KL}(p(x) \| q(x)) := \sum_{x \in S^x} p(x) [\ln p(x) - \ln q(x)].
\] [2]

In Eq. 2, \( x \) is assumed to be discrete, but almost the same definition is applicable for a continuous variable by replacing the summation with an integral. Furthermore, after making the assumption of MF where \( q^\Sigma,\theta(S, \theta) = q^\Sigma(\Sigma) q^\theta(\theta) \), the optimization problem on the right-hand side of Eq. 1 is solved iteratively by setting
\[
q^\Sigma_{t+1}(\Sigma) = \exp \left( \int_{\theta \in S^\theta} d\theta \ q^\theta_{t+1}(\theta) \ln \left( \hat{p} Z^\Sigma(\Sigma; \theta) \right) \right),
\] [3]
\[
q^\theta_{t+1}(\theta) = \exp \left( \sum_{\Sigma \in S^\Sigma} q^\Sigma_t(\Sigma) \ln \left( \hat{p} Z^\Sigma(\Sigma; \theta) \right) \right),
\] [4]

where \( q^\Sigma_t(\Sigma) \) and \( q^\theta_t(\theta) \) are the distributions of \( \Sigma \) and \( \theta \) at the \( t \)-th iteration, respectively (13, 14). Once we get the posterior distributions of \( \Sigma \) and \( \theta \), we can utilize them for inference problems.

**Motivations of Quantization and a Noncommutative Term**

The optimization problem in Eq. 1, however, is still highly nonconvex with multiple local minima, and finding good solutions is a challenging task. We explain this difficulty of VB from the viewpoint of quantum statistical mechanics and then show how this seeming escalation of complexity introduced by viewing a Bayesian problem as a quantum system leads to a better solution to the original VB problem. In statistical physics, the probability \( w_\beta(\beta) \) of finding a system in a configuration with energy \( E_\beta \) is given by \( w_\beta(\beta) := e^{-\beta E_\beta} / Z(\beta) \), where \( \beta := (k_B T)^{-1} \), \( k_B \) is the Boltzmann constant, \( T \) is the temperature of a heat bath to which the system is attached, and \( Z(\beta) := \sum_{n=0}^{\infty} e^{-\beta E_\beta} \). We can now reverse directions, and, given the VB problem, we can construct a virtual physical system such that \( \hat{p} Z^\Sigma(\Sigma; \theta) \) is the probability of being in configuration \( Z^\Sigma(\Sigma; \theta) \) conditioned by \( \theta \). Then, this system is defined by energy levels \( \epsilon_{\Sigma;\theta} := -\frac{1}{\beta} \ln \hat{p} Z^\Sigma(\Sigma; \theta) \). Since the next step is to construct a quantum system, it is more convenient to use the concept of a Hamiltonian, which specifies the energy level corresponding to every configuration of a system; for our classical system, the Hamiltonian is identical to the energy levels. We first define two Hamiltonians corresponding to the probabilities, \( \hat{p} Z^\Sigma(\Sigma; \theta) \) and \( \hat{p} \theta(\theta) \):
\[
\hat{H}^\Sigma_{\text{cl}} := -\ln \hat{p} Z^\Sigma(\Sigma; \theta),
\] [5]
\[
\hat{H}^\theta_{\text{pr}} := -\ln \hat{p} \theta(\theta).
\] [6]

Next, we convert this classical physical system to quantum ones using the canonical quantization approach (17). We denote the projection operator on \( \Sigma \) and \( \theta \) by \( \hat{\rho}^{\Sigma,\theta} := |\Sigma, \theta \rangle \langle \Sigma, \theta| \); then, we can write the Hamiltonian operators of Eqs. 5 and 6 as
\[
\hat{H}^\Sigma_{\text{cl}} := \sum_{\Sigma \in S^\Sigma} \int_{\theta \in S^\theta} d\theta \ \hat{H}^\Sigma_{\text{cl}}(\Sigma, \theta),
\] [7]
\[
\hat{H}^\theta_{\text{pr}} := \sum_{\Sigma \in S^\Sigma} \int_{\theta \in S^\theta} d\theta \ \hat{H}^\theta_{\text{pr}}(\Sigma, \theta),
\] [8]

where \( S^\Sigma \) and \( S^\theta \) are the domains of \( \Sigma \) and \( \theta \), respectively. Note that the dimension of the Hamiltonian is \( K^N \) if \( \theta \) is not quantized and infinity if \( \theta \) is quantized. These Hamiltonians are still diagonal, and thus, the system is still classical. Each, diagonal element is by definition, \( \epsilon_{\Sigma;\theta} := -\ln \hat{p} Z^\Sigma(\Sigma; \theta) \). Since we are soon going to develop the framework for estimating these probabilities by defining a nondiagonal Hamiltonian, it is useful to introduce the notation of the Gibbs operator
\[
\hat{f}(\beta_{\text{pr}}, \beta) := \exp \left( -\beta_{\text{pr}} \hat{H}^\theta_{\text{pr}} - \beta \hat{H}^\Sigma_{\text{cl}} \right),
\] [9]
and rewrite the probabilities back in terms of the Hamiltonian notation. For simplicity, we consider the case of a noninformative prior distribution, so that \( \hat{H}^\theta_{\text{pr}} \) is not a function of \( \theta \). Since we are still dealing with a diagonal Hamiltonian, we can rewrite Eq. 1 in the Hamiltonian formulation:
\[
\hat{p}_\beta^{\Sigma,\theta}(\Sigma, \theta) = \text{arg min}_{\hat{\rho}^{\Sigma,\theta}} \mathcal{S}(\hat{\rho}^{\Sigma,\theta} \| \hat{p} Z(\beta)) |_{\beta=1},
\] [10]
where \( \mathcal{Z}(\beta) \) is the partition function at \( \beta : \mathcal{Z}(\beta) := \text{Tr} \left[ e^{-\beta \hat{H}^\Sigma_{\text{cl}}} \right] \) and \( \beta \) is the inverse temperature. Furthermore, \( \mathcal{S}(\| \cdot \|) \) is the quantum relative entropy, which is a quantum extension of the KL divergence, Eq. 2, given by
\[
\mathcal{S}(\hat{\rho} \| \hat{\sigma}) := \text{Tr} \left[ \hat{\rho} \ln \hat{\rho} - \hat{\rho} \ln \hat{\sigma} \right].
\] [11]

The optimization problem in Eq. 10 is as difficult as Eq. 1 since they are equivalent. Let us first consider a simpler problem by taking the limit \( \beta = \infty \) in Eq. 10; then, we have a
\[
|0; \text{cl}) \langle 0; \text{cl}| = \text{arg min}_{\hat{\rho}^{\Sigma,\theta}} \lim_{\beta \to \infty} \mathcal{S}(\hat{\rho}^{\Sigma,\theta} \| \hat{p} Z(\beta)) \rangle, \] [12]
where \(|0; \text{cl}) \rangle \langle 0; \text{cl}| \) is the ground state of \( \hat{H}^\Sigma_{\text{cl}} \). Eq. 12 implies that at \( \beta = \infty \), Eq. 1 becomes the problem of finding the ground state of the data-defined Hamiltonian in Eq. 7. As explained next, we can use a variant of the quantum annealing technique to approximate such a ground state, denoted as \(|0; \text{cl}) \rangle \langle 0; \text{cl}| \).

We next consider the relationship between the populations of the canonical distributions at \( \beta = \infty \) and at \( \beta = 0 \) and discuss how such canonical distributions might evolve as \( \beta \) is changed adiabatically to \( \beta = 1 \). Note that a canonical distribution at \( \beta = 1 \) corresponds to an optimal solution to the VB problem. Fig. 1B and C show schematic representations of populations of the canonical distributions at \( \beta = \infty \) and at \( \beta = 1 \), respectively. As shown in Fig. 1B and C, once we get \(|0; \text{cl}) \rangle \langle 0; \text{cl}| \), we obtain \( \hat{p}_\beta^{\Sigma,\theta}(\Sigma, \theta) \) in Eq. 10 by pumping up the population of the

---

* SI Appendix, section 2 for the details of the canonical distribution.
We denote the energy level of state, then one could raise temperature slowly to reach the canonical distribution at the change of the energy spectrum of initialization. In contrast, for other methods based on classical statistics or Monte Carlo methods, the challenge is to start with a canonical distribution at any \( s \) uses a variant of quantum annealing to approximate the ground state at close to \( \beta \) starts at a very high temperature (i.e., \( \beta_0 \approx 0 \)) and gets stuck at a saddle point or starts with a random initialization of the distribution at a finite temperature (which would not be the canonical distribution for that temperature) getting easily stuck in local minima and leading to different estimations sensitive to the initial choice.

Fig. 1. Quantum advantage in VB explained using schematics. (A) a typical energy spectrum of \( \hat{H} = \sum_{n=0}^{\infty} \epsilon_n |n\rangle \langle n| \) populations of the canonical distributions at \( \beta = \frac{1}{T} = \infty, 1, 0 \), respectively (where \( T \) is the temperature of the bath attached to the system), and (F) that of a typical noncanonical distribution. We denote the energy level of \( |n\rangle \) by \( \epsilon_n \) for \( n = 0, 1, 2, \ldots \) and assume that \( \epsilon_0 \leq \epsilon_1 \leq \cdots \leq \epsilon_n \). A mixed state is written as \( \hat{\rho} = \sum_{n=0}^{\infty} w_n |n\rangle \langle n| \), where \( w_n \geq 0 \) for \( n = 0, 1, 2, \ldots \) and \( \sum_{n=0}^{\infty} w_n = 1 \). At \( \beta = \infty \), we have \( w_0 = 1 \) and \( w_n = 0 \) for \( n = 1, 2, \ldots \) while we have \( w_0 = \text{const.} \) for \( n = 0, 1, 2, \ldots \) at \( \beta = 0 \). (F) Schematic of the change of the energy spectrum of \( \hat{H}(s) \) from \( s_{\text{fin}} \) to \( s_{\text{ini}} \). By construction, the optimal solution to the VB problem corresponds to the canonical distribution of the corresponding physical system at \( \beta = 1 \). If one could start with the system in a canonical distribution at zero temperature (\( \beta \gg 1 \)), which is the ground state, then one could raise temperature slowly to reach the canonical distribution at \( \beta = 1 \) and, hence, obtain the optimum solution to the VB problem. QAVB uses a variant of quantum annealing to approximate the ground state at close to \( \beta \approx \infty \) and then increases the temperature to \( \beta = 1 \), leading to a closer approximation to the canonical distribution. Moreover, it requires only a single run (especially with \( s_0 = 1 \) as in Algorithm 1) without any dependence on initialization. In contrast, for other methods based on classical statistics or Monte Carlo methods, the challenge is to start with a canonical distribution at any \( \beta_0 < 1 \) and avoid having to cool the temperature where the system will get stuck in local minima. For example, the deterministic annealing method either starts with a very high initial temperature (where the canonical distribution is trivially known, i.e., uniform) and gets stuck at a saddle point or starts with a random initialization of the distribution at a finite temperature (which would not be the canonical distribution for that temperature) getting easily stuck in local minima and leading to different estimations sensitive to the initial choice.

Quantum Annealing Variational Bayes (QAVB) Inference

We describe QAVB by following ref. 16. In general, QA (18–20) is a method to find the ground state of a given Hamiltonian by using the adiabatic theorem, as shown in Fig. 1F. If we can design a parametrized Hamiltonian \( \hat{H}(\theta) \) such that \( \hat{H}(\theta_{\text{fin}}) \) is solvable and \( \hat{H}(\theta_{\text{fin}}) = \hat{H}_{\text{pr}} + \hat{H}_{\text{cl}}^{\theta(0)} \), then one can apply QA to approximate the desired ground state. In the case of QA, the dynamics are described by the Schrödinger equation; then, the adiabatic theorem holds for the dynamics of a time-dependent system. However, in our case, the state evolution follows the MF equation, and a similar property is not known. The analysis of adiabatic evolution in QAVB is one of the goals of this paper and is addressed in a later section.

In the rest of this paper, we formulate QAVB by adding a noncommutative term to the Hamiltonians of VB, Eq. 7 and Eq. 8, and confirm its validity. By using Eq. 7 and Eq. 8, we then define the following Gibbs operator:

\[
\hat{f}(\beta, s) := \exp \left( -\hat{H}_\text{pr} + \beta(1 - s)\hat{H}_\text{cl}^{\theta(0)} - \beta s \hat{H}_{\text{qu}}^{\theta(0)} \right). \tag{13}
\]

Here, the third term of Eq. 13 is given by \( \hat{H}_\text{qu}^{\theta(0)} := \sum_{i=1}^{N} \hat{H}_\text{qu}^{\theta(0)} \), and each term on the right-hand side satisfies the following noncommutative relation:

\[
\left[ \hat{H}_\text{qu}^{\theta(0)} \left( \int_{j=1}^{N} \int \hat{\sigma}_j \left( \int \sigma_i \right) \right) \hat{f}(\beta, s) \right] \neq 0. \tag{14}
\]

Here, \( \hat{\sigma}_j \) is a matrix such that \( \hat{\sigma}_j |\sigma_j\rangle = |\sigma_j\rangle |\sigma_j\rangle_i \) and \( \hat{f}(\cdot) \) is the identity operator for the corresponding Hilbert space. Using Eq. 11 and Eq. 13, we consider the following quantum relative entropy:

\[
S \left( \hat{\rho}^{\Sigma, \theta} \| \hat{f}(\beta, s) \right) := \text{Tr}_{\Sigma, \theta} \left( \hat{\rho}^{\Sigma, \theta} \left( \ln \hat{\rho}^{\Sigma, \theta} - \ln \hat{f}(\beta, s) \right) \right). \tag{15}
\]

where \( \mathcal{Z}(\beta, s) \) is the partition function given by \( \mathcal{Z}(\beta, s) := \text{Tr}_{\Sigma, \theta} \hat{f}(\beta, s) \). By minimizing Eq. 15 with respect to \( \hat{\rho}^{\Sigma, \theta} \), we can estimate the distribution of \( \theta \). However, the minimization problem of Eq. 15 is quite difficult; then, we utilize the following decomposition:

\[
\hat{\rho}^{\Sigma, \theta} \approx \hat{\rho}^{\Sigma} \otimes \hat{\rho}^{\theta}. \tag{16}
\]

Eq. 16 is often called the MF approximation. By performing the variational calculation of Eq. 15 with Eq. 16, we obtain the following update equations:

\[
\tilde{\rho}^{\Sigma}_{i+1} \propto \exp \left( \text{Tr}_{\Sigma} \left[ \left( \hat{f}(\beta, s_i) \right) \ln \hat{f}(\beta, s_i) \right] \right), \tag{17}
\]

\[
\tilde{\rho}^{\theta}_{i+1} \propto \exp \left( \text{Tr}_{\Sigma} \left[ \left( \hat{f}(\beta, s_i) \right) \ln \hat{f}(\beta, s_i) \right] \right). \tag{18}
\]

Finally, we summarize this algorithm in Algorithm 1. Note that the setting of \( s_0 = 1 \) in the algorithm ensures that there is...
Algorithm 1: Quantum annealing variational Bayes (QAVB)
inference with \( s_0 = 1 \).

1: set \( t \leftarrow 0 \) and \( p_{0}^{\beta}(\theta) \)
2: fix annealing schedules \( \{s_t\} \) and \( \{\beta_t\} \) such that \( s_0 = 1 \) and \( \beta_0 \gg 1.0 \)
3: while convergence criterion is not satisfied do
4: compute \( \rho_{t+1}^{\beta} \) in Eq. 18
5: compute \( \hat{\rho}_{t+1}^{\Sigma} \) in Eq. 17
6: \( t \leftarrow t + 1 \)
7: end while

no dependence of the results on the initial choice of \( \hat{\rho}_{0}^{\Sigma} \), and hence, this variant of the QAVB is executed only once for a given problem. In contrast, for DAVB and VB (also for QAVB where \( s_0 < 1 \); ref. 16), results are highly sensitive to the initial conditions and good outcomes are obtained with low probability.

There are multiple candidates for \( H_{qu}^{\Sigma} \) that satisfy Eq. 14. In numerical simulations, we use the following \( \hat{H}_{qu}^{\Sigma} \):

\[
\hat{H}_{qu}^{\Sigma} = \left( \sum_{j=1}^{i-1} \mathbb{I}^{q_j} \right) \otimes \left( \sum_{k=1}^{K} |\sigma_j = k\rangle \langle \sigma_j = k + 1 | + |\sigma_j = k + 1\rangle \langle \sigma_j = k | \right) \otimes \left( \sum_{j=i+1}^{N} \mathbb{I}^{q_j} \right) \otimes \mathbb{I}^{\theta}, \tag{19}
\]

where \( |\sigma_j = K + 1\rangle = |\sigma_j = 1\rangle \). To run QAVB, we also need to fix an annealing schedule; so, it is quite important to construct an efficient one. However, there are an infinite number of possible annealing schedules; so we need to limit ourselves. In ref. 16, the following annealing schedules for \( s_t \) and \( \beta_t = 1/T_t \), where \( T_t \) is the temperature of the bath to which the system is attached at time \( t \), are adopted:

\[
s_t = s_0 \times \max(1 - t/T_1, 0.0), \tag{20}
\]

\[
\beta_t = \begin{cases} 
\beta_0 & (t \leq t_1), \\
1 + \frac{(\beta_0 - 1)(t_2 - t)}{t_2 - t_1} & (t_1 \leq t \leq t_2), \\
1.0 & (t \geq t_2).
\end{cases} \tag{21}
\]

Note that Eq. 20 and Eq. 21 are characterized by four parameters: \( s_0, \beta_0, t_1, \) and \( t_2 \). Furthermore, the performance of QAVB on \( s_0 \) and \( \beta_0 \) is investigated in ref. 16, and it shows that \( s_0 = 1.0 \) and \( \beta_0 = 30.0 \) are effective. In Fig. 2A, we plot the annealing schedules described by Eqs. 20 and 21 with \( \beta_0 = 30.0, s_0 = 1.0, t_1 = 300, \) and \( t_2 = 350 \).

Mechanisms of QAVB

To discuss the dynamics of an estimate by QAVB, we focus on the annealing schedule described by Eqs. 20 and 21 with \( \beta_0 \gg 1.0 \) and \( s_0 = 1.0 \) since, as we see later, QAVB with this annealing schedule shows high performance. The annealing schedule can be divided into two parts. First quantum fluctuations are gradually decreased until they disappear at low temperature, and then, the temperature \( \beta \) is raised to 1, at which point the cost functions of QAVB and VB are identical. We develop a highly likely mechanism of QAVB on the basis of this decomposition as follows.

Due to the nature of the canonical distribution, the ground state of \( \hat{H}_{cl}^{\Sigma,\theta} \) dominates the density operator at finite but large \( \beta \gg 1 \). In the QA part of the annealing schedule, the state is expected to gradually vary from the ground state of the Hamiltonian, \( \hat{H}_{cl}^{\Sigma,\theta} \), that has a trivial ground state (by design) to that of the Hamiltonian of interest, \( \hat{H}_{cl}^{\Sigma,\theta} \) in Eq. 5. Of course, given the parameterized form of \( \hat{\rho}^{\Sigma,\theta} \) used in VB, and the MF approximation, \( \hat{\rho}^{\Sigma,\theta} = \hat{\rho}^{\Sigma} \otimes \hat{\rho}^{\theta} \) one can only approximate the ground state. Picking more expressive functional forms or, as shown in the numerical section, increasing the number of clusters \( K \) in the GMM estimation problem can improve the expressive power and lead to better approximation and improved performance.

Furthermore, the ground state corresponds to the hard clustering assignment, in the sense that each data point is assigned to exactly one categorical value. This follows from the observation.

Fig. 2. (A) Annealing schedules described by Eqs. 20 and 21 with \( \beta_0 = 30.0, s_0 = 1.0, t_1 = 300, \) and \( t_2 = 350 \). (B) Two-dimensional dataset generated by ten Gaussian functions. Each data point has a label. (C) Gaussian functions at step 112 estimated by DAVB with \( \beta_0 = 0.0010, s_0 = 0.0, t_1 = 10, \) and \( t_2 = 100 \). Only one Gaussian mode dominates; the rest have \( s_t = 0 \). This shows that when DAVB starts at a high temperature and is slowly cooled, it gets stuck in a saddle point.
Note that QAVB approximates the ground state well and obtains a good hard-clustering solution. QAVB at the end of the QA part is also very close to that at convergence. This demonstrates that QAVB can achieve almost optimal performance for a wide range of model parameter settings. For the two-dimensional data illustrated in Fig. 2, the best achievable success rates of QAVB at convergence and at the end of the QA part are presented here; for this dataset, the maximum possible success rate as obtained from the generative derivations is 0.9833 ± 0.0045. Numerical computations for the two-dimensional data illustrated in Fig. 2 are presented here; for this dataset, the maximum possible success rate as obtained from the generative model is 0.98 and is shown by the black dotted lines. The update steps are identical to that of QAVB when δ = 0; [SI Appendix](#) for derivations. However, if we start DAVB with high temperature (i.e., β ≈ 0), we cannot avoid SSB, and if we start it with low temperature, then the final estimate depends strongly on the initial configuration. Such deficiencies motivated us to set a target success rate, pcr, of 0.85 and 0.95, respectively. (C) Kmin(τ1), which is the minimum value of K to achieve a given pcr, as τ1 is varied, and (D) τ1min(K), which is the minimum value of τ1 to achieve a given pcr. For example, for τ1 = 500, one can match the best performance for K = 12. Note that in (C) and (D), the values are not monotonically decreasing. Note that when K is set larger than 10 (actual number of Gaussians), the posterior probabilities of only 10 modes match the corresponding ground-truth values, and the rest of the K − 10 modes have almost zero posterior probabilities. Similar results are shown for a 3-D dataset in Fig. 6.
Comparative average performance of DAVB and its dependence on the initial temperature \( \beta_0 \). Note that in DAVB, the performance depends strongly on the initial distribution, i.e., \( \hat{\beta}_0 \). For the two-dimensional dataset, we profile the dependence of the average prediction rate of DAVB on \( \beta_0 \) for (A) \( K = 20 \) and (B) \( K = 30 \). The number of times that achieves \( p_{cr} = 0.95 \) out of 1,000 runs for (C) \( K = 20 \) and (D) \( K = 30 \) is also plotted. As is expected, the average performance is inferior to that of QAVB, which for a wide range of choices of parameters gives a near-optimal result in a single run.

develop QAVB and analyze its dynamics. We show here why QAVB has a different dynamics, allowing it to outperform other methods.

Numerical Simulations

For demonstrating quantum advantage in VB inference and to showcase the dynamics of QAVB, we apply QAVB to the well-known clustering problem using the Gaussian mixture model (GMM). In the numerical simulations, two datasets are investigated: two-dimensional and three-dimensional datasets generated by the GMM\(^1\). These low-dimensional datasets are sufficient to demonstrate the various factors that contribute to the successful dynamics of QAVB. For applications of QAVB to higher dimensional datasets, please ref. 16. In Fig. 2B, the first dataset is shown. To quantify performance, we define the prediction rate, or success rate, as the ratio of how many hidden variables are correctly estimated (i.e., how many data points are assigned to the same Gaussian as in the model that generated the data) to the total number of data points. Note that there is an arbitrariness on the permutation of hidden variables; thus, we use the maximum value with respect to the permutation as the prediction rate.

We use Eqs. 20 and 21 and set \( \beta_0 = 30.0 \) and \( s_0 = 1.0 \) for the annealing schedule of the experiments. With these settings, Algorithm 1 is run only once for any given choices of \( K \), \( \tau_1 \), and \( \tau_2 \). The prediction rate is a function of the hyperparameters \( K \) and \( \tau_1 \), and we hereafter denote it by \( p_{\text{succ}}(K, \tau_1) \). Clearly, the duration of the QA steps, \( \tau_1 \), determines how closely one can track the ground state and plays a crucial role. The number of clusters, \( K \), determines the number of parameters and, hence, the expressive power of \( \hat{\rho}^{\Sigma, \hat{\beta}} \) in approximating the ground state. One of the goals of our experiments is to show the tradeoffs between these two hyperparameters. Since \( p_{\text{succ}}(K, \tau_1) \) is not monotonic with respect to \( K \) and \( \tau_1 \), we define \( p_{\text{succ}}^{\text{min}}(K, \tau_1) := \max_{K' \leq K} p_{\text{succ}}(K', \tau_1) \) and \( p_{\text{succ}}^{\text{max}}(K, \tau_1) := \max_{\tau_1' \leq \tau_1} p_{\text{succ}}(K, \tau_1') \). In Fig. 3A and B, we plot the dependence of \( p_{\text{succ}}(K, \tau_1) \) on \( \tau_1 \) and the dependence of \( p_{\text{succ}}^{\text{min}}(K, \tau_1') \) on \( K \), respectively. These results show that for sufficiently large \( \tau_1 \) and \( K \), QAVB shows a high prediction rate that almost matches the upper bound set by the generative model, with full knowledge. Given prediction criterion \( p_{cr} \), we then define \( p_{\text{succ}}^{\text{min}}(\tau_1) := \arg \min_{\tau_1} p_{\text{succ}}(K, \tau_1) \) and \( p_{\text{succ}}^{\text{max}}(K) := \arg \min_{\tau_1} p_{\text{succ}}(K, \tau_1) \) subject to \( p_{\text{succ}}(K, \tau_1) \geq p_{cr} \). In Fig. 3C and D, we plot \( p_{\text{succ}}^{\text{min}}(\tau_1) \) and \( p_{\text{succ}}^{\text{max}}(K) \), respectively. These figures show that, to achieve \( p_{cr} = 0.85, 0.95 \), relatively small \( K \) and \( \tau_1 \) are enough.

Next, we turn our attention to DAVB. We again use Eq. 21 for \( \beta_0 \) in DAVB and set \( \tau_1 = 10 \) and \( \tau_2 = 200, 300, 300 \). In Fig. 4A and B, we plot the dependence of the average prediction rate of DAVB on \( \beta_0 \) for \( K = 20, 30 \), respectively. And, in Fig. 4C and D, we also plot the number of times that achieves \( p_{cr} = 0.95 \) on \( \beta_0 \) for \( K = 20, 30 \), respectively. These figures show that the average prediction rate of DAVB is much lower than that of QAVB, and DAVB rarely achieves \( p_{cr} = 0.95 \).

\(^1\)SI Appendix, section 3) document for the details of GMM. The prior and posterior distributions for the GMM are described in SI Appendix, section 4.
We next present numerical results on a three-dimensional dataset. In Fig. 5, we plot the dependence of $p_{\text{suc}}^K(K, \tau_1)$ on $\tau_1$ and the dependence of $p_{\tau_1}^s(K, \tau_1)$ on $K$, respectively.

In Fig. 6 A and B, we plot the dependence of the average prediction rate of DAVB and the number of times that achieves $p_{\tau_0} = 0.95$ on $\beta_0$, respectively. Here, we set $\tau_1 = 10$, $\tau_2 = 300$, and $K = 20$.

The numerical result on the three-dimensional dataset is consistent with the case of the two-dimensional dataset though it is quantitatively different from the case of the two-dimensional dataset.

To understand the dynamics of QAVB, it is instructive to study cluster assignments of QAVB at the end of the QA part of the annealing schedule at $t = \tau_1$ and at convergence. We show the cluster assignments of QAVB with $\beta_0 = 30.0$ in Fig. 7 A and B. The estimates at the end of the QA part are almost as those at convergence. Only a few of the clusters in the ground-truth data are split. Thus, the process of raising temperature is not that important. This is expected due to the absence of SSB, and it is quite reasonable to focus on the QA part. Fig. 7 A and B show that, in the case of low temperature, QAVB successfully estimates the ground state, while QAVB does not in the case of high temperature. These results are also consistent with the discussions of the possible mechanism of QAVB.

In Table 1, we show the success rates of QAVB at convergence and at the end of the QA part and the best possible success rates with full knowledge of the generative model. Ten datasets were created by using the same generative model to create the dataset shown in Fig. 2, and then the mean and SD of the performance were computed. The success rate of QAVB at convergence is very close to that of the generative model, and that of QAVB at the end of the QA part is also very close to them. This observation reflects two points. The first one is simply that QAVB is successful for the generative model under consideration. The second one is that the soft clustering defined via the minimization problem of the KL divergence, which is Eq. 10 at $\beta = 1$ and $s = 0$, and the hard clustering defined via that at $\beta \gg 1$ are similar. In the case of the hard clustering problem, however, the data points are classified into a larger number of clusters. That is, out of the $K$ Gaussians ($K > 10$ where the generative model has 10 Gaussians), more than 10 have $\pi_j > 0.01$. Thus, the success rate of QAVB at the end of the QA part is slightly worse than at convergence.

**Adiabatic-Theorem-like Property: Similarity and Differences Between QA and QAVB**

In QA, the total Hamiltonian is constructed by the convex combination of a Hamiltonian that describes an optimization problem of interest and a noncommutative Hamiltonian that can be easily diagonalized. Similar to QA, we construct Eq. 13 by the convex combination of two Hamiltonians. On the other hand, the main difference is that QA solves the Schrödinger equation, but QAVB solves the MF equation. Thus, the adiabatic theorem (17) does not directly hold. We next analytically
Assuming that $\hat{\rho}_t^\Sigma = |0; \beta_{t-1}, s_{t-1}; \Sigma \rangle \langle 0; \beta_{t-1}, s_{t-1}; \Sigma|$ and $|0; \beta_{t-1}, s_{t-1}; \Sigma | 0; \beta_t, s_t; \Sigma \rangle \approx 1$, Eq. 25 becomes

$$
\ln \hat{\rho}_{t+1}^\theta = \text{Tr}_\Sigma \left[ \left( |0; \beta_{t-1}, s_{t-1}; \Sigma \rangle \langle 0; \beta_{t-1}, s_{t-1}; \Sigma | \otimes \hat{\rho}_t^\theta \right) \times \sum_n \epsilon_n^{\Sigma}(\beta, s) |n; \beta_t, s_t; \Sigma \rangle \langle n; \beta_t, s_t; \Sigma | \otimes |n; \beta_t, s_t; \theta \rangle \langle n; \beta_t, s_t; \theta | \right] + \text{const.} [26]
$$

and a similar computation can also be done for Eq. 24. Note that a constant multiple does not affect the physical property of $\hat{\rho}_{t+1}^\theta$. In the numerical simulations of QAVB, we varied $s$ at fixed $\beta$ at the first QA part; then, the above assumption is reasonable. Thus, the discussion here analytically gives the reason why QAVB shows high performance. In particular, it explains the mechanism by which QAVB gives the ground state of $\hat{\rho}_{t+1}^\Sigma$ in Eq. 5.

**Time Complexity of QAVB and QAVB as Quantum Dynamics**

The main focus of this paper is to quantify how much better QAVB can perform compared with VB and provide analytical results on its dynamics. From the viewpoint of practical applications, its computational complexity is also an important metric. In the case of a classical computer, the time complexity of VB with respect to the number of clusters $K$ is $O(K)$ since VB has single loops on $K$. On the other hand, QAVB requires one to compute the exponentials of $K \times K$ matrices; thus, the time complexity of QAVB with respect to $K$ is $O(K^3)$. Note that, similarly to VB, the time complexity of QAVB with respect to the number of data points is $O(N)$; thus, it practically works on a classical computer.

The above situation changes if we assume a quantum computer. The simulations of a quantum system does not increase time complexity compared with that of a classical system. Furthermore, if we can find a local Hamiltonian that describes QAVB, then we can expect a quantum speedup with respect to $K$ (21). From this viewpoint, it is worth considering physical implementations of QAVB. From Eqs. 17 and 18, the relationship between $\hat{\rho}_{t+1}^\Sigma$ and $\hat{\rho}_{t+1}^\theta$ is written as

$$
\hat{\rho}_{t+1}^\Sigma = \frac{1}{Z_{t+1}} \exp \left( \text{Tr}_{\beta_t} \left[ \left( \hat{\rho}\Sigma^\otimes \hat{\rho}_t^\theta \right) \ln \hat{f}(\beta_t, s_t) \right] \right),
$$

where

$$
Z_{t+1} = \text{Tr}_{\beta_t} \left[ \exp \left( \text{Tr}_{\beta_t} \left[ \left( \hat{\rho}\Sigma^\otimes \hat{\rho}_t^\theta \right) \ln \hat{f}(\beta_t, s_t) \right] \right) \right].
$$
In Eq. 29, two types of operations are involved: the exponential operation and partial trace. We show that these two operations are both CPTP because it can be realized in a physical process (22).

First, we discuss that the exponential map is CP when the input density operator is positive semidefinite. Here, we basically follow ref. 23. It is enough to say that the composite operator $\exp(\mathcal{A}_n)$ is positive for $n = 1, 2, \ldots$ where $\mathcal{A}_n$ is an arbitrary $n$-dimensional operator. As described in ref. 23, the family of positive definite operators is closed under point-wise addition and point-wise multiplication; thus, $\exp(\mathcal{A}_n)$ is positive for $n = 1, 2, \ldots$; thus, the exponential map is completely positive. Note that a Hamiltonian is not necessarily positive semidefinite, but we can always add a constant shift such that the Hamiltonian becomes positive semidefinite. The map of interest is TP because of the partition function, though the exponential map itself is not TP.

Next, we turn our attention to partial trace. We can say that partial trace operation is completely positive by constructing a Kraus operator directly. Let us consider $\hat{K}_{\alpha} = \hat{I}_{A} \otimes |\alpha\rangle$. In general, a density operator for subsystems $A$ and $B$ has the form $\rho_{AB} := \sum_{ij} \lambda_{ij} |ij\rangle \langle ij| \otimes |\mu\rangle \langle \mu|$, and taking partial trace with respect to subsystem $B$ yields $\text{Tr}_B[\rho_{AB}] = \sum_{ij} \lambda_{ij} \sum_{\mu} \langle ij| \langle \mu| \langle \alpha| \mu \rangle |\alpha\rangle = \sum_{ij} \lambda_{ij} |ij\rangle \langle ij|$. On the other hand, $\hat{K}_{\alpha}$ leads to $\sum_{\alpha} \hat{K}_{\alpha} \rho_{AB} \hat{K}_{\alpha}^\dagger = \sum_{ij} \lambda_{ij} |ij\rangle \langle ij| \hat{K}_{\alpha} \rho_{AB} |\alpha\rangle \langle \alpha| \hat{K}_{\alpha}^\dagger = \text{Tr}_B[\rho_{AB}]$. Thus, we have shown that $\hat{K}_{\alpha}$ is the Kraus operator for partial trace. For more details, refer to ref. 22. Thus, we have shown that Eq. 29 is CPTP. In other words, QAVB is physically implementable.

Discussions

In this paper, we have analyzed the dynamics of QAVB by developing an analytical framework and providing numerical simulations to support the analytical results. In particular, we developed a theoretical framework to understand why there is a quantum advantage in variational bayesian inference. Next, via numerical analysis, we confirmed that the QA part of QAVB is essential by showing that an estimate at the end of the QA part is almost the same as an estimate at convergence at finite temperature. Thus, an optimal solution to the VB problem is essentially obtained at the end of the QA part, and increasing temperature does not affect the estimates very much. We also showed that the estimate at the end of the QA part of the annealing part corresponds to the hard clustering assignment. Second, we developed an adiabatic-theorem-like result that shows that the QA also holds in the case of the MF dynamics. Then, we explained that this generalized QA framework is why QAVB is efficient and gives a quantum advantage. Finally, we discussed the physical realizability of QAVB by showing that QAVB can be expressed as a CPTP map. This discussion tells us that QAVB can be realized in a quantum system. We expect this work to motivate physics-inspired algorithms and further research on emerging fields at the intersection of physics and machine learning.

We have provided physics-based arguments for the quantum advantage in variational Bayes inference, and we have left a rigorous mathematical proof as future work. Rigorous mathematical proofs for recently proposed algorithms which are widely considered to have quantum advantage are also lacking. For example, the quantum approximate optimization algorithm (QAOA) (24) and the variational quantum eigensolver (VQE) (25), which are considered to be equivalent to each other, have attracted much attention as methods to efficiently utilize NISQ devices, and a large number of their variants have been proposed. The main theoretical support for them is that, for $N \to \infty$, the QAOA realizes the adiabatic evolution, where $N$ is the number of layers in a circuit (24). However, the proof cannot be applied to the QAOA of finite layers; in other words, their computational advantage is not known for the practical setup. As discussed in this paper, the proposed algorithm is expected to have a practical advantage for ML.

Data, Materials, and Software Availability. All study data are included in the article and/or SI Appendix. The code for the algorithms proposed in this paper and the code used for generating data can be found at the following github link: https://github.com/hmihayara512/QAVB (26).

Author affiliations: *Department of Electrical and Computer Engineering, Henry Samueli School of Engineering and Applied Science, University of California, Los Angeles, CA 90095

1. S. Lloyd, M. Mohseni, P. Rebentrost, Quantum principal component analysis. Nat. Phys. 10, 631–634 (2014).
2. I. Kerenedis, A. Prakash, Quantum recommendation systems. arXiv [Preprint] (2016). http://arxiv.org/abs/1603.08675 (Accessed 5 July 2023).
3. E. Tang, “A quantum-inspired classical algorithm for recommendation systems” in Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing (2019), pp. 217–228.
4. E. Tang, Quantum-inspired classical algorithms for principal component analysis and supervised clustering. arXiv [Preprint] (2018). http://arxiv.org/abs/1811.00414 (Accessed 5 July 2023).
5. K. Mitali, M. Negro, M. Kitagawa, K. Fuji, Quantum circuit learning. Phys. Rev. A 99, 023209 (2018).
6. M. Schuld, A. Bocharov, K. M. Svore, N. Wiebe, Circuit-centric quantum classifiers. Phys. Rev. A 101, 032308 (2020).
7. H. Miyahara, V. Roychowdhury, Ansa...